

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

MULTI-SOURCE MULTIMODAL DEEP  
LEARNING TO IMPROVE SITUATION  
AWARENESS: AN APPLICATION OF  
EMERGENCY TRAFFIC MANAGEMENT

A THESIS PRESENTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY  
IN  
EMERGENCY MANAGEMENT  
AT MASSEY UNIVERSITY, WELLINGTON,  
NEW ZEALAND.

Hewa Algiriyage Rangika Nilani

2023

# Contents

<b>Abstract</b>	<b>xiii</b>
<b>Acknowledgements</b>	<b>xv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Research Questions and Objectives . . . . .	3
1.2 Research Context . . . . .	3
1.3 Research Outline . . . . .	4
1.4 The Structure of the Thesis . . . . .	6
<b>2 Multi-source Multimodal Big Data and Deep Learning for Disaster Research: A Systematic Review</b>	<b>9</b>
Abstract . . . . .	9
2.1 Introduction . . . . .	10
2.2 Research Question Synthesis . . . . .	11
2.2.1 The First Component of Learning: The Target Function . . . . .	12
2.2.2 The Second Component of Learning: The Training Data . . . . .	12
2.2.3 The Third and Fourth Components of Learning: The Learning Algorithm and Hypothesis Test . . . . .	13
2.2.4 The Fifth Component of Learning: The Final Hypothesis . . . . .	14
2.2.5 The Final Analysis . . . . .	14
2.3 Methodology . . . . .	15
2.3.1 Develop a research plan . . . . .	16
2.3.2 Search the relevant articles . . . . .	17
2.3.3 Apply exclusion criteria . . . . .	17
2.3.4 Extract relevant data from the selected articles . . . . .	17
2.3.5 Analyse data using the Knowledge Discovery in Databases (KDD) process . .	18
2.3.5.1 Association Rule Mining . . . . .	18
2.4 <b>RQ<sub>1</sub></b> : What types of DR problems have been addressed by DL approaches? . . . .	19
2.5 <b>RQ<sub>2</sub></b> : How have the training datasets been extracted, preprocessed, and used in DL-based approaches for DR tasks? . . . . .	22
2.5.1 <i>RQ<sub>2.1</sub> What types of DR data have been used?</i> . . . . .	22
2.5.2 <i>RQ<sub>2.2</sub> What sources have been used to extract data, and how have data been extracted?</i> . . . . .	23
2.5.3 <i>RQ<sub>2.3</sub> How have data been preprocessed before applying the DL models?</i> . . .	25
2.6 <b>RQ<sub>3</sub></b> : What DL models are used to support DR tasks? . . . . .	27

2.6.1	<i>RQ<sub>3.1</sub> What types of DL architectures are used?</i>	27
2.6.2	<i>RQ<sub>3.2</sub> What training processes are used to optimize DL models?</i>	29
2.6.3	<i>RQ<sub>3.3</sub> What methods are used to avoid overfitting and underfitting?</i>	30
2.7	<b>RQ<sub>4</sub>: How well do DL approaches perform in supporting various DR tasks?</b>	31
2.7.1	<i>RQ<sub>4.1</sub> What evaluation matrices are used to evaluate the performance of DL models?</i>	31
2.7.2	<i>RQ<sub>4.2</sub> What “baseline” models have been compared?</i>	33
2.8	<b>RQ<sub>5</sub>: What are the underlying challenges and replicability of DL for DR studies?</b>	34
2.9	Opportunities, directions and future research challenges	35
2.10	Results of the Association Rule Mining	36
2.11	Flowchart and guidelines for applying DL in future DR research	37
2.12	Conclusions	39
<b>3</b>	<b>Research Methodology and Design</b>	<b>40</b>
3.1	Research Philosophy	40
3.2	Design Science Research (DSR)	42
3.2.1	The relevance cycle	42
3.2.2	The rigor cycle	43
3.2.3	The design cycle	43
3.2.4	Method Framework for Design Science Research	44
3.3	Research Methodology	45
3.4	Research Methods	47
3.4.1	Data collection methods	47
3.4.2	Data analysis methods	49
3.4.2.1	Quantitative data analysis	49
3.4.2.2	Qualitative data analysis	50
3.5	Chapter Summary	51
<b>4</b>	<b>Requirements Definition, Design and Development of the Artefact</b>	<b>52</b>
4.1	Explicate Problem	52
4.2	Define Requirements	55
4.3	Design and Develop Artefact	60
4.4	Chapter Summary	63
<b>5</b>	<b>Traffic Flow Estimation from CCTV Data</b>	<b>64</b>
	Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images	64
	Abstract	65
5.1	Introduction	65
5.2	Related Work	66
5.3	Methodology	69
5.3.1	Data set	70
5.3.1.1	Data processing	70
5.3.2	Experiments	70
5.3.2.1	Experiment 1	70



5.3.2.2	Experiment 2 . . . . .	71
5.4	Results and Discussion . . . . .	71
5.5	Conclusion . . . . .	73
Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance		
	Video Footage . . . . .	74
	Abstract . . . . .	74
5.6	Introduction . . . . .	74
5.7	Related work . . . . .	76
5.8	Methodology . . . . .	77
5.8.1	Dataset . . . . .	77
5.8.2	Vehicle detection . . . . .	78
5.8.3	Vehicle tracking . . . . .	79
5.8.4	Vehicle movement direction estimation and traffic flow counting . . . . .	79
5.9	Results . . . . .	80
5.10	Conclusion . . . . .	82
5.11	Summary . . . . .	82
<b>6</b>	<b>Deep Learning-based Short Term Traffic Flow Prediction with Weather Data</b>	<b>83</b>
	Abstract . . . . .	83
6.1	Introduction . . . . .	84
6.2	Related Work . . . . .	85
6.3	Bi-LSTM for Traffic Flow Prediction . . . . .	87
6.3.1	LSTM Networks . . . . .	87
6.3.2	Bi-LSTM Networks . . . . .	88
6.3.3	Dataset . . . . .	89
6.3.3.1	Traffic Data . . . . .	90
6.3.3.2	Weather Data . . . . .	90
6.3.4	Problem Formulation . . . . .	94
6.3.5	Experiments . . . . .	94
6.4	Experimental Results and Discussion . . . . .	95
6.5	Conclusions . . . . .	96
<b>7</b>	<b>DEES - A real-time system for event extraction from disaster-related web text</b>	<b>98</b>
	DEES - A real-time system for event extraction from disaster-related web text . . . . .	98
7.1	Introduction . . . . .	99
7.2	Related Work . . . . .	101
7.3	DEES : Description of Methodos and System . . . . .	102
7.3.1	News and tweet extraction . . . . .	103
7.3.2	Noise filtering . . . . .	103
7.3.3	Relevant tweet identification . . . . .	104
7.3.4	Clustering . . . . .	104
7.3.5	Candidate extraction . . . . .	104
7.3.6	Candidate scoring . . . . .	106
7.4	Evaluation . . . . .	109
7.5	Discussion and Conclusion . . . . .	111

Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison . . . . .	112
Abstract . . . . .	112
7.6 Abstract . . . . .	113
7.7 Introduction . . . . .	113
7.8 Related Work . . . . .	115
7.9 Methodology . . . . .	117
7.9.1 Dataset . . . . .	117
7.9.2 Models . . . . .	119
7.9.3 Experiments . . . . .	121
7.10 Results and Discussion . . . . .	122
7.11 Conclusion . . . . .	124
<b>8 Real-time Information Extraction from Multi-source Multimodal Data: An Ap- plication of Emergency Traffic Management</b>	<b>125</b>
Abstract . . . . .	125
8.1 Introduction . . . . .	126
8.2 Related Work . . . . .	128
8.3 Methodology . . . . .	131
8.3.1 Text data handling - $MoD_1$ . . . . .	132
8.3.1.1 Text data extraction - $SMoD_{1.1}$ . . . . .	132
8.3.1.2 Noise filtering - $SMoD_{1.2}$ . . . . .	133
8.3.1.3 Relevant tweet identification - $SMoD_{1.3}$ . . . . .	133
8.3.1.4 Clustering - $SMoD_{1.4}$ . . . . .	133
8.3.2 Visual data handling - $MoD_2$ . . . . .	134
8.3.2.1 Visual data extraction - $SMoD_{2.1}$ . . . . .	134
8.3.2.2 Visual data classification - $SMoD_{2.2}$ . . . . .	134
8.3.2.3 Visual data captioning - $SMoD_{2.3}$ . . . . .	134
8.3.3 Candidate extraction - $MoD_4$ . . . . .	136
8.3.3.1 Candidate extraction from news, tweets and image captions - $SMoD_{4.1}$	136
8.3.4 Candidate scoring and event template creation - $MoD_5$ . . . . .	140
8.3.4.1 Candidate scoring for news, tweets, and image captions - $SMoD_{5.1}$	140
8.3.4.2 Impact information scoring - $SMoD_{5.2}$ . . . . .	143
8.3.4.3 Event template creation from news, tweets, and image captions - $SMoD_{5.3}$ . . . . .	143
8.4 Evaluation . . . . .	144
8.5 Discussion and Conclusions . . . . .	146
<b>9 Prototype Evaluation</b>	<b>148</b>
9.1 Evaluation in DSR research . . . . .	148
9.2 Analyse Evaluation Context . . . . .	151
9.3 Select Evaluation Goals and Strategy . . . . .	152
9.4 Design and Carry Out Evaluation . . . . .	156
9.4.1 Round 1 Evaluations . . . . .	156
9.4.2 Findings of the Round 1 interviews . . . . .	158
9.4.3 Round 2 Evaluations . . . . .	169

9.4.4	Findings of the Round 2 interviews . . . . .	172
9.5	Chapter Summary . . . . .	178
<b>10</b>	<b>Discussion and Conclusion</b>	<b>180</b>
10.1	Motivation for study . . . . .	180
10.2	Addressing the research questions . . . . .	181
10.2.1	Research Question 1: How have different deep learning algorithms been applied to data from various sources to support disaster response tasks? . . . . .	181
10.2.2	Research Question 2: How can data from multiple sources be fused to support disaster response? . . . . .	182
10.2.3	How can the integration of multi-source multimodal data effectively support disaster response by cross-validating social media data? . . . . .	183
10.3	Contributions of the study . . . . .	183
10.3.1	Contribution to research . . . . .	183
10.3.1.1	Systematic literature review: . . . . .	184
10.3.1.2	Traffic flow estimation from CCTV images: . . . . .	184
10.3.1.3	Traffic flow estimation from CCTV footage: . . . . .	185
10.3.1.4	Short-term traffic flow prediction: . . . . .	186
10.3.1.5	Event extraction from multi-source unimodal data: . . . . .	187
10.3.1.6	Disaster-related tweet classification: . . . . .	188
10.3.1.7	Real-time information extraction from multi-source multimodal data for disaster response . . . . .	189
10.3.2	Contribution to practice . . . . .	190
10.4	Research Impacts . . . . .	193
10.5	Research Implications, limitations and future work . . . . .	195
10.6	Conclusion . . . . .	196
	<b>Appendices</b>	<b>197</b>
	<b>A Deep Learning - A short overview</b>	<b>198</b>
	<b>B Information Sheet</b>	<b>207</b>
	<b>C Consent Form</b>	<b>210</b>
	<b>D Prototype Evaluation User Guide and Interview Questions - Round 1</b>	<b>212</b>
	<b>E Interview Questions - Round 2</b>	<b>222</b>
	<b>F Human Ethics Notification</b>	<b>228</b>
	<b>G DRC 16 Forms</b>	<b>231</b>
	<b>References</b>	<b>239</b>

# List of Tables

- 1.1 Research outline. . . . . 4
- 2.1 Attributes in the data extraction form . . . . . 18
- 2.2 Main DR tasks of the analysed articles . . . . . 22
- 2.3 Disaster data collection methods . . . . . 24
- 2.4 Data preprocessing steps. . . . . 26
- 2.5 Best Accuracy Scores for DR tasks. . . . . 32
- 2.6 Some association rules extracted from the analysed papers . . . . . 36
- 3.1 Alternative research methodologies. . . . . 46
- 3.2 Comparison of data collection methods in qualitative research. . . . . 48
- 4.1 Background of participants selected for the interviews. . . . . 55
- 4.2 Findings from interview data. . . . . 57
- 5.1 Popular object detection models and their objectives . . . . . 68
- 5.2 Data set before and after pre-processing . . . . . 70
- 5.3 Performance and accuracy of the three models for our CCTV data set . . . . . 71
- 5.4 Details of the image dataset used to train YOLOv4 . . . . . 78
- 5.5 Details of the analysed CCTV videos (hr: hours, mins: minutes and secs: seconds) . 79
- 5.6 Man Average Precision (mAP) of vehicle detector classes . . . . . 80
- 5.7 Number of vehicles counted by humans (ground-truth), automatically by our algorithm and the accuracy for video 01 (day), video 02 (evening) and video 03 (night). . 81
- 6.1 Traffic Flow Data . . . . . 91
- 6.2 Weather Data . . . . . 91
- 6.3 Traffic flow prediction results - univariate time series model . . . . . 95
- 6.4 Traffic flow prediction results - fusion model . . . . . 95
- 7.1 Absolute and relative time expressions . . . . . 105
- 7.2 Dependencies and POS of the news headline “Motorcycles crash leaves one dead near Whakatāne” . . . . . 105
- 7.3 Candidates extracted from news headlines . . . . . 106
- 7.4 Candidates extracted from tweets . . . . . 106
- 7.5 Candidates extracted from news body text . . . . . 106
- 7.6 Scores used to determine the best candidates among the candidate set extracted from news headlines, body, tweets, and image captions . . . . . 107
- 7.7 Position score calculation for *where* candidates of the news in Tables 7.3 and 7.5 . . 107

7.8	Frequency score calculation for <i>where</i> candidates of the news and tweet examples in Tables 7.3, 7.4 and 7.5 . . . . .	107
7.9	Location relatedness score calculation for <i>where</i> candidates of the news and tweet examples in Tables 7.3, 7.4 and 7.5. . . . .	108
7.10	Final <i>where</i> candidate selection of the news and tweet examples in Tables 7.3, 7.4 and 7.5. . . . .	108
7.11	Details of the evaluation dataset before preprocessing . . . . .	109
7.12	Generalised precision scores of 3W . . . . .	111
7.13	Tweets text data . . . . .	111
7.14	News + tweets text data . . . . .	111
7.15	The summary tweet classification studies. . . . .	116
7.16	Related and Not-Related labelled tweets grouped by disaster category . . . . .	118
7.17	In-disaster, out-disaster and cross-disaster experimental datasets . . . . .	121
7.18	Average F1-scores of the DL and ML models for the in-disaster experiments (The best scores are highlighted in grey, and the three best performing ML models and the best performing DL model are underlined). . . . .	122
7.19	Average F1-scores of the ML and DL models (LSTM-fastText (DL <sup>1</sup> ), CNN-fastText (DL <sup>2</sup> ), LSTM-GloVe (DL <sup>3</sup> ), CNN-GloVe (DL <sup>4</sup> ), LSTM-Word2Vec (DL <sup>5</sup> ) and DL-Word2Vec (DL <sup>6</sup> )) for the Out-disaster experiments. The best scores are highlighted in grey. . . . .	123
7.20	Average F1-scores of the ML and DL models (LSTM-fastText (DL <sup>1</sup> ), CNN-fastText (DL <sup>2</sup> ), LSTM-GloVe (DL <sup>3</sup> ), CNN-GloVe (DL <sup>4</sup> ), LSTM-Word2Vec (DL <sup>5</sup> ) and DL-Word2Vec (DL <sup>6</sup> )) for the Out-disaster experiments. The best scores are highlighted in grey. . . . .	123
8.1	Applications using multimodal data for disaster response . . . . .	130
8.2	Dependencies and POS of the news headline “One dead after serious crash in Auckland”	137
8.3	Absolute and relative time expressions. . . . .	138
8.4	Candidates extracted from news headlines . . . . .	139
8.5	Candidates extracted from tweets . . . . .	140
8.6	Candidates extracted from news body text . . . . .	140
8.7	Candidates extracted from image captions . . . . .	141
8.8	Scores used to determine the best candidates among the candidate set extracted from news headlines, body, tweets, and image captions . . . . .	141
8.9	Position score calculation for <i>where</i> candidates of the news in Tables 8.4 and 8.6 . .	141
8.10	Frequency score calculation for <i>where</i> candidates of the examples in Tables 8.4, 8.5, 8.6 and 8.7 . . . . .	142
8.11	Location relatedness score calculation for <i>where</i> candidates of the news and tweet examples in Tables 8.4, 8.5 and 8.6. . . . .	142
8.12	Final <i>where</i> candidate selection of the examples in Tables 8.4, 8.5, 8.6 and 8.7 . . . .	143
8.13	Traffic accident-related impacts during 2020 and 2021 in New Zealand . . . . .	143
8.14	Scoring of impact information . . . . .	143
8.15	Generalised precision scores of 3W+Impact . . . . .	145
8.16	Tweets text data . . . . .	145
8.17	News + tweets text data . . . . .	145

8.18	News + tweets text and visual data . . . . .	145
9.1	Summary of the selected experts for the prototype evaluation . . . . .	152
9.2	Quality attributes of multiple quality models . . . . .	154
9.3	Research strategies and methods for different evaluation strategies adapted from Venable et al. . . . .	155
9.4	Results of the preliminary interview analysis . . . . .	158
9.5	Problems of the artefact identified from the cognitive walkthrough . . . . .	160
9.6	Suggestions received for the improvement of artefact through cognitive walkthrough	163
9.7	Mapping of the problems and suggestions to sub-categories of FURPS model and the sub-categories of the thematic analysis process . . . . .	165

# List of Figures

1.1	The use of different data modalities in manuscripts. Manuscript numbers are as listed in Table 1.1. . . . .	6
1.2	Thesis chapter outline. . . . .	7
2.1	The components of learning as proposed by Abu Moftha . . . . .	12
2.2	Literature selection process. . . . .	16
2.3	Publication venues of the articles . . . . .	17
2.4	Papers published per year according to DR task . . . . .	20
2.5	Taxonomy of DR tasks . . . . .	21
2.6	Data types used for DR task . . . . .	23
2.7	Sources used to extract data types . . . . .	24
2.8	DL architectures used by DR tasks except for CNN as a single architecture . . . . .	27
2.9	Usage of CNN by DR tasks . . . . .	28
2.10	DL architectures used by DR tasks by year . . . . .	28
2.11	Pre-trained DL networks used by DR tasks. . . . .	29
2.12	Methods used to avoid overfitting and underfitting by DR tasks . . . . .	30
2.13	Flowchart for conducting DL for DR research . . . . .	38
3.1	Design science research cycles adapted from Hevner et al. . . . .	42
3.2	Method framework for DSR introduced by Johannesson et al. . . . .	45
3.3	A decision flow chart for data collection in machine learning research . . . . .	49
3.4	The relationships between research methodologies and methods. The selected methodologies and methods are highlighted in blue letters. . . . .	51
4.1	Explicate Problem activity . . . . .	54
4.2	Define Requirements activity . . . . .	60
4.3	Design sketch of the artefact . . . . .	61
4.4	Design and Development activity . . . . .	62
4.5	The relationship between algorithms developed in Chapters 5 - 8 . . . . .	63
5.1	R-CNN architecture . . . . .	67
5.2	Fast R-CNN architecture . . . . .	67
5.3	YOLO object detection . . . . .	67
5.4	Methodology. . . . .	69
5.5	A sample CCTV image . . . . .	70
5.6	Camera "unavailable" image . . . . .	70

5.7	Vehicle Detection (a) faster R-CNN-Day (b) mask R-CNN-Day (c) YOLOv3 R-CNN-Day (d) faster R-CNN-Night (e) mask R-CNN-Night (f) YOLOv3-Night (g) faster R-CNN-Blur (h) mask R-CNN-Blur (i) YOLOv3-Blur . . . . .	72
5.8	Left Lane. . . . .	73
5.9	Right Lane. . . . .	73
5.10	A sample of the obtained vehicle counts . . . . .	73
5.11	A plot showing a sample of traffic flow . . . . .	73
5.12	Two-stage detection vs single-stage detection (a) R-CNN architecture (b) YOLO object detection . . . . .	76
5.13	Methodology for Real-time traffic flow estimation. . . . .	78
5.14	Drawing of the location analysed - Line coordinates (L1[0], L1[1], L2[0], L2[1]) and bounding box properties of a vehicle object (x, y, width (w), height (h)). . . . .	79
5.15	Traffic flow estimation from video footage. . . . .	80
5.16	Live plots of traffic flow (a) Directional traffic flow (b) Traffic flow by vehicle class. . . . .	81
6.1	Architecture of a LSTM network . . . . .	87
6.2	The architecture of Bi-LSTM Network . . . . .	89
6.3	Proposed model architecture for the traffic and weather data fused model and traffic only model. . . . .	90
6.4	(a) Selected location “SH73 Yaldhurst Rd” in the map. (b) Region of interest selection for the traffic flow counting using CCTV images . . . . .	91
6.5	A plot of all-weather parameters for the considered duration. . . . .	92
6.6	Pearson coefficient calculation. . . . .	93
7.1	The First box consists of a news article, the title (bold), and the second box consists of four tweets. Highlighted phrases in each sentence represent the answers to what, when and where questions . . . . .	100
7.2	Real-time event extraction process. 1. News and tweet extraction, 2. Relevant tweet identification, 3. Noise filtering, 4. Clustering, 5. Candidate extraction, and 6. Candidate scoring. . . . .	102
7.3	Dependency parsing of news headline “Motorcycles crash leaves one dead near Whakatāne”	105
7.4	Home screen showing events . . . . .	109
7.5	Screen showing more details of news and tweets . . . . .	109
7.6	Scores used for identifying best candidates among the candidate set extracted from news headlines, body, and tweets . . . . .	110
7.7	Wordcloud representation of news headlines . . . . .	110
7.8	Wordcloud representation of tweets content . . . . .	110
7.9	The the distribution of length of news headlines and tweets, in terms of words . . . . .	111
7.10	The Olteanu categorization for tweets . . . . .	114
7.11	Related and Not-related tweets by category. . . . .	119
7.12	Related and Not-related tweets in the dataset. . . . .	119
7.13	Illustration of (a) CNN and (b) Bidirectional LSTM of twitter text classification task.	120
8.1	Online news, images and tweet text with complementary information . . . . .	127
8.2	The proposed architecture . . . . .	131



8.3	Inception-v3-based disaster-related image classification model . . . . .	134
8.4	VGG-16-based image captioning model . . . . .	136
8.5	Available and auto-generated captions for two images downloaded from online news and tweets . . . . .	136
8.6	Dependency parsing of news headline “One dead after serious crash in Auckland”. The ROOT node is highlighted in orange. . . . .	137
8.7	Relationship patterns of ROOT node and other nodes . . . . .	138
8.8	Impact information visualization in event templates . . . . .	144
8.9	Symbols used to show impact information . . . . .	144
9.1	Evaluation framework proposed by Pries-Heje et al. . . . .	149
9.2	Demonstrate Artefact 8.1 . . . . .	150
9.3	Evaluate Artefact activity . . . . .	151
9.4	Evaluation process used in this research study . . . . .	156
9.5	Software artefact evaluated during Round 1 interviews . . . . .	157
9.6	Initial screen and view more screen . . . . .	171
9.7	Third screen of the software prototype . . . . .	172
9.8	Results of Functionality Evaluation . . . . .	173
9.9	Results of Usability Evaluation . . . . .	174
9.10	Results of Reliability Evaluation . . . . .	175
9.11	Results of Performance Evaluation . . . . .	176
9.12	Results of Performance Evaluation . . . . .	177
9.13	Results of Satisfaction Questions . . . . .	178
10.1	Initial screen of the software prototype . . . . .	190
10.2	View more screen of the software prototype . . . . .	191
10.3	Final screen of the software prototype . . . . .	192
10.4	An overview of how the software prototype supports all three levels of SA. . . . .	193

# Abstract

Traditionally, disaster management has placed a great emphasis on institutional warning systems, and people have been treated as victims rather than active participants. However, with the evolution of communication technology, today, the general public significantly contributes towards performing disaster management tasks challenging traditional hierarchies in information distribution and acquisition. With mobile phones and Social Media (SM) platforms widely being used, people in disaster scenes act as non-technical sensors that provide contextual information in multiple modalities (e.g., text, image, audio and video) through these content-sharing applications. Research has shown that the general public has extensively used SM applications to report injuries or deaths, damage to infrastructure and utilities, caution, evacuation needs and missing or trapped people during disasters.

Disaster responders significantly depend on data for their Situation Awareness (SA) or the dynamic understanding of “the big picture” in space and time for decision-making. However, despite the benefits, processing SM data for disaster response brings multiple challenges. Among them, the most significant challenge is that SM data contain rumours, fake information and false information. Thus, responding agencies have concerns regarding utilising SM for disaster response. Therefore, a high volume of important, real-time data that is very useful for disaster responders’ SA gets wasted.

In addition to SM, many other data sources produce information during disasters, including CCTV monitoring, emergency call centres, and online news. The data from these sources come in multiple modalities such as text, images, video, audio and meta-data. To date, researchers have investigated how such data can be automatically processed for disaster response using machine learning and deep learning approaches using a single source/ single modality of data, and only a few have investigated the use of multiple sources and modalities. Furthermore, there is currently no real-time system designed and tested for real-world scenarios to improve responder SA while cross-validating and exploiting SM data. This doctoral project, written within a “PhD-thesis-with-publication” format, addresses this gap by investigating the use of SM data for disaster response while improving reliability through validating data from multiple sources in real-time.

This doctoral research was guided by Design Science Research (DSR), which studies the creation of artefacts to solve practical problems of general interest. An artefact: a software prototype that integrates multisource multimodal data for disaster response was developed adopting a 5-stage design science method framework proposed by Johannesson et al. [175] as the roadmap for designing, developing and evaluating. First, the initial research problem was clearly stated, positioned, and root causes were identified. During this stage, the problem area was narrowed down to Emergency traffic management instead of all disaster types. This was done considering the real-time nature and data availability for the artefact’s design, development and evaluation.

Second, the requirements for developing the software artefacts were captured using the interviewing technique. Interviews were conducted with stakeholders from a number of disaster and emergency management and transport and traffic agencies in New Zealand. Moreover, domain knowledge and experimental information were captured by analysing academic literature. Third, the artefact was designed and developed. The fourth and final step was focused on the demonstration and evaluation of the artefact.

The outcomes of this doctoral research underpin the potential for using validated SM data to enhance the responder's SA. Furthermore, the research explored appropriate ways to fuse text, visual and voice data in real-time, to provide a comprehensive picture for disaster responders. The achievement of data integration was made through multiple components. First, methodologies and algorithms were developed to estimate traffic flow from CCTV images and CCTV footage by counting vehicle objects. These outcomes extend the previous work by annotating a large New Zealand-based vehicle dataset for object detection and developing an algorithm for vehicle counting by vehicle class and movement direction. Second, a novel deep learning architecture is proposed for making short-term traffic flow predictions using weather data. Previous research has mostly used only traffic data for traffic flow prediction. This research goes beyond previous work by including the correlation between traffic flow and weather conditions. Third, an event extraction system is proposed to extract event templates from online news and SM text data, answering What (semantic), Where (spatial) and When (temporal) questions. Therefore, this doctoral project provides several contributions to the body of knowledge for deep learning and disaster research. In addition, an important practical outcome of this research is an extensible event extraction system for any disaster capable of generating event templates by integrating text and visual formats from online news and SM data that could assist disaster responders' SA.

# Acknowledgements

I spent most of my PhD journey in a global pandemic, and I would not have been able to finish it without the love and support of amazing people around me.

I would want to begin by thanking my husband, Rangana Sampath, who left everything behind, including his loving family and well-established career, to come to New Zealand to assist me. Ranga, I am eternally thankful to you for being my shadow through all of my ups and downs. Without your immense support, this thesis would just have been a dream.

The reason and motivation behind my hard work were you Ranuka. You are my ultimate source of happiness and love. You were just three years when I started my PhD. However, you understood that your mother is busy at times and helped me at your best. Thanks a lot, my little prince, for your commitment.

To my little bundle of joy, Ranumi, it was not easy having you during the very final stage of my PhD through a complicated c-section surgery. I can not remember how many times I thought about giving up my PhD after your arrival. Thank you for your love and emotions, my little angel, for making me feel stronger every day.

I owe huge gratitude to my parents and brothers, whom I missed while doing my PhD. I was not able to see you for more than four years as COVID-19 travel restrictions affected international air travel. You always encouraged me to see the bright side amidst the multiple challenges I faced. I truly would not be able to submit this thesis without your love and encouragement. I also extend gratitude to my mother-in-law, sister-in-law and little Liyana for understanding and supporting me in many ways throughout the PhD journey.

I feel extremely privileged to have been offered an opportunity to study at Massey University, New Zealand and would like to thank the AHEAD project, Sri Lanka, for funding this PhD project. Further appreciation goes to my peers and colleagues at the University of Kelaniya for encouraging and helping me in many ways. As an international student whose first language is not English, academic writing was not an easy task for me. I would like to thank Dr Kendra Marston for her expertise in editing, commenting and highlighting the most important places in my writing. I would definitely like to comment that it was a turning point in my writing journey, the decision that I made to participate in Massey Universities' "get Published" course. A special thank goes to Dr Collin Bjork for providing me with many materials, video tutorials, guidelines and a platform to discuss writing issues with fellow PhD students.

I was fortunate to have an amazing set of friends around me at the Joint Centre for Disaster Research (JCDR), Massey University, Wellington. The informal weekly meetings you organized provided me with a valuable opportunity to discuss research matters. I am grateful for the support and encouragement from my friends and fellow PhD students, including Dr Nancy Brown, Ms Lesley Gray, Dr Marion Tan, Dr Miles Crawford, Ms Lisa McLaren, Dr Ashleigh Rushton, Dr Syed

Yasir Imtiaz, Dr Sara Harrison, Dr Bruce Pepperell, Mr Richard Mowll and Ms Ayisha Siddiqua. Thank you very much for being with me during the time that I was struggling with my research. A heartfelt "thank you" to Dr Marion Tan and Dr Sara Harrison for taking the time to read the chapters of my thesis and provide feedback. My sincere gratitude is also extended to all staff at the JCDR, who helped me in many ways. I truly enjoyed being a PhD student at the centre.

I would like to express my most profound appreciation to all officers from many governments and private authorities who joined my interviews. Thank you very much for spending your valuable time giving the essential information and feedback for my research. I consider every meeting significantly important, and I will forever be thankful to you all.

I finalize my acknowledgements by giving a heartfelt thank you to the most amazing four supervisors who made my PhD journey a successful one. I am extremely grateful to my primary supervisor, Dr Raj Prasanna, for his understanding, support and guidance throughout my PhD journey. The weekly meetings you organised helped significantly to extend my skills and knowledge as a researcher. Dr Kristin Stock always gave me feedback to improve my work with her expertise in computational methodologies. You motivated me to extend my thinking, which always resulted in high-quality research. Dr Emma Hudson-Doyle went through all the pieces of my writing and always provided me with constructive feedback. You continuously encouraged me to see beyond what I thought I was capable of achieving. A big thank you to Prof David Johnston for providing me with many platforms to present my research. I am forever grateful to you all for nurturing me as the researcher I have always wanted to become.

# Chapter 1

## Introduction

Different types of disasters, such as floods, hurricanes, earthquakes, and wildfires impact human lives and cause economic losses around the world every year. The frequency of disasters has increased over the last decade due to a variety of factors, including global warming, climate change, and rapid population growth [323]. Disaster response is the process of taking actions to minimise the negative effects of such disasters on people, animals, and the environment. Quick access to information on the situation is essential for responders to initiate activities such as evacuating and rescuing people in affected areas, allocating resources, prioritising relief efforts, and communicating situational information to the general public and other responders [108, 131, 45]. Situation Awareness (SA) or the dynamic understanding of “the big picture” in space and time is crucial for them to make more informed and effective decisions [98].

The increased use of mobile devices, particularly Social Media (SM) applications, has led to a significant increase in the amount of data generated during disasters. Given the rapid and widespread dissemination of information, there is growing interest in exploiting SM data for crisis event detection [124], understanding public sentiment [130], damage assessment [164, 412] and actionable information gathering for improving SA [339, 185]. However, despite their usefulness, the majority of these data are not available to decision-makers, mainly due to issues with data quality and data processing [14, 54, 7, 37, 63]. For example, researchers who have analysed SM data during disasters have found that rumours, misinformation, and false information are very common among SM messages [163]. Therefore, responding organisations have concerns about the trustworthiness of information available in SM channels [165, 195]. As a result, a significant amount of timely, valuable information on SM platforms gets wasted without being properly utilised.

Apart from SM applications, there are other sources such as remote sensing, wireless sensor networks, Internet of Things (IoT), online citizen responses, CCTV monitoring, and call centre recordings, which produce heterogeneous data during disasters [12, 313, 411]. Therefore, an overwhelming volume of multimodal data (e.g., text, image, audio, video, and meta-data) is generated within seconds of a disaster. As the magnitude and capabilities of data generation and collection grow exponentially, multi-source multimodal data can be used to create many useful avenues to cross-validate SM content and effectively utilise them in improving SA [12, 313]. Furthermore, a computerised application capable of combining such data can assist emergency management personnel by [309, 308, 61]:

- reducing information scarcity;
- reducing the uncertainty of decision-making;

- providing cross-validated results from multiple sources and hence increasing the reliability;
- reducing variations in decisions made by different individuals by providing a common interface;
- reducing the burden of integrating data manually;
- assisting in improving the accuracy of data through integration;
- reducing cognitive load;

Recent research has studied the use of Deep Learning (DL) algorithms to integrate different data modalities in applications such as audio-visual speech recognition, image captioning, event detection, and multimedia retrieval [271, 352, 40] (a short introduction to DL techniques is provided in Appendix A). Different features characterise different modalities of data (e.g., the visual modality has features such as colour, texture, and shape while the audio modality has pitch). The goal of multimodal data fusion is to relate these different features into a single joint feature space [190, 243]. DL techniques have become more popular for data fusion tasks due to their fast performance, the ability for automating feature engineering, the possibility to fill in missing modalities, and the option to obtain joint representation even in the absence of some modalities [352].

In the disaster domain, research has used DL to fuse multimodal data to investigate problems such as disaster event detection [20], disaster damage assessment [7, 132, 261], and disaster-related information filtering and classification [1, 54, 109, 152]. However, these methods have been mostly explored using static datasets available offline [1, 202, 7]. Considering the time-critical nature of disaster environments, one of the main concerns is that the responding organisations need real-time information for their decision-making tasks [309]. To date, there has been little research done on real-time multimodal data extraction and fusion. The research that has been done includes systems such as the Advanced System for Emergency Management (ASyEM) [107] and the Quelloffene Integrierte Multimedia Analyse (QuOIMA) [329]. However, both these systems have four main limitations:

- ASyEM uses multimodal data from a variety of sensors, including infrastructure sensors (e.g., video cameras), mobile system sensors (e.g., drones), and SM applications. On the other hand, QuOIMA uses multimodal data mainly from SM applications such as Facebook and Twitter. However, neither article explicitly describes how they combined multiple data modalities. Furthermore, it is not evident that any of those systems utilized DL for data fusion.
- Both systems are at the proposal stage; hence, they do not evaluate how successful the fusion algorithms are except for ASyEM running a simple test scenario. Therefore, they don't provide adequate details for future researchers to be able to use the systems as benchmarks.
- The focus of both systems has been to collect multimedia data, process them and share targeted information among different disaster responding units for their SA. However, the authors highlight the inability to test and validate the systems in various real-world situations. As a result, neither ASyEM nor QuOIMA provides examples of the use of the proposed architectures for real-time applications.
- None of these systems emphasizes the reliability issues associated with SM and suggests solutions to enhance the credibility of SM content in order to support responders' SA.

## 1.1 Research Questions and Objectives

The limitations highlighted above indicate that there has not been a proper system developed to effectively utilize SM data and to support disaster responders for their SA in real-time. Therefore, the identified research gaps led to the formulation of the following Research Questions (RQs):

**Research Question 1:** How have different deep learning algorithms been applied to data from various sources to support disaster response tasks?

**Research Question 2:** How can data from multiple sources be fused to support disaster response?

**Research Question 3:** How can the integration of multi-source multimodal data effectively support disaster response by cross-validating social media data?

This study intends to create and develop a real-time system that can incorporate multi-source multimodal data to assist disaster responders in addressing the above-mentioned problems. As a result, the data accessible via SM channels will be cross-validated, increasing trustworthiness, and responding organizations will be able to successfully use timely relevant SM data for their SA. The research questions identified above will be answered by achieving the following research objectives:

**Objective 1:** To identify deep learning algorithms that can be used to analyze different modalities of data extracted from different sources for disaster response.

**Objective 2:** To develop algorithms that are capable of integrating multi-source unimodal data (e.g., SM-text with news-text).

This is a necessary first step to achieve prior to the development of extended algorithms that address multiple modalities.

**Objective 3:** To develop algorithms that are capable of integrating multi-source multimodal data.

## 1.2 Research Context

United Nations Office for Disaster Risk Reduction (UNDRR) defines a disaster as a significant disruption in the operation of a community or society of any size [89]. Disasters are caused by the interaction of hazardous events with conditions of exposure, vulnerability, and capacity, resulting in human, material, economic, and environmental losses and impacts [89]. For instance, flooding, earthquakes, landslides, and terrorist attacks were among the most frequent disasters that occurred globally over the previous decade [90]. Sendai Framework for Disaster Risk Reduction 2015-2030 categorizes such disasters as small-scale, large-scale, frequent and infrequent, slow-onset, and sudden-onset based on the size of the affected population, probability of occurrences and the type of emergence of the disaster [340]. Sudden-onset disasters such as vehicle crashes cause roadways to close, creating traffic congestion. According to statistics released by the Ministry of Transport, there were 11,449 traffic accidents in New Zealand, including 2,449 that caused serious



and fatal injuries in 2019<sup>1</sup>. Multiple costs are incurred as a result, including more driving time, unexpected delays, missed appointments, increased fuel use and higher vehicle emissions. These added costs cause substantial economic loss and damage to the environment.

Managing traffic emergencies is one of the most challenging tasks that disaster managers face, requiring extensive planning to ensure safe, responsive, efficient, and sustainable transportation for everyone [296, 104]. During such emergencies, the road network becomes congested, making evacuation impossible and rescue personnel and supplies unable to be transported [21, 260]. Wrong decisions made without a clear picture of the situation have resulted in multiple unfortunate incidents and dozens of human casualties during mass evacuations. For example, during hurricane Rita in the Gulf of Mexico, 107 deaths were caused while evacuation, mostly due to a lack of adequate traffic control [60]. More recent examples include the diversity of traffic impacts from rain storms in late August 2022, ranging from significant flooding and prolonged road closure and damage across Nelson-Tasman through to the smaller-scale slips of road cuttings across the Wellington region that caused local, temporary, but significant traffic congestion in New Zeland [148].

The artefact developed during this PhD research was designed to function during any disaster. However, choosing a disaster scenario in which the researcher could access data and evaluate real-time algorithms was necessary. Sudden-onset disasters such as traffic incidents are prevalent, and data regarding these events for training DL algorithms was readily accessible via SM, CCTV surveillance, traffic monitoring agencies, and the media. Furthermore, the artefact developed from this research can be used in traffic operations to minimize the previously mentioned costs. Therefore, emergency traffic management was selected as a suitable application area to develop and test the algorithms.

### 1.3 Research Outline

This research study was conducted following Massey University’s thesis by publication model [91]. Four peer-reviewed journal articles answer the research questions in Chapters 2, 6, 7 and 8. In addition to the journal articles, three conference publications detail the development of the various components that underpin the final software system. The conference papers are presented in Chapters 5 and 6. The alignment of the research questions, objectives, manuscripts, chapters, and publication status is presented in Table 1.1.

**Table 1.1** Research outline.

Research Question	Research Objective	Manuscript	Chapter	Status
How have different deep learning algorithms been applied to data from various sources to support disaster response tasks?	To identify deep learning algorithms that can be used to analyze different modalities of data extracted from different sources for disaster response.	1 <sup>st</sup> Manuscript: “Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review”	2	Published <sup>2</sup>

<sup>1</sup><https://www.transport.govt.nz/statistics-and-insights/safety-annual-statistics/>

<sup>2</sup>Journal: SN Computer Science 2022

		2 <sup>nd</sup> Manuscript* :“Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images”	5	Published <sup>3</sup>
		3 <sup>rd</sup> Manuscript* :“Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage”	5	Published <sup>4</sup>
How can data from multiple sources be used to support disaster response?	To develop algorithms that are capable of integrating multi-source unimodal data.	4 <sup>th</sup> Manuscript: “Deep Learning-based Short Term Traffic Flow Prediction with Weather Data”	6	Ready to submit
		5 <sup>th</sup> Manuscript: “DEES - A real-time system for event extraction from disaster-related web text”	7	Accepted <sup>5</sup>
		6 <sup>th</sup> Manuscript#:“Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison”	7	Published <sup>6</sup>
How can the integration of multi-source multimodal data effectively support disaster response by cross-validating social media data?	To develop algorithms that are capable of integrating multi-source multimodal data.	7 <sup>th</sup> Manuscript: “Real-time information extraction from multi-source multimodal data for disaster response”	8	Ready to submit

\* This conference article provides details of traffic flow estimation that was used as the input for 4<sup>th</sup> paper.

# This conference paper provides the methodology for disaster-related tweet classification that is required for 5<sup>th</sup> manuscript.

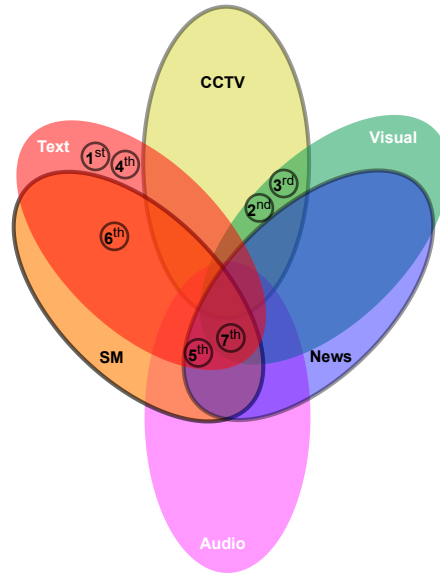
<sup>3</sup>Conference: ISCRAM 2020

<sup>4</sup>Conference: ISCRAM 2021

<sup>5</sup>Journal: Social Network Analysis and Mining 2022

<sup>6</sup>Conference: ISCRAM 2021

Figure 1.1 illustrates how manuscripts address the use of different data sources and modalities.



**Figure 1.1** The use of different data modalities in manuscripts. Manuscript numbers are as listed in Table 1.1.

## 1.4 The Structure of the Thesis

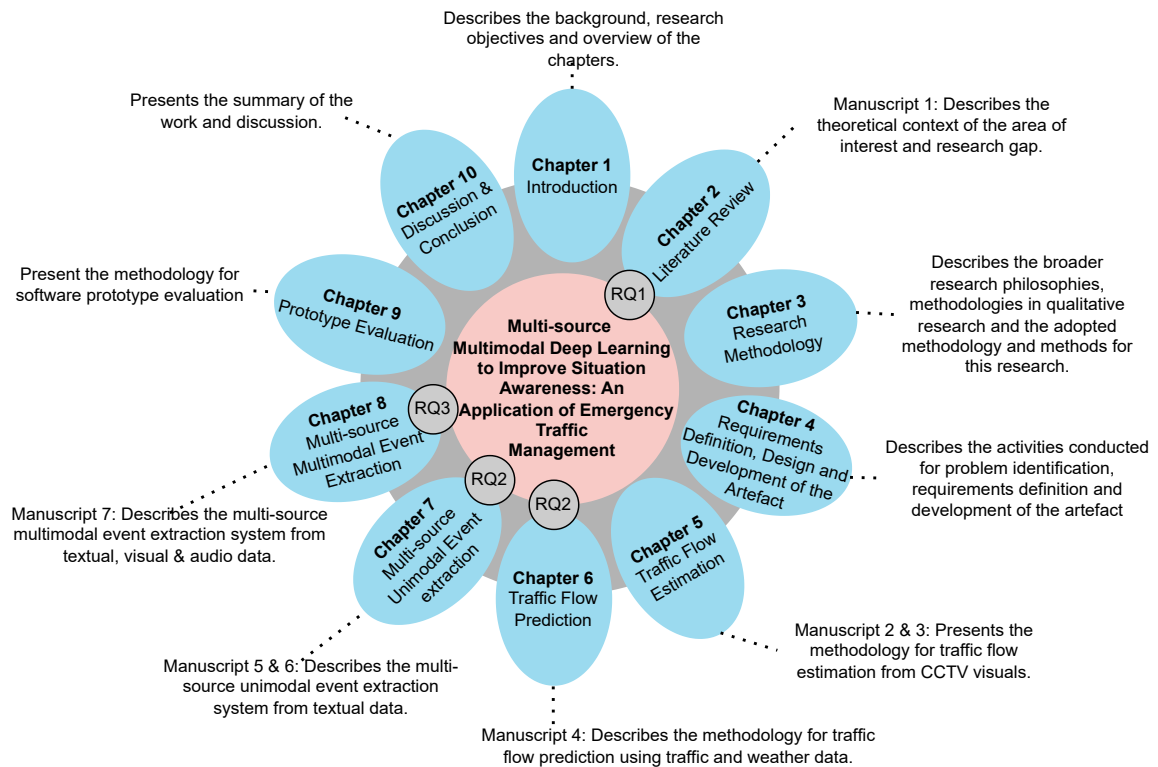
This thesis includes ten chapters. The first, fourth, and seventh manuscripts are individual chapters. Both manuscripts 2 and 3 are in a single chapter (Chapter 5). Manuscripts 5 and 6 are included in Chapter 7. Additionally, the thesis contains an introductory chapter, a research philosophy and methodology chapter, a chapter describing the adaptation of the first three activities of the design science method framework, a prototype evaluation chapter, and an overall discussion chapter. Figure 1.2 illustrates the thesis chapter outline.

### Chapter 2; Manuscript 01

In Chapter 2 a review of the literature on DL in disaster research is presented. The published paper answers the first research question, “How have different deep learning algorithms been applied to data from various sources to support disaster response tasks?”. The chapter presents a systematic review of 83 articles to identify the successes, challenges, and opportunities associated with the use of DL algorithms for disaster data processing. As a result of the analysis, key areas that have received little attention have been identified, and a flowchart outlining recommendations for future research are developed to maximize the benefits of DL for disaster response tasks. Moreover, the chapter identifies the gap in the literature that is the focus of this study.

### Chapter 3

Chapter 3 provides the rationale for the research framework and justifies the selection of Design Science Research as the appropriate approach to conduct the research. Furthermore, it introduces the five-stage method framework for design science research developed by Johannesson et al. [175]. The stages include Explicate Problem, Define Requirements, Design and Develop Artefact, Demonstrate Artefact and Evaluate Artefact. The chapter also covers the research philosophy and the subsequent methods used in the entire research process.



**Figure 1.2** Thesis chapter outline.

## Chapter 4

Chapter 4 describes how the first three activities of Johannesson's design science method framework have been adopted for this doctoral research. The activities include Explicate Problem, Define Requirements and Design of the artefact.

The development of the artefact is discussed in Chapter 5 to 8.

## Chapter 5 Manuscripts 02 and 03

The algorithms for counting traffic flow from CCTV images and footage are described in Chapter 5. The chapter includes two articles published in peer-reviewed conferences. Traffic flow data is the primary source of data for the design and development of the artefact. There were no traffic datasets collected in New Zealand that could support the project at the time of the research. As a result, the researcher developed DL-based algorithms for obtaining traffic data detailed in the chapter.

## Chapter 6; Manuscript 04

In chapter 6, a DL-based algorithm is introduced to make short-term traffic flow predictions. The paper answers the second research question "How can data from multiple sources be used to support disaster response?". The chapter shows that the accuracy of DL-based traffic flow prediction improves when fusing weather data as an additional input. Experimental results are discussed in detail to show that the newly proposed DL architecture predicts the short-term traffic flow given weather data with greater accuracy than previous work.

## Chapter 7; Manuscripts 05 and 06

Chapter 7 consists of the algorithms proposed for multi-source unimodal event extraction for disaster response. This paper explores the answers to the second research question "How can data

from multiple sources be used to support disaster response?”. The article recognises SM data as a significant source of information for real-time disaster response, despite many obstructions such as fake news and rumours that prevent user-generated SM content from being used directly. As a result, the paper introduces a new algorithm for extracting disaster events from both online news and tweets text data. The chapter also includes a conference paper that describes the methodology for disaster-related tweet classification adapted for the event extraction algorithm.

### **Chapter 8; Manuscript 07**

The final research article that brings together the algorithms proposed for the multi-source multi-modal event extraction for disaster response is presented in Chapter 8. This paper finds the answers to Research Question Three “How can multi-source multimodal be used to effectively support disaster response by cross-validating SM data?”. The importance of exploiting SM data for situation awareness is emphasised, and the ways to improve the trustworthiness of such data are explored by fusing multi-source multimodal data. The article presents the architecture of Multi-Source Multi-modal Event Extraction System for Disaster Response (M<sub>2</sub>E<sub>2</sub>S for DR) that focuses on integrating text and images from SM and online news to extract answers for the *What (semantic)*, *Where (spacial)* and *When (temporal)* (3W) questions, as well as impact information.

### **Chapter 9**

Chapter 9 describes Demonstrate Artefact and Evaluate Artefact activities of Johannesson’s design science method framework. The methods used to do these activities are presented in detail, along with the results of the evaluation. The artefact was evaluated using two rounds of interviews with expert users within the design science research framework.

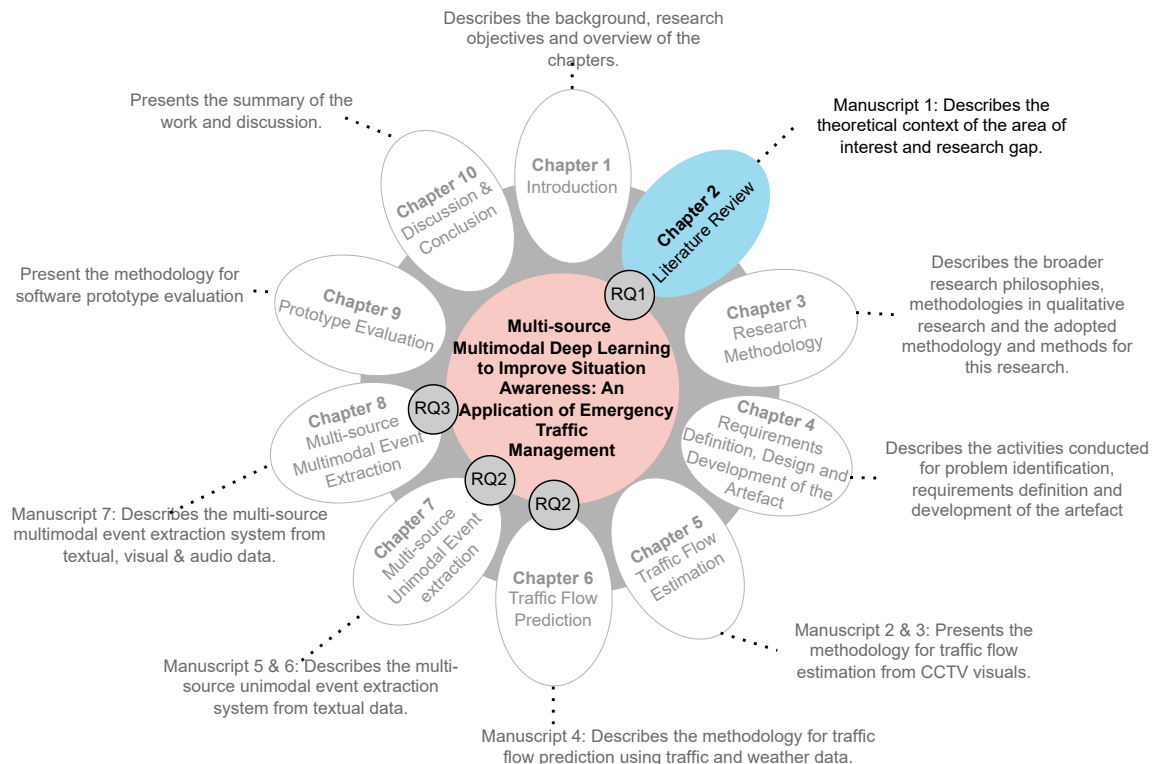
### **Chapter 10**

The final chapter summarises the main findings of the doctoral research and presents some limitations and opportunities for future work. The research questions and objectives are revisited, and the significant contributions to the research and practice are highlighted.

The following chapter of the thesis, Chapter 2, contains the literature review that determined the research gap for this doctoral research.

## Chapter 2

# Multi-source Multimodal Big Data and Deep Learning for Disaster Research: A Systematic Review



This chapter presents the first manuscript containing a systematic literature review of DL and DR research. This review seeks the answer to the first research question “How have different deep learning algorithms been applied to data from various sources to support disaster response tasks?”.

The article presented in this chapter was published in SN Computer Science. The paper was submitted on 28<sup>th</sup> July 2021 and after revisions, was published on 27<sup>th</sup> November 2021. The article is published as: Algiriyage, N., Prasanna, R., Stock, K. et al. Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review. SN COMPUT. SCI. 3, 92 (2022). <https://doi.org/10.1007/s42979-021-00971-4>

## Abstract

Mechanisms for sharing information in a disaster situation have drastically changed due to new technological innovations throughout the world. The use of social media applications and collaborative technologies for information sharing have become increasingly popular. With these advancements, the amount of data collected increases daily in different modalities, such as text, audio, video, and images. However, to date, practical Disaster Response (DR) activities are mostly depended on textual information, such as situation reports and email content, and the benefit of other media is often not realised. Deep Learning (DL) algorithms have recently demonstrated promising results in extracting knowledge from multiple modalities of data, but the use of DL approaches for DR tasks has thus far mostly been pursued in an academic context. This paper conducts a systematic review of 83 articles to identify the successes, current and future challenges, and opportunities in using DL for DR tasks. Our analysis is centred around the components of learning, a set of aspects that govern the application of Machine Learning (ML) for a given problem domain. A flowchart and guidance for future research are developed as an outcome of the analysis to ensure the benefits of DL for DR activities are utilized.

## 2.1 Introduction

Disasters, whether natural or human-induced, often result in loss of lives, property, or damage that can impose a significant impact on communities over a long period. With the proliferation of smart mobile devices, people are now increasingly using social media applications during disasters to share updates, check on loved ones, or inform authorities of issues that need to be addressed (e.g., damaged infrastructure, stranded livestock). Besides physical sensors and many other sources; human sensors, such as people who use smart mobile devices, generate massive amounts of data in different modalities (such as text, audio, video, and images) during a crisis. Such datasets are generally characterised as *multimodal* [40].

Disaster Response (DR) tasks bring together groups of officials who often serve different organizations and represent different positions, and their information requirements remain complex, dynamic, and ad hoc [308]. Also, it is beyond the capacity of the individual human brain to combine different forms of data in real-time and process them to form meaningful information in a complex and fast-moving situation [309]. Therefore, the main challenge faced by emergency responders is effectively extracting, analyzing, and interpreting the enormous range of multimodal data that is available from different sources within a short time period. As a result, emergency responders still depend mostly on text-based reports prepared by field officers for their decision-making processes, avoiding many other sources that could provide them with useful information.

Previously, the DR research community applied classical Machine Learning (ML) techniques to automate DR activities [291, 4]. However, the requirement of ML algorithms for handcrafted features prevented the timely use of such models. Furthermore, the research processes with these methods were labour intensive and time-consuming [267]. More recently, Deep Neural Networks, which rely less on handcrafted features, instead learning directly from input data, have been used extensively to learn high-level representations through deep features and have proven to be highly effective in many application areas such as speech recognition, image captioning, and emotion recognition [40, 363, 33, 221]. As DL techniques gain popularity among researchers, there is a

timely need to discuss the potential for their use for DR activities. Researchers and practitioners need to understand what has been done in the literature and the current knowledge gaps to make further improvements. Thus, this article analyses and systematically reviews the intersection of the two research fields (DL for DR).

We have organized our review around the *components of learning* as proposed by Abu-Mostafa [366] and used by Watson et al. [397] for their systematic review. Abu-Mostafa [366] demonstrated the application of five components of learning for any ML problem. These components provide a clear mapping to establish a roadmap for investigating DL approaches in DR research. Our objective is to identify application scenarios, best practices and future research directions in using DL to support DR activities. Therefore, we synthesize five main research questions (RQs) and eight sub-questions that support the main RQs according to the components of learning. To answer the RQs, we create a data extraction form having 15 attributes such as DR Task, Data Type, Data Source, and DL Architecture. We create a taxonomy of DR tasks in response to the first RQ, which is then utilized to derive answers for the next RQs. Finally, we use the Knowledge Discovery in Databases (KDD) process to uncover hidden relationships among extracted values for the attributes in the extraction form. Based on our findings, we propose a flowchart with guidelines for DR researchers to follow when using DL models in future research.

We found multiple review articles that discussed the use of multimodal data for disaster response (for example, [12, 313]), outlining applications and challenges. However, many of these have not explicitly considered using DL for feature extraction. We also observed other review articles focused on individual data sources. For example, the studies [235, 279, 392, 335, 22, 198] addressed the frameworks, methodologies, technologies, future trends, and applications for disaster response while using social media datasets. Among other reviews, Gomez et al. [117] analyzed remotely sensed UAV data, considering cases of different disaster scenarios. Overall, these reviews are especially focused on addressing a single source of data and how it can be used for disaster response. The more recent article by Sun et al. [359] provides an overview of using Artificial Intelligence (AI) methods for disaster management. Our work significantly differs from the work by Sun et al. in a number of ways. Firstly, we analyze the articles systematically, adopting the learning components as proposed by Abu Moftha [366]. Secondly, our analysis is confined to trending DL techniques as a subset of AI. Thirdly, we provide a wider discussion on the datasets, preprocessing, DL architectures, hyperparameter tuning, challenges and solutions in processing data for the DL task, and clarify future research directions.

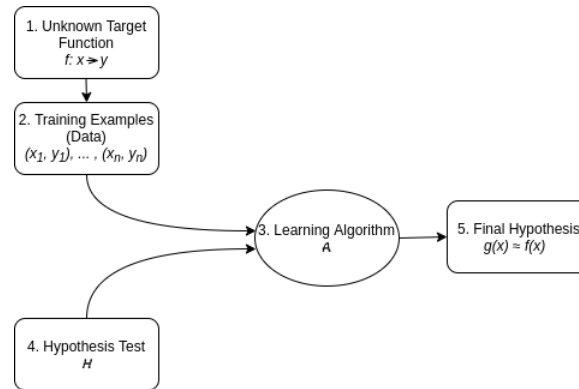
The remainder of this article is organized as follows. We first provide a synthesis of the research questions in Section 2.2. Section 2.3 outlines the methodology used to analyze the literature. Sections 2.4 to Section 2.8 provide the analysis of the research questions and Section 2.9 summarises opportunities and future research challenges. Section 2.10 discusses the relationships extracted during the KDD process. In Section 2.11 a flow chart is provided with recommendations for future research. Finally, in Section 2.12 we broadly discuss research gaps and conclusions. An online appendix contains the full details of the analysis process, as well as the resources [23].

## 2.2 Research Question Synthesis

Our overarching objectives during this study are to identify research challenges and best practices, and provide directions for future research while using DL methods for DR tasks. Therefore,



we have centralized our analysis around the elements of learning (see Fig. 2.1) and formulated the main RQs accordingly. As a result, we ensure that our analysis effectively captures the essential components of DL applications while also allowing us to perform a descriptive content analysis across these components. Furthermore, we formulated sub-questions supporting the main RQs to analyze more details. The next subsections discuss the formulation of the main RQs and sub-questions according to the components of learning.



**Figure 2.1** The components of learning as proposed by Abu Moftha [366]

### 2.2.1 The First Component of Learning: The Target Function

The first component of the learning problem is an “unknown target function ( $f : x \rightarrow y$ )” as illustrated in Fig. 2.1, which represents the relationship between known input ( $x$ ) and output ( $y$ ). The Target Function is the optimal function that we are attempting to approximate with our learning model. Therefore, the first component of learning enables the researcher to identify main application areas in the research field. As a result, we formulated our first research question to identify target functions in the DR domain, as follows:

**RQ<sub>1</sub>**: What types of DR problems have been addressed by DL approaches?

**RQ<sub>1</sub>** aims to discover DR tasks that have been investigated previously using DL methodologies. Furthermore, the answers to our first RQ provide a taxonomy for analyzing the next research questions.

### 2.2.2 The Second Component of Learning: The Training Data

The second component of learning is the historical data (training data), required by the algorithm to learn the unknown target function. A thorough understanding of the training data leads to insights about the target function, possible features, and DL architecture design. Furthermore, the quality of the output of a DL model is directly coupled with the provided training data. Therefore, our second question is formulated to understand training data.

**RQ<sub>2</sub>**: How have the training datasets been extracted, preprocessed and used in DL-based approaches for DR tasks?

Our goal during this question is to capture the types of training data, the extraction sources, and the preprocessing techniques applied to prepare them for the DL tasks. To support and allow a deeper understanding of the main RQ, we examine this through three sub-questions.

- $RQ_{2.1}$  What types of DR data have been used?
- $RQ_{2.2}$  What sources have been used to extract data, and how have data been extracted?
- $RQ_{2.3}$  How have data been preprocessed before applying the DL models?

The answers we extract during questions  $RQ_{2.1}$  and  $RQ_{2.2}$  will enable future researchers to see what types and sources of data have been explored in previous studies and what data has not yet been investigated. Furthermore,  $RQ_{2.3}$  provides them with the details of preprocessing techniques that have been followed during the studies.

### 2.2.3 The Third and Fourth Components of Learning: The Learning Algorithm and Hypothesis Test

According to Abu Moftha [366], the third and fourth learning elements are known as the “learning model”. The learning model consists of the learning algorithm and the hypothesis set. A learning algorithm tries to define a model to fit a given dataset. For example, the algorithm generally uses a probability distribution over the input data to approximate the optimal hypothesis from the hypothesis set. The hypothesis set consists of all the hypotheses to which the input data are mapped. Therefore, the learning algorithm and the hypothesis set are tightly coupled. Considering together the learning algorithm and hypothesis set, we formulate our third RQ as follows.

**RQ<sub>3</sub>:** What DL models are used to support DR tasks?

We aim to identify and evaluate the various DL models that have been applied for DR tasks. Hence, we consider three further sub-questions to capture specific architectures and types of DL models.

- $RQ_{3.1}$  What types of DL architectures are used?
- $RQ_{3.2}$  What types of learning algorithms and training processes are used?
- $RQ_{3.3}$  What methods are used to avoid overfitting and underfitting?

The answers to  $RQ_{3.1}$  provide DL architectures that has been adopted for various DR tasks. Our goal is to determine whether certain DL architectures are preferred by researchers and the reasons for those trends. As a part of the analysis, we capture how transfer learning approaches have been adopted to address algorithm training and performance issues. During  $RQ_{3.2}$  we intend to examine the types of learning algorithms and the training processes involved, including how parameter optimization has been achieved. Moreover, in  $RQ_{3.3}$ , we aim to analyze the methods used to combat overfitting and underfitting. Answers to both  $RQ_{3.2}$  and  $RQ_{3.3}$  will provide future researchers with an idea of how parameter tuning and optimization has been applied in DL for DR research to improve the accuracy of the output.

### 2.2.4 The Fifth Component of Learning: The Final Hypothesis

The final component of learning is the “final hypothesis”. This is the target function learnt by the algorithm to predict unseen data points. Through this component of learning, we aim to analyze the effectiveness of the algorithm at achieving the hypothesis for the selected DR task. Therefore, our fourth RQ is formulated as follows:

**RQ<sub>4</sub>**: How well do DL approaches perform in supporting various DR tasks?

During the analysis for **RQ<sub>4</sub>**, we derive the metrics used to evaluate the performance of DL models. Future researchers can utilize these matrices and extract values to compare the results achieved by their models. Additionally, we examine two sub-questions to perform a deeper evaluation of the selected question.

- *RQ<sub>4.1</sub>* What evaluation matrices are used to evaluate the performance of DL models?
- *RQ<sub>4.2</sub>* What “baseline” models have been compared?

Our intention with *RQ<sub>4.1</sub>* is to derive a taxonomy of performance matrices used by the analyzed studies, while *RQ<sub>4.2</sub>* will identify those “baseline” models that have been criticized and allow future researchers to select those appropriate for comparison of their results.

### 2.2.5 The Final Analysis

Our fifth RQ is designed to identify and characterize underlying problems that arise when utilizing DL models for DR tasks. Our goal is to provide researchers with challenges faced by the DR research community in employing DL-based approaches. This will enable future research to be designed in a way that addresses or avoids these challenges and better utilizes DL algorithms to support DR tasks. Furthermore, we aim to analyze the replicability of DL models and architectures. Researchers are more likely to re-implement, improve, or compare new models if the existing DL architectures are easily replicable, which will eventually increase the quality and quantity of DL for DR research. Thus, our final RQ is formulated as follows:

**RQ<sub>5</sub>**: What are the underlying challenges and the replicability of DL for DR studies?

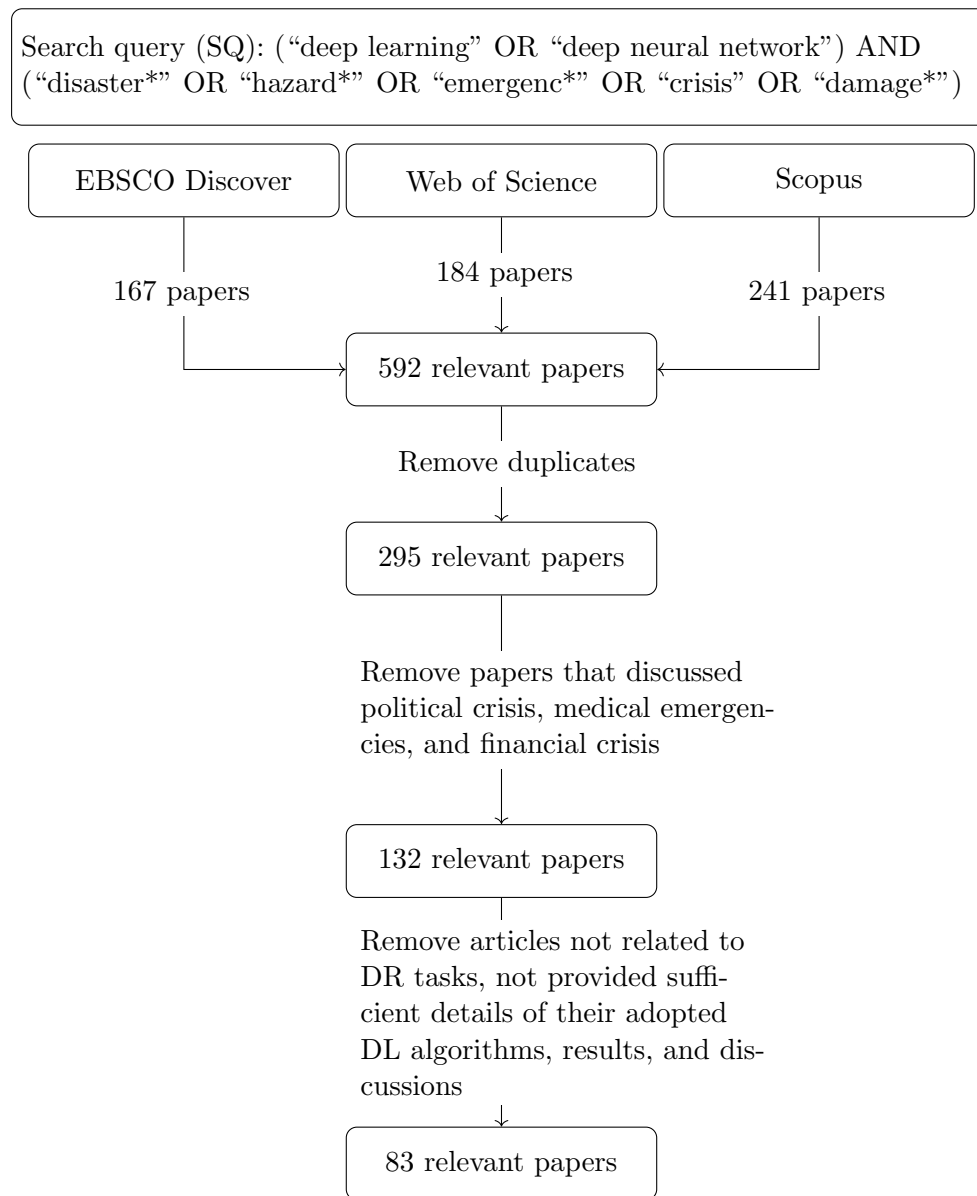
In summary, the systematic literature review (SLR) conducted in this paper answers the following research questions:

- **RQ<sub>1</sub>:** What types of DR problems have been addressed by DL approaches?
- **RQ<sub>2</sub>:** How have the training datasets been extracted, preprocessed, and used in DL-based approaches for DR tasks?
  - *RQ<sub>2.1</sub> What types of DR data have been used?*
  - *RQ<sub>2.2</sub> What sources have been used to extract data, and how have data been extracted?*
  - *RQ<sub>2.3</sub> How have data been preprocessed before applying the DL models?*
- **RQ<sub>3</sub>:** What DL models are used to support DR tasks?
  - *RQ<sub>3.1</sub> What types of DL architectures are used?*
  - *RQ<sub>3.2</sub> What types of learning algorithms and training processes are used?*
  - *RQ<sub>3.3</sub> What methods are used to avoid overfitting and underfitting?*
- **RQ<sub>4</sub>:** How well do DL approaches perform in supporting various DR tasks?
  - *RQ<sub>4.1</sub> What evaluation matrices are used to evaluate the performance of DL models?*
  - *RQ<sub>4.2</sub> What “baseline” models have been compared?*
- **RQ<sub>5</sub>:** What are the underlying challenges and the replicability of DL for DR studies?

## 2.3 Methodology

Multiple techniques have been proposed to understand the content of a body of scholarly literature, including scoping reviews, umbrella reviews, or systematic reviews [119]. Among them, the systematic review aims to exhaustively and comprehensively search for research evidence on a topic area and appraise and synthesize it thoroughly [119]. In this analysis, we are interested in identifying the gaps in the research and whether there are opportunities for researchers and practitioners to investigate new problems that have not yet been addressed in the DR domain using DL. We, therefore, consider a systematic review to be the most appropriate approach to find answers to the above formulated RQs. To the best of our knowledge, this is the first systematic review that investigates the intersection of the DL and DR research fields. Our study adopts the following steps to guide the SLR process, as highlighted by Yigitcanlar et al. [408].

1. Develop a research plan;
2. Search for relevant articles;
3. Apply exclusion criteria;
4. Extract relevant data from the selected articles;
5. Analyse the literature data;



**Figure 2.2** Literature selection process.

### 2.3.1 Develop a research plan

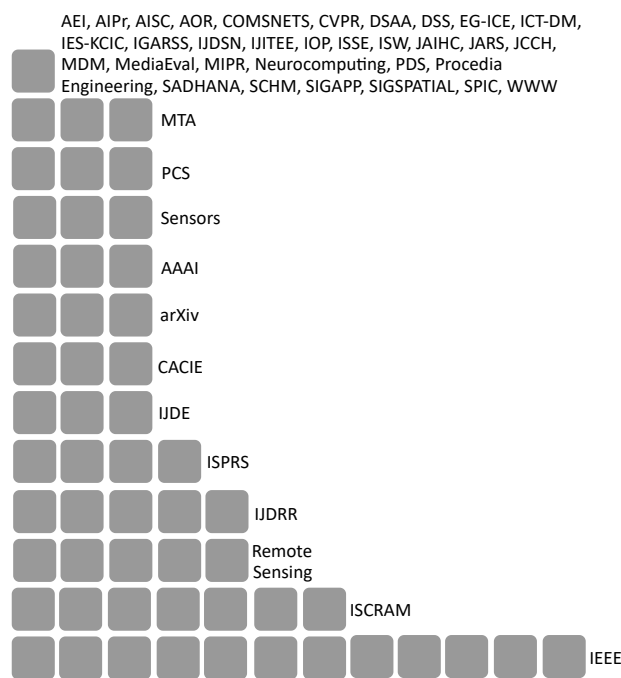
As the first step for carrying out the SLR, a research plan was developed, including research aim, keywords, and a set of inclusion and exclusion criteria. The research aim was to identify the usage of DL techniques on disaster data to support DR tasks as outlined in RQs 1 to 5. Hence, “disaster” and “deep learning” were selected as the search keywords. The search also included variants of these keywords. The alternate search terms for “disaster” included ‘hazard’, ‘emergency’, ‘crisis’, and ‘damage’. Also, ‘deep neural network’ was used as an alternative keyword for DL. Some research has considered “machine learning” as an alternative keyword for DL. However, since we were particularly interested in Deep Neural Networks, we omitted “machine learning” as a keyword in the search. The inclusion criteria limited the sources to peer-reviewed academic publications available online in a full-text format and relevant to the research aims. The exclusion criteria were determined as publications in languages other than English; grey literature, such as government or industry reports; and non-academic research.

### 2.3.2 Search the relevant articles

In the second step, the search for relevant articles was conducted using a keyword search in each of the following databases: Scopus, Web of Science, and the EBSCO Discovery Service on April 2, 2021. Articles published since April 2011 were considered because a scan of existing literature suggested that there was not much literature related to DL in disaster research before then. The initial search produced 592 results.

### 2.3.3 Apply exclusion criteria

In this step, the results were filtered to remove duplicates between the databases, which reduced the number to 295 unique articles. We used a simple Python script to remove duplicates using the title of the article. We confined our scope to only papers that discuss natural or human-induced disasters. Therefore, the abstracts were manually read and removed if they discussed political crises, medical emergencies or financial crises. We also removed articles that did not provide sufficient details related to the attributes in our extraction form (see Table 2.1). Finally, 83 articles were selected for the review. Fig. 2.2 illustrates the process and the steps that we followed to filter the results and the quantity of papers returned at each step. Moreover, we provide the publication venues of the 83 articles in Fig 2.3.



**Figure 2.3** Publication venues of the articles. The number of grey boxes corresponds to the number of articles published in each publication venue. Full publication venue names are available in our online Appendix [23]

### 2.3.4 Extract relevant data from the selected articles

The next step in our methodology was to extract relevant data from the selected articles. We developed a data extraction form including the information shown in Table 2.1. The extracted information was collected manually and added to a Google sheet and later downloaded as a tab-separated (.tsv) file for the data analysis steps. The extracted data sheet is available in the online appendix [23].

**Table 2.1** Attributes in the data extraction form

Article Published Year	Venue	DR Task Addressed
Input Data Modality	Data Source	Data Extraction Technique
Data Preprocessing Technique	Size of the Dataset	Type of Learning
DL Architecture used	Learning Algorithm	Evaluation Metrics
Replicability	Baseline	Combating Overfitting and Underfitting

### 2.3.5 Analyse data using the Knowledge Discovery in Databases (KDD) process

The final step in our SLR methodology was to analyze the extracted data. We used the steps discussed in [397], namely data collection, initial coding and focused coding. After the coding process, we used the Knowledge Discovery in Databases (KDD) process to understand relationships among attributes in the extraction form. The KDD process is used to extract knowledge from databases using five steps: selection, preprocessing, transformation, data mining, and interpretation/evaluation [103]. We combined data preprocessing and transformation into one step as both steps involve preparing data for the mining step. The details of each stage are listed as follows.

- **Selection:** This stage is related to the selection of relevant data for the analysis. As described in the previous section, we selected 83 articles and extracted 15 attributes from them for the analysis.
- **Preprocessing:** In this stage, we cleaned the extracted values by removing noise such as misspellings, incorrect punctuation and mismatching coding. We noticed that a number of variations on particular terms, and standardized these to ensure appropriate matching (e.g., ConvNet/CNN, F-measure/F1-value/F-score/F1-score).
- **Data mining:** The third stage is related to identifying relationships among extracted data. We applied Association Rule Mining to derive relationships discussed further in Section 2.3.5.1.
- **Interpretation/Evaluation:** We interpret the findings of the KDD process in Section 2.10. These relationships demonstrate actionable knowledge for future researchers from the 83 articles analyzed through the SLR process.

#### 2.3.5.1 Association Rule Mining

We followed the association rule mining process introduced by Samia et al. [208] for literature analysis. Our association rules are extracted using the Apriori algorithm. Association rules help to discover relationships in categorical datasets. For instance, the rules generated during the process identify frequent patterns in the dataset. Associations are generally represented by “Support”, “Confidence”, and “Lift”. We illustrate this using the values in the *Data Source* column in the extraction form. “Support” and “Confidence” are the two indicators evaluating the interestingness of a given rule.  $supp(Twitter)$  is the fraction of articles for which **Twitter** appears in the *Data Source* column of the extraction form as given in Equation 2.1.

$$supp(Twitter) = \frac{\text{Number of Articles in which } Twitter \text{ appears in the Data Source column}}{\text{Total Number of Articles}} \quad (2.1)$$

If we consider the values in both the *Data Source* and the *Data Type* columns of the extraction form, the association rule  $Twitter \rightarrow Text$  means that each time **Twitter** appears in the *Data Source* column, **Text** appears in the *Data Type* column (see Equation 2.2).

$$conf(Twitter \rightarrow Text) = \frac{supp(Twitter \cup Text)}{supp(Twitter)} \quad (2.2)$$

“Lift” measures how likely it is that item **Text** is found in the *Data Type* column when **Twitter** is found in the *Data Source* column as given in Equation 2.3. A “Lift” value greater than 1 means that item **Twitter** is likely to appear in the *Data Source* column if **Text** appears in the *Data Type* column, while a value less than 1 means that **Twitter** is unlikely to appear if **Text** appears in the respective columns.

$$lift(Twitter \rightarrow Text) = \frac{supp(Twitter \cup Text)}{supp(Twitter) \times supp(Text)} \quad (2.3)$$

These associations can provide a guidance for future researchers during the planning stages of a project applying DL to DR research, supporting them in choosing different attributes such as data source, deep learning algorithm and learning types. We used the Python apyori library<sup>1</sup> to discover association rules, details of which are presented in the online appendix [23].

## 2.4 RQ<sub>1</sub>: What types of DR problems have been addressed by DL approaches?

This RQ explores the types of DR problems that have been investigated with DL models. We derived a taxonomy of DR tasks to capture relationships between other learning components, as illustrated in Fig. 2.5. From the 83 papers that we analysed, we identified nine main DR tasks (level-1 of the taxonomy) that have been addressed using DL approaches. Fig. 2.4 shows the number of papers published in each year by the main DR tasks. During the ten-year duration of studies we analysed, unsurprisingly, little work was undertaken between 2011 and 2015. There was a sudden interest in exploring DL architectures in the DR domain from 2017 onwards. This interest coincides with the introduction of popular DL frameworks such as Keras<sup>2</sup> and TensorFlow<sup>3</sup> in 2015 and PyTorch<sup>4</sup> in 2016. *Disaster event detection* was the first task to be explored using DL algorithms. Among the other tasks, *Disaster damage assessment*, *Disaster related information filtering* and *Disaster related information classification* were explored in 2017. Remotely sensed images were the main source of data for multiple early studies that used DL approaches. Early research may have used remotely sensed data for various reasons. Firstly, in 2011, Google Earth<sup>5</sup> launched a platform that allowed researchers to download massive volumes of satellite imagery. This inspired researchers to investigate remotely sensed data for DR tasks. Furthermore, researchers were also able to successfully employ DL approaches since these images were available in larger quantities. Secondly, the advancement of computer vision techniques, such as DL structures pre-trained on huge datasets, made visual data processing easier.

The number of studies combining DL and DR tasks rapidly increased from 2017 to 2018, more

<sup>1</sup>Python Apriori algorithm implementation v1.1.2, <https://pypi.org/project/apyori/>

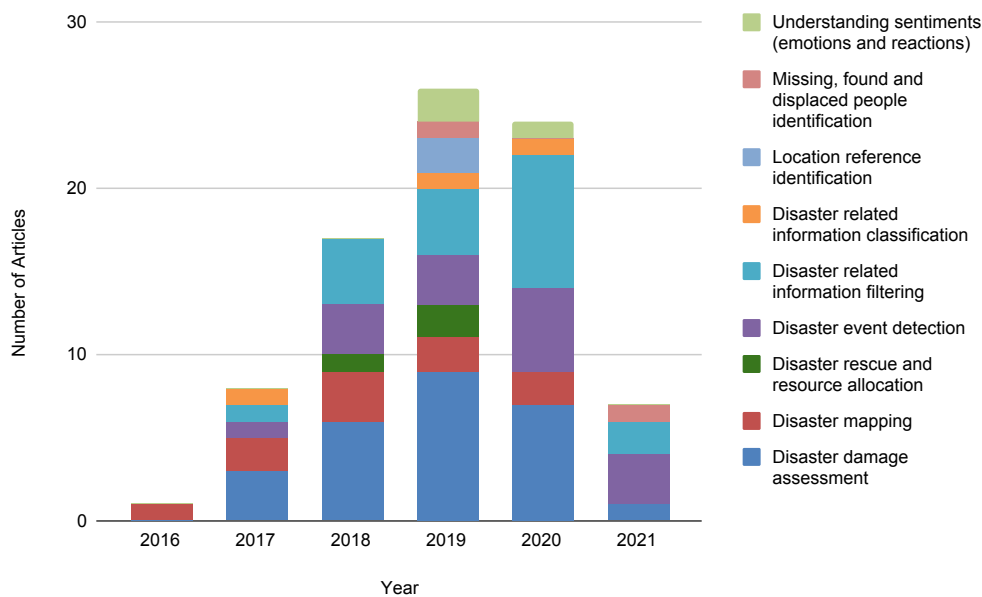
<sup>2</sup>Keras, <https://keras.io/>

<sup>3</sup>TensorFlow, <https://www.tensorflow.org/>

<sup>4</sup>PyTorch, <https://pytorch.org/>

<sup>5</sup>Google Earth, <https://earth.google.com/web/>





**Figure 2.4** Papers published per year according to DR task

than doubling. Furthermore, researchers extended their interest to explore multiple DR tasks over time, including *Disaster rescue and resource allocation*, *Location reference identification*, and *Understanding sentiments*. However, we see a slight drop in the number of articles published in 2020. This inconsistency may be due to the COVID-19 global pandemic and the physical and mental challenges that researchers encountered. We notice a significant amount of literature emerging during the first quarter of 2021, potentially representing a Covid-19 lag effect in publication.

*Disaster damage assessment* has been the most popular DR task analysed using DL approaches over the years, with 26 articles out of the 83 exploring this. There are three likely reasons for the popularity of *Disaster damage assessment*. Firstly, there is quite a strong driver and a clear need for damage assessment as it is urgently needed following an event, and there is little time for manual data collection. Secondly, the high availability of training datasets extracted from social media and remote sensing platforms were able to be used in supervised learning approaches. Thirdly, there is a clear mapping between training data and the target function (e.g., images of cracked buildings). This mapping helps researchers when designing DL-based applications to extract effective features. We observed an increasing interest in *Disaster related information filtering* and *Disaster related information classification* tasks. These DR tasks are mainly based on text datasets extracted from Twitter. A possible explanation for this trend could be the increased popularity of using Twitter as a communication channel during disasters. Moreover, the advancement of Natural Language Processing (NLP) techniques with the increased availability of annotated data corpora aids further developments in the information filtering and classification tasks.

DR tasks such as *Missing, found and displaced people identification* and *Location reference identification* had received less attention from researchers, resulting in a total of 4 articles out of the 83 reviewed. The lack of availability of large-scale training datasets and annotated data to train supervised learning approaches could be the main reasons for the reduced popularity of these DR tasks. We summarise the papers addressing each of the main DR tasks in Table 2.2.

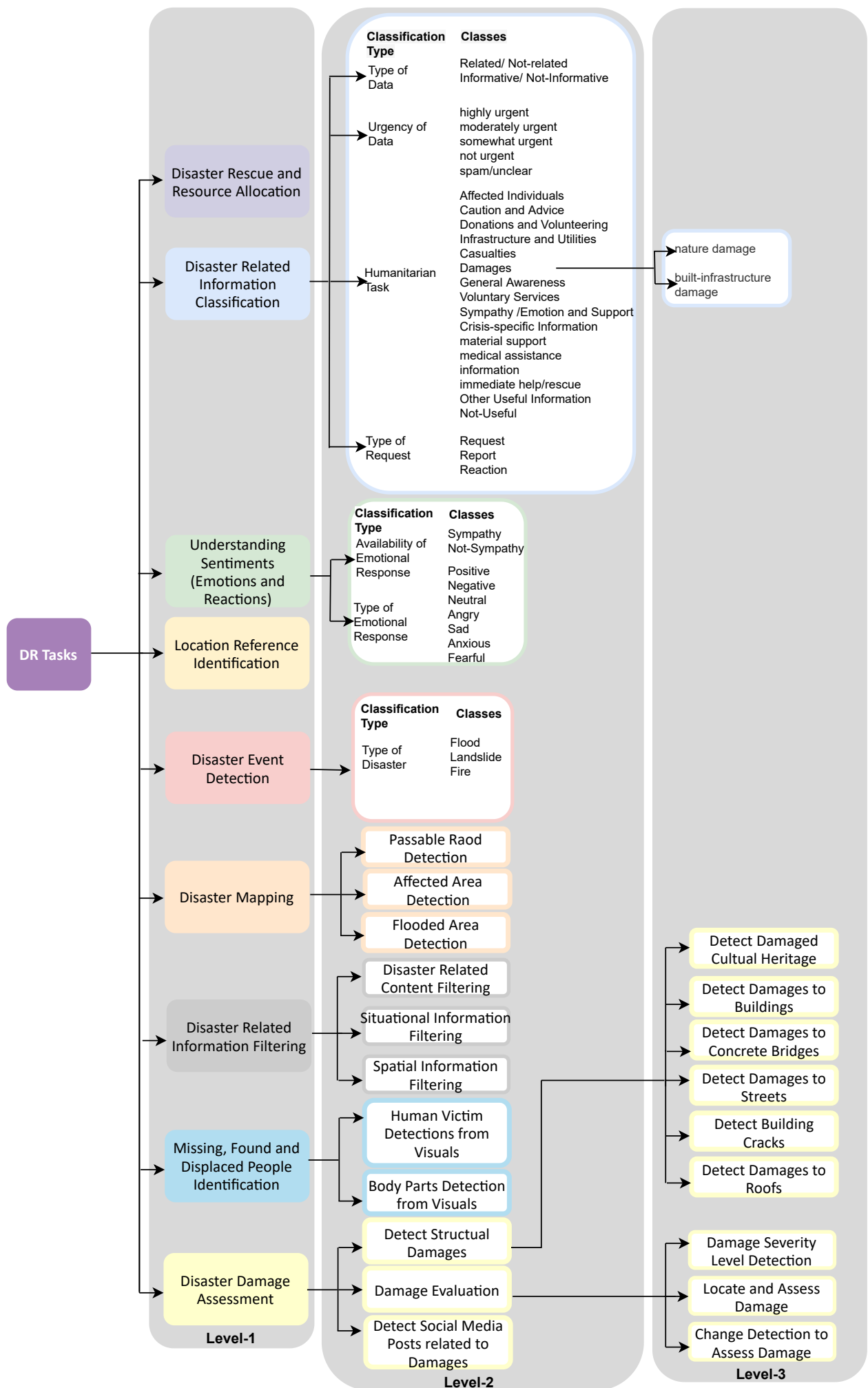


Figure 2.5 Taxonomy of DR tasks

**Table 2.2** Main DR tasks of the analysed articles

DR task	Articles
Disaster related information filtering	[14, 55, 109, 140, 152, 153, 206, 217, 242, 241, 237, 240, 259, 268, 287, 299, 298, 302, 350]
Disaster damage assessment	[7, 37, 63, 66, 68, 94, 112, 114, 132, 144, 176, 203, 219, 222, 218, 223, 244, 257, 261, 270, 272, 311, 341, 357, 381, 421]
Disaster event detection	[20, 32, 34, 189, 228, 229, 233, 262, 276, 418]
Location reference identification	[200, 342]
Missing, found and displaced people identification	[134, 258]
Disaster mapping	[8, 265, 316]
Disaster rescue and resource allocation	[44, 46, 85]
Understanding sentiments (emotions and reactions)	[220, 334, 406]
Disaster related information classification	[1, 10, 16, 56, 65, 179, 202, 199, 238, 306, 307, 326, 327, 367, 281]

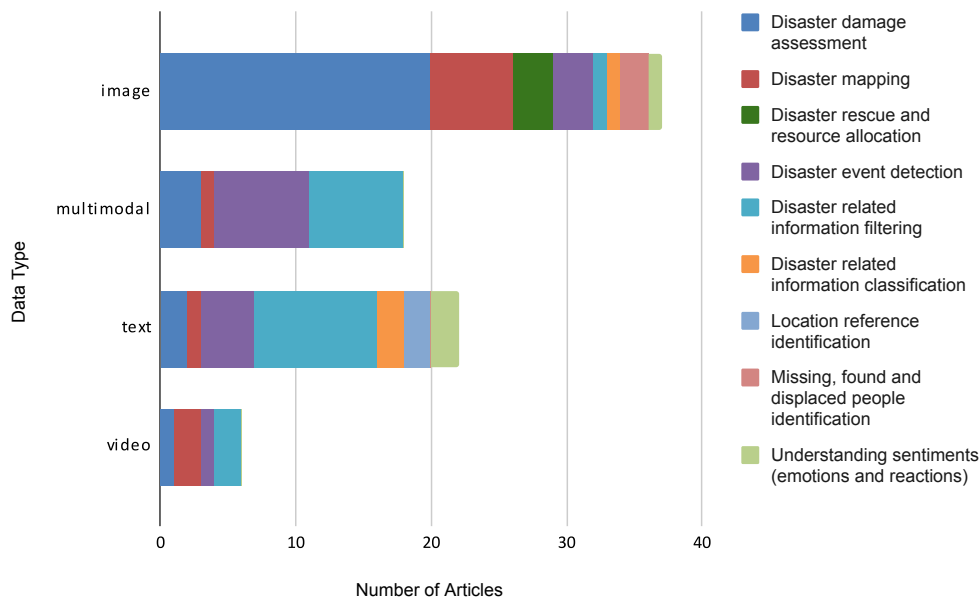
## 2.5 RQ<sub>2</sub>: How have the training datasets been extracted, preprocessed, and used in DL-based approaches for DR tasks?

For this research question, we analyze the types of disaster data that have been used by DL models to support disaster response. The accuracy and effectiveness of DL algorithms depend on the training dataset and its clarity. Therefore, we aim to understand the various types of disaster data used by DL approaches, the sources and methods employed to extract them, and the preprocessing steps. All of these points are important in understanding and designing DL approaches for DR tasks.

### 2.5.1 RQ<sub>2.1</sub> *What types of DR data have been used?*

Our analysis of the types of data that have been used for DR tasks using DL approaches reveals relationships between DR tasks and data types, illustrated in Fig. 2.6. Among the 83 articles analysed, 37 used images as the data source. Surprisingly, in practise, disaster responders rely significantly on textual data sources such as emails and field reports [132]. This finding indicates that these approaches have been mostly pursued in academic contexts. We assume multiple reasons contributing to the popularity of using image data for DR tasks: firstly, the power of visuals in conveying messages over textual content; secondly, the availability of pre-trained networks and the use of transfer learning techniques for image feature extraction and thirdly, easy accessibility of image datasets through web search and web databases. *Disaster damage assessment* is the most popular DR task among the studies that used image datasets.

Text data was used by 22 of the 83 articles and is more prominent in *Disaster related information filtering* and *Disaster related information classification* tasks. Currently available, annotated



**Figure 2.6** Data types used for DR task

disaster-related text data repositories (particularly using social media data) provide a clear guide for specific target problems. As a result, many researchers have used text data for supervised learning approaches in information filtering and classification applications.

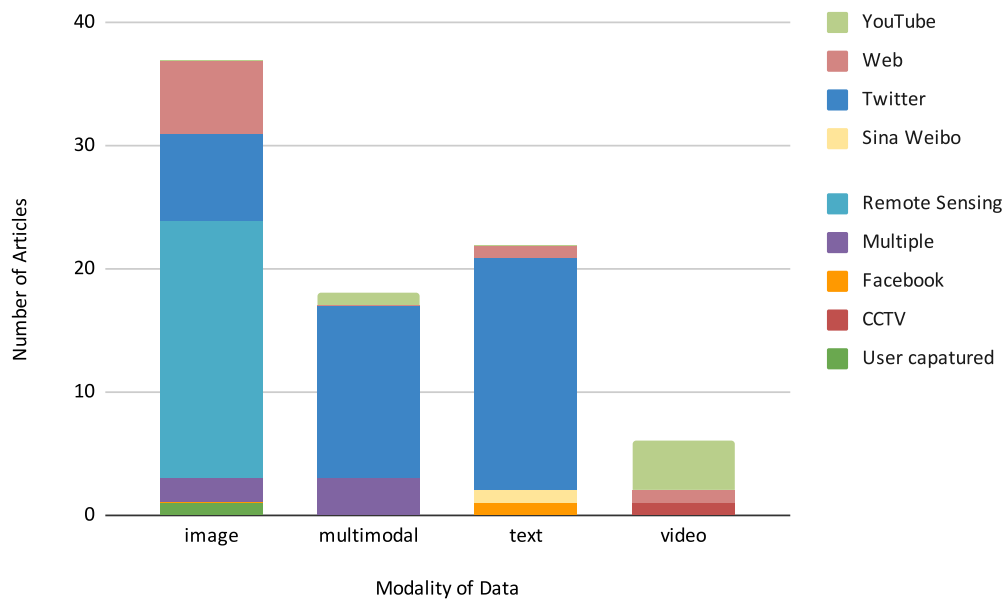
There has been little interest in using video datasets for DR tasks. Only 6 articles discussed the usage of video datasets for *Disaster related information filtering*, *Classification*, and *Disaster event detection* tasks. The possible reasons for this can be difficulties in storing and moving, and the need for special computing facilities for analysing video data such as Graphical Processing Units (GPUs).

We observed a significant interest in using multimodal data to extract information for DR tasks between 2018 and 2020. Multimodal data has been used for *Disaster related information filtering*, *Disaster related information classification*, *Disaster damage assessment* and *Disaster event detection* contributing to 18 articles in the analysed papers. We assume the popularity of multimodal DL networks depends on three reasons. Firstly, the combination of multiple modalities leads to more complementary information than learning from a single data modality. Secondly, multimodal learning helps to integrate data from different sources and provides access to large quantities of data. Thirdly, the more recent development of multimodal DL networks shows improved results over unimodal analysis.

### 2.5.2 RQ<sub>2.2</sub> *What sources have been used to extract data, and how have data been extracted?*

In this RQ, we analyse the sources (including accessible disaster data repositories) used to extract data used in DL models.

Image data has mainly been extracted using remote sensing from sources such as satellites, aerial vehicles and LiDAR. Apart from that, Twitter and the Web have been used by 7 and 6 articles respectively to extract image datasets (we grouped research that extracted data from websites and Google search under *Web*). Twitter has been the prominent source of text information, and was



**Figure 2.7** Sources used to extract data types

used for a total of 19 articles out of the total 83 (and out of the 22 articles that used text data) analysed. The growing number of human-annotated disaster-related Twitter data repositories is likely to have increased the amount of research using them with DL approaches. We observed that 5 articles used a combination of multiple sources to extract data, such as Twitter, web mining, Baidu, Flickr, Instagram, and Facebook. Most notably, Facebook was rarely used (1/83) as a source due to its data extraction limitations (e.g., the requirement of prior approval from Facebook to use public feed Application Programming Interface (API) [312]). Fig. 2.7 shows the sources used to extract different modalities of data.

Researchers have employed multiple techniques to extract data from different sources. Twitter data has been extracted through the Twitter Streaming API using general or specific keywords (e.g., earthquake, Nepal Earthquake) and a spatial bounding box covering the impacted area is often used while extracting tweets. However, it is notable that a total of 28 articles downloaded data from annotated Twitter repositories from previous research such as CrisisNLP<sup>6</sup> and CrisisLex<sup>7</sup>, indicating the importance of annotated data repositories catering for DR problems. Web mining and web databases were used in 22 articles to download data. Workshops and conferences, for example, MediaEval<sup>8</sup>, have provided researchers with annotated dataests and meta-data for target problems. Table 2.3 summarizes the different data collection methods.

**Table 2.3** Disaster data collection methods

Data Extraction Method	Articles
Artificial Intelligence for Disaster Response (AIDR)	[16, 272]
Baidu API	[406]

<sup>6</sup>CrisisNLP datasets, <https://crisisnlp.qcri.org/>

<sup>7</sup>CrisisLex datasets, <https://crisislex.org/>

<sup>8</sup>MediaEval datasets, <http://www.multimediaeval.org/>

Cameras mounted on satellite, airborne and UAV	[46, 316, 63, 381, 37, 94, 222, 270]
Copernicus EMS program	[176]
CrisisLex	[268, 55, 56, 276, 261]
CrisisMMD	[242, 202, 241, 109, 307, 217, 238, 1, 199, 7]
CrisisNLP	[10, 11, 220, 276, 261]
Facebook page crawling	[302]
Flicker API	[132]
GNIP (Social media data re-seller)	[342, 85]
Google Earth	[32, 228, 229, 265, 66]
LiDAR	[341]
Previous research	[237, 34, 218, 223, 421, 219]
Twitter API	[16, 179, 350, 334, 200, 287, 240, 140, 44, 327, 259, 132]
Web database	[326, 134, 206, 262, 189, 357, 114, 112, 144, 257, 65]
Web mining	[367, 306, 258, 153, 152, 14, 299, 203, 233, 272, 261, 68, 244, 219, 298, 418]
Workshop/Conference	[41, 8, 20, 311]

---

### 2.5.3 $RQ_{2.3}$ *How have data been preprocessed before applying the DL models?*

To address  $RQ_{2.3}$ , we derive a taxonomy of preprocessing steps that researchers have used to clean raw data for use in DL algorithms. Cleaning and transforming data to be used effectively by DL models are critical steps towards improved performance. However, 19 articles out of 83 analysed did not explicitly mention the preprocessing steps that were undertaken.

We observe three common preprocessing steps across the articles analyzed: filtering, annotation, and dataset splitting. Data filtering helps reduce noise in raw data. Annotation deals with labelling the data depending on the target function. A total of 10 of the 83 articles employed external annotators or hired them through annotation service providers such as Figure Eight<sup>9</sup> (formerly known as CrowdFlower). The annotated datasets are generally split into train, test, and validation sets during the preprocessing steps. The training data sets are used to train the DL model, while test datasets are used to provide unseen data to be classified by the model as a test. The validation set is used to tune hyperparameters of the DL model.

Our analysis identified that the design of the preprocessing steps largely depends upon the modality of data. For example, text data preprocessing steps included tokenizing, lowercasing, stemming, lemmatization and removal of stop words, tokens having less than 3 characters, sentences having less than 3 words, user mentions, punctuation, extra spaces, line breaks, emojis, emoticons, special characters, symbols, hashtags, numbers, and duplicates. Text normalization using the Out of Vocabulary (OOV) dictionary is used to replace slang, mistakenly added words, abbreviations, and misspellings. Image data preparation steps included data filtering, duplicate removal, patch generation, resizing, pixel value normalization, and image augmentations. Video data preprocessing included clipping to extract keyframes, shot boundary detection and removal of duplicates and

<sup>9</sup>Figure Eight external annotation service, <https://appen.com/>

blurred and noisy frames. Table 2.4 illustrates the preprocessing steps involved in preparing raw data for DL algorithms, as found in the analyzed articles.

**Table 2.4** Data preprocessing steps.

Modality	Preprocessing step	Description/ Example
Text	Tokenizing	Tokenization is the process of breaking sentences in to smaller chunks (e.g., words).
	Lowercasing	Lowercasing tweet text is used to merge similar words and reduce the dimensionality of the problem.
	Removal of Stop words	Stopwords are a set of frequently used words such as “the”, “in”, and “a” that are not required to analyse them for a analysis task.
	Removal of URLs and User mentions	Tweets generally consist user handlers and embedded URLs. During preprocessing they are removed or replaced with <USER >and <URL >respectively.
	Removal of Hashtags	Hashtags are words or phrases chosen by users to connect specific themes such as events and topics (e.g., #NepalEq).
	Removal of punctuation, whitespaces, linebreaks	Punctuations (e.g., “!@#” ;:”), whitespaces and linebreaks are removed as they do not contain valuable information for a analysis task.
	Removal of Numbers	Numerical values included in tweets are removed if they do not contain any information for the analysis task.
	Removal of words shorter than 3 characters	Shorter words such as “oh”, “omg” and “hmm” are not useful for the analysis task and therefore, removed.
	Replacing contractions	The user-generated Twitter posts mostly contain shorten phrases (e.g., <i>I’d, didn’t and I’ll’ve</i> ). During the contraction mapping, these words are mapped into their original format (e.g., <i>I would, did not, I will have</i> ).
	Stemming and lematization	Stemming and lemmatization are used to convert a word into its root format. The stemming process cuts off the ends of words without considering the context, while lemmatization considers the context (e.g., <i>felt to feel</i> ).
	Remove sentences having less than three words	Remove very short sentences.
Image	Manual filtering	Manually check images to remove unwanted.
	Patch generation	Select arbitrary shaped regions from an original image.
	Resizing	
	Pixel value normalization	Pixel values of an image normally are between 0-255. During the normalization, values are converted to be in a specified range such as [1-0].

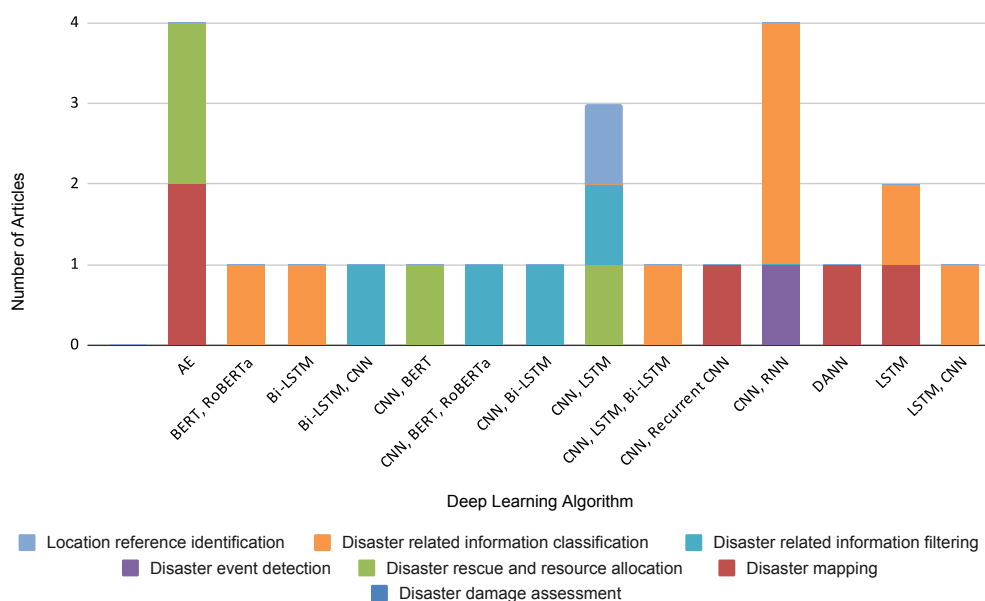
	Image transformation	(e.g., rotation, translation, rescaling, flipping, shearing, and stretching).
Video	Manual filtering	Manually check videos to remove unwanted.
	Shot boundary detection	A shot is an unbroken sequence of frames and a shot boundary is determined by the change of color histogram features.
	Clipping to extract key frames	Extract frames in the middle of each shot as key frames.
	Removal noisy frames	Remove duplicates and blurred frame

## 2.6 RQ<sub>3</sub>: What DL models are used to support DR tasks?

In this section, we analyze the types of DL architectures used for DR tasks and learning algorithms. Our aim is to identify the relationship between DR tasks and the DL architectures. We provide a short overview of different deep learning architectures in our online appendix [23].

### 2.6.1 RQ<sub>3.1</sub> *What types of DL architectures are used?*

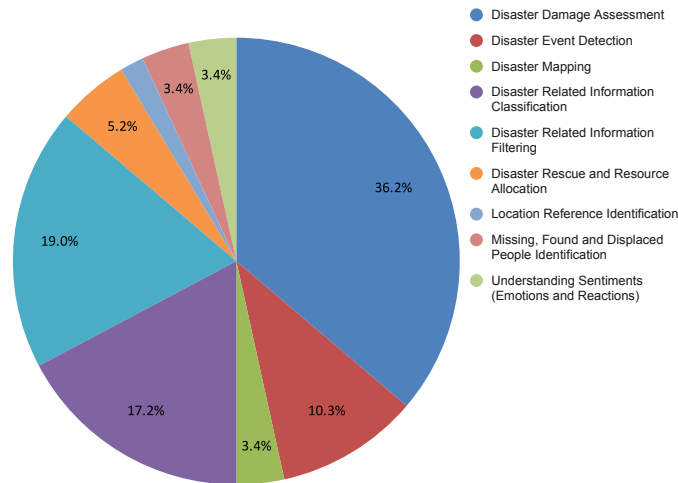
Through this question, we analyze types of DL architectures used to extract features for DR tasks. We observed that six main DL architectures had been used, namely Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory Networks (LSTMs) and its variant Bi-directional LSTMs (Bi-LSTMs), Domain Adversarial Neural Networks (DANNs), and AutoEncoders (AEs) across the studies we analyzed. Moreover, popular language models like Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimized BERT Pre-training Approach (RoBERTa) have been used for Natural Language Processing (NLP) tasks.



**Figure 2.8** DL architectures used by DR tasks except for CNN as a single architecture

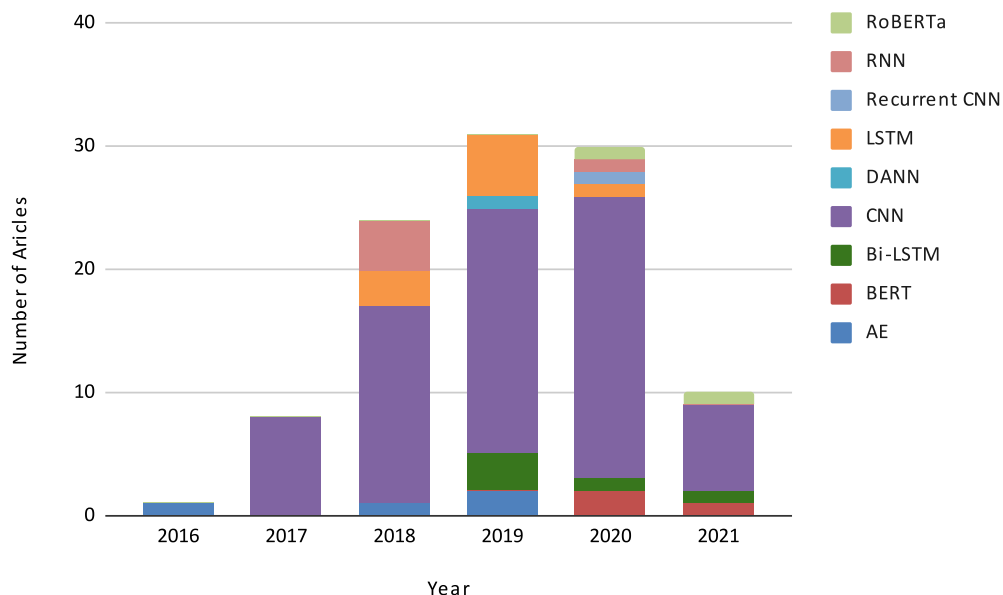
Fig. 2.8 shows the usage of DL algorithms according to the DR tasks excluding CNNs. We demonstrate the application of the CNN algorithm for DR tasks in a separate diagram (see Fig.





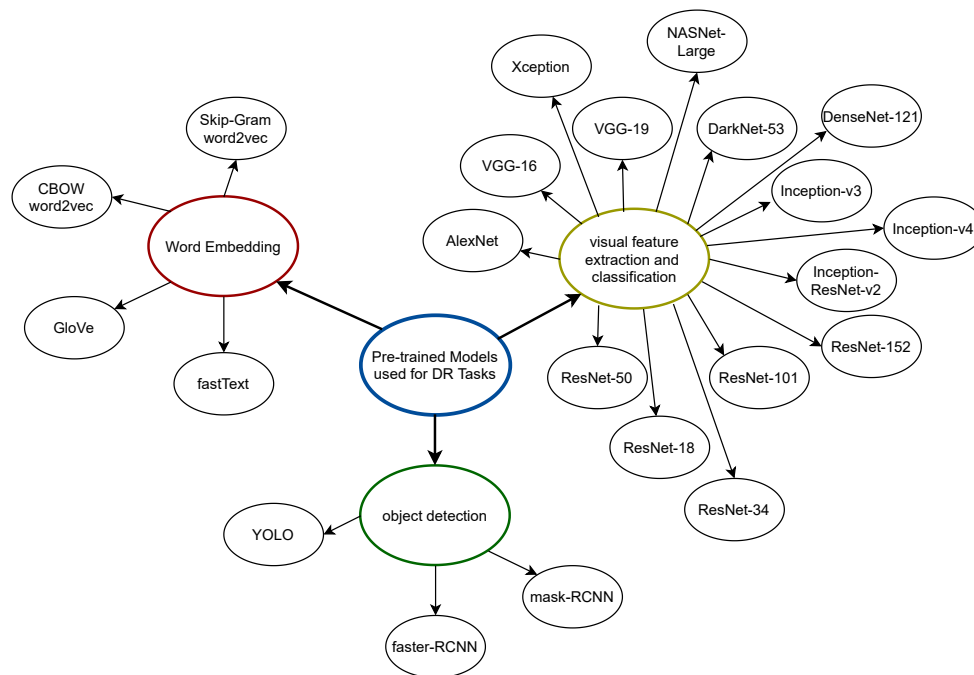
**Figure 2.9** Usage of CNN by DR tasks

2.9), and we present the usage of DL architectures based on publication year in Fig. 2.10. There has been a significant growing interest in using CNNs over the years across all DR tasks in 71 out of 83 articles analyzed. We consider it likely that CNNs have been adopted largely due to their capability in learning features automatically, parameter sharing and dimensionality reduction [345]. However, CNNs have performed poor for identifying word order in a sentence for text classification tasks [237]. Moreover, the computational cost (e.g., training time) for CNNs has been considerable, particularly when the training dataset is large.



**Figure 2.10** DL architectures used by DR tasks by year

RNNs, LSTMs, and Bi-LSTMs have been used to analyze varying length sequence data such as sentences (e.g., tweet text). Although RNNs have been successful in many sequence prediction tasks, it has issues in learning long term dependencies due to the vanishing gradient problem. This problem occurs from the gradient propagation of the recurrent network over many layers [237]. LSTM networks have been proposed to overcome these drawbacks and have shown better results



**Figure 2.11** Pre-trained DL networks used by DR tasks.

for multiple text classification tasks [306]. Recent studies have demonstrated more improved results using Bi-LSTMs. One of the major advantages of using Bi-LSTMs is that they can capture and deal with long-range dependencies having variable lengths by analyzing information in both directions of a sequence (e.g., past and future entries) [140, 179].

We observe that many studies adopt DL models pre-trained on larger data sets such as Places365<sup>10</sup> and ImageNet<sup>11</sup>. Fifty-one of the analyzed papers used pre-trained DL networks for word embeddings, visual feature extraction, object detection and classification. The advantage of adopting a pre-trained model is that it saves time and resources relative to training a model from scratch. Fig. 2.11 provides a taxonomy of pre-trained networks adopted by our analyzed studies.

In addition, we observed that 17 studies adopted multiple DL architectures. This is very common in research that uses different modalities of data. For example, CNNs are often used to extract image features, while RNNs, LSTMs or Bi-LSTMs are used for text feature extraction.

### 2.6.2 $RQ_{3.2}$ *What training processes are used to optimize DL models?*

In this RQ, we analyze the processes used to train DL algorithms focusing on optimization and error calculation.

All but four of the 83 articles used supervised learning as the training type for the selected DR problem. In supervised learning, the DL algorithm extracts features to associate data with the required classification labels. Therefore, a labelled training dataset is required. In contrast, unsupervised learning assigns a class label by grouping similar data together based on extracted features. Therefore, unsupervised approaches do not require labelled training data. Semi-supervised approaches use partially labelled data sets. However, both unsupervised and semi-supervised approaches were rarely used in the analyzed articles resulting in only 4/83. The current favour for supervised learning approaches is mostly due to the readily available labelled datasets. However,

<sup>10</sup>Places365 dataset, <http://places2.csail.mit.edu/download.html>

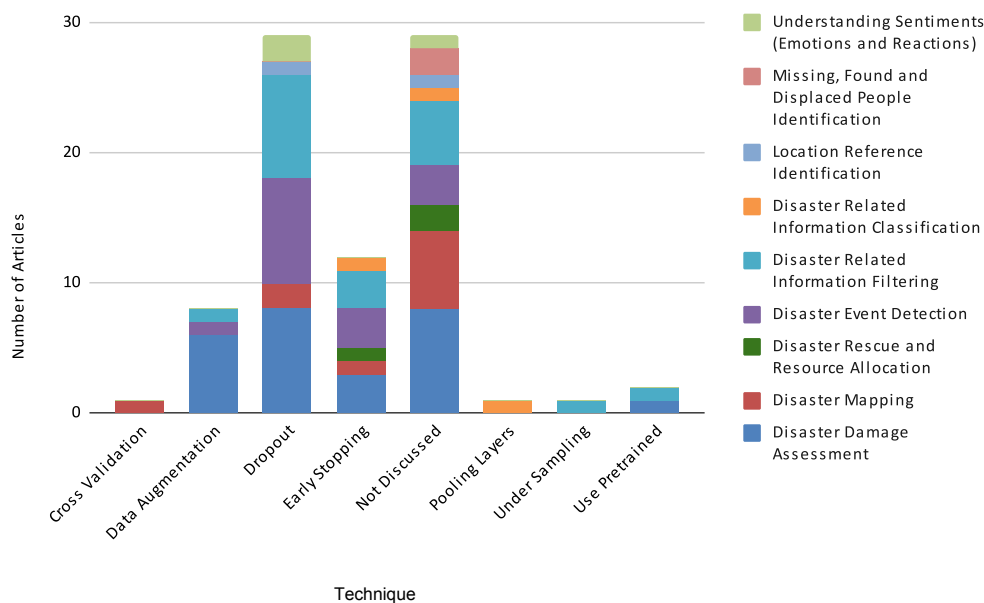
<sup>11</sup>ImageNet dataset, <https://image-net.org/>

those outdated datasets would not reflect the temporal variations, and therefore, more improvements are required for DL architectures to make approximations without training.

The classical gradient descent algorithm was the most frequently adopted learning algorithm in the articles we analyzed for updating weights during backpropagation. Although researchers widely use gradient descent, the computational complexity is considerable because the entire dataset is considered every time the parameters are updated [305]. Multiple other algorithms such as Adaptive Moment Estimation (Adam), Adadelta, and RMSProp algorithms were proposed to overcome this issue. These new techniques have been used for optimization by 45 articles. The selection of optimization algorithm significantly affects the results of the model. However, we could observe that only 31% of the analyzed articles explicitly mention the optimization process and the algorithms they used.

Our analysis found that multiple algorithms have been adopted to calculate the error rate. Categorical cross-entropy is the most frequently used loss function, while negative log-likelihood was adopted by one article. The objective of a loss function is to optimize and tune weights in deep neural network layers. However, only 22 of the papers discussed the error function.

### 2.6.3 $RQ_{3.3}$ What methods are used to avoid overfitting and underfitting



**Figure 2.12** Methods used to avoid overfitting and underfitting by DR tasks

Two common problems associated with generalizing a trained DL model are known as “overfitting” and “underfitting”. Overfitting happens when the model learns training data extremely well but is not able to perform well on unseen data [136]. In contrast, an underfitted model fails to learn training data well and hence performs poorly on new unseen data. This happens due to the lack of capacity of the model or not having sufficient training iterations [171]. In both these cases, the model is not generalized well for the target problem.

To combat overfitting and underfitting, we observed that research had used multiple techniques such as *Dropout*, *Batch normalization*, *Early stopping*, *Pooling layers*, *Cross-validation*, *Undersampling*, *Pre-trained weights* and *Data augmentation*. Fig. 2.12 illustrates these methods by DR tasks.

A total of 24 articles used *Dropout* layers and 12 articles used *Early stopping* to avoid overfitting. Dropout layers ignore nodes in the hidden layer when training the neural network, and therefore, it prevents all neurons in a layer from optimizing their weights [353]. However, the batch normalization technique was proposed to achieve higher accuracy with fewer training steps, eliminating the need for Dropout [169]. During model training, the Early stopping technique evaluates the performance of the model on the validation dataset. The training process is stopped when the accuracy starts decreasing. As a result, however, this technique prevents the use of all available training data. Rice et al. [324] provide remedies for overfitting using a series of experimental evaluations.

Addressing underfitting while training DL models is a complex task, and these are not well-defined techniques [397]. We observed that 2 articles used pre-trained weights to avoid underfitting. However, 29 of the analyzed articles did not discuss the methods used for combating overfitting or underfitting.

## 2.7 RQ<sub>4</sub>: How well do DL approaches perform in supporting various DR tasks?

In this RQ, we analyze the effectiveness of DL approaches for DR tasks, including reviewing the evaluation matrices and baseline models and comparing results achieved.

### 2.7.1 RQ<sub>4.1</sub> *What evaluation matrices are used to evaluate the performance of DL models?*

Through this question, we explore the different performance matrices adopted by the studies we analysed. Our aim is to identify how the existing research evaluated their results. Evaluation of the performance of a model is a core function when employing DL algorithms, as it helps to improve the model constructively. We observed that 76 of the 83 articles had adopted standard performance evaluation matrices such as precision, recall, accuracy, and F1-score (see the definition of these metrics matrices in equations 2.4 to 2.9.). These measures are based on the “true positive”, “false positive”, “true negative”, and “false negative” values, which evaluate the correctness of the results.

		True condition			
Predicted condition	True positive $T_p$	False positive $F_p$	Precision/Positive Predictive Value (PPV) $\frac{T_p}{T_p+F_p} \times 100\%$		
	False negative $F_n$	True negative $T_n$	Negative Predictive Value (NPV) $\frac{T_n}{T_n+F_n} \times 100\%$		
		Sensitivity/Recall Rate (RR) $\frac{T_p}{T_p+F_n} \times 100\%$	Specificity Rate (SR) $\frac{T_n}{T_n+F_p} \times 100\%$		

$$\text{Precision/Positive Predictive Value (PPV)} = \frac{T_p}{T_p + F_p} \times 100\% \quad (2.4)$$

$$\text{Recall/Sensitivity} = \frac{T_p}{T_p + F_n} \times 100\% \quad (2.5)$$

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \times 100\% \quad (2.6)$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \quad (2.7)$$

$$\text{Specificity/True Negative Rate (TNR)} = \frac{T_n}{T_n + F_p} \times 100\% \quad (2.8)$$

$$\text{Negative Predictive Value (NPV)} = \frac{T_n}{T_n + F_n} \times 100\% \quad (2.9)$$

We also observed that Area Under the Receiver Operating Characteristic (ROC) curve value has been used by 6 articles. The ROC curve plots the values between sensitivity and (1- specificity). Sixty-four of the analysed articles presented their performance using more than one metric, while all of the remaining 19 used one metric only. Other metrics used by our analysed articles include Average Precision (AP), and Intersection over Union (IoU). Our analysis suggests that researchers primarily selected performance metrics based on the baseline work that they selected as a comparison for their results. Therefore, it is essential to use standard metrics so other researchers can compare and contrast results in future studies. Table 2.5 shows the best accuracy scores obtained for level-1 and level-2 DR tasks in our taxonomy, revealing that across most tasks DL performs very well, with slightly lower success rates for sub-tasks such as *Damage evaluation* and *Spatial information filtering*.

**Table 2.5** Best Accuracy Scores for DR tasks.

Author	DR task	Sub-task	Best Accuracy Score			
			Precision	Recall	Accuracy	F1-score
[220]	Understanding Sentiments (Emotions and Reactions)	Classification-binary (e.g., Sympathy vs Non-Sympathy)	0.95		0.71	0.76
[334]		Classification-multiclass (e.g., Angry, sad, anxious, fearful)	0.90	0.88	0.93	0.89
[134]	Missing, Found and Displaced People Identification	Human Victim Detection from Visuals			1.00	
[258]		Body Parts Detection from Visuals	0.96	0.99	0.95	
[200]	Location Reference Identification		0.97	0.95	0.96	

[8]	Disaster Mapping	Passable Road Detection					0.65
[265]		Affected Area Detection					0.92
[85]	Disaster Rescue and Resource Allocation		0.94	0.92	0.98		0.87
[85]	Disaster Event Detection	Flood detection					0.86
[228]		Landslide detection	0.98	0.97	0.97		
[418]		Early fire detection					1.00
[63]	Disaster Damage Assessment	Structural damage detection	0.88	0.95	0.99		0.91
[132]		Damage evaluation	0.85	0.78	0.99		
[7]		damage related social media posts detection	0.99	0.99			0.99
[238]	Disaster Related Information Classification	Classification-binary (e.g., Informative vs Not-Informative)				0.96	0.96
[10]		Classification-multiclass (e.g., Affected Individuals, Casualties, Damages)				0.97	
[206]	Disaster Related Information Filtering	Disaster Related Content Filtering	0.92	0.91			0.92
[237]		Situational Information Filtering		0.99	0.66		0.74
[140]		Spatial Information Filtering	0.85	0.82			0.84

### 2.7.2 RQ<sub>4.2</sub> What “baseline” models have been compared?

This question explores the benchmarks that have been chosen by the analysed articles. We observed that the vast majority of the analysed articles self-generated their own benchmark. Specifically, 35 of the studies evaluated the performance of their proposed approach against self-generated tests, while 25 evaluated DL approaches against classical ML approaches. We consider it likely that this is because, until recently, there have not been many DL-based approaches with which to compare. Moreover, the majority of the studies have not published their adopted models or code for future researchers to easily implement and evaluate. Only 12 of the articles selected DL-methods proposed by previous research as baselines. We see that some benchmarks have also been compared in multiple articles as described in our online appendix [23].

## 2.8 RQ<sub>5</sub>: What are the underlying challenges and replicability of DL for DR studies?

In RQ<sub>5</sub> we analyse the challenges researchers face in employing DL algorithms for DR studies and how well the current work can be adopted in future research. We aim to identify common challenges and provide future researchers with knowledge to better design future DL-based projects. Furthermore, we provide the details of research available for replication and reproduction in future research.

We observed that the challenges mostly depend on the data types and sources, including the following, which were extracted from 61 research articles:

1. **Data annotation:** Early studies using supervised approaches found very few publicly available annotated datasets. Therefore, they downloaded their own datasets and recruited people to annotate them. This took a massive amount of time and resources and delayed experiments. Furthermore, multi-label problems (one data item can belong to one or more informative categories), task subjectivity (difficulty in agreeing on one informative class), and conflicting annotation by human annotators were major issues. Even though many annotated datasets are available recently, data incompleteness and bias are common problems in processing DR data.
2. **High-level of noise:** Due to the high volume of heterogeneous data collected from social media platforms in the wake of disasters, the level of noise in the resulting data sets is extremely high (for example, spam, bots, data duplication). Furthermore, the content is informal, mostly using colloquial language, and very brief with casual acronyms and sometimes with non-literal language devices, like sarcasm, metaphors, and double entendre. Thus, it is challenging to train a DL model that can correctly interpret the intention of human expressions of this kind.
3. **High variability:** High variability in image quality resulting from different sensors and environmental conditions (for example, mist, cloud cover, and poor illumination) is challenging when applying DL models. Moreover, debris and damaged buildings look completely different depending on the disaster and structure of the building (e.g. concrete buildings, masonry buildings, or buildings made from natural materials), and are characterised by different features and patterns when captured in an image. As a result, the replicability of an already implemented solution for such a task is very low.
4. **Semantic segmentation:** Semantic segmentation of images to differentiate ground objects, such as roads and trees, from intact and damaged buildings, is a major challenge while using satellite, airborne and UAV imagery.

Despite these challenges, we observed that a very limited number of studies had made available their datasets, annotations, and implementation code for future research. For example, only 5 of the analysed articles made their resources publicly available. This trend results in researchers generating their own baseline and hence reducing research quality and the evolution of the field. Therefore, there is a considerable gap for researchers in adopting previous research as baselines.

## 2.9 Opportunities, directions and future research challenges

With the rapid change of climate and human-induced global warming, the variety and frequency of disasters have increased at a rate that has not happened before [70]. As a result, managing disasters while reducing their impacts on the communities and environment would be one of the main problems of the next decade. The increasing number of smart mobile devices and their embedded sensors enable the generation of a massive amount of heterogeneous data within a significantly shorter time than seen previously during disasters [153, 1]. Therefore, there is an immediate need for robust methods to automatically analyze and fuse such multimodal datasets and provide consolidated information to assist disaster management.

Data from different sources and formats bring complementary information regarding an event and lead to more robust inferences. Thus, future DL models will require analysis of heterogeneous, incomplete, and high-dimensional data sets to fill the missing information gaps in each data source or modality [305]. Multiple studies have explored the use of multimodal data for understanding the big picture of a disaster event [389, 281, 1, 306, 7]. However, more and more advanced DL approaches are required to solve core challenges in multimodal deep learning, such as missing data, dealing with different noise levels and effective fusing of heterogeneous data [40].

To address this problem, we identify that training data acquisition and preprocessing plays a major role when employing DL approaches. For example, large-scale human-annotated datasets are required to train DL algorithms to successfully predict the class label for unseen data. While a few annotated data repositories have been created (e.g., CrisisNLP, CrisisMMD, and CrisisLex), more datasets are required to reflect temporal variations. Furthermore, there are still no large-scale benchmark datasets incorporating a variety of disaster data types except for CrisisMMD [18]. Therefore, the current research is mostly limited to small-scale home-grown datasets covering specific disaster types.

This leads to the next challenge of data irregularities occurring in datasets and which reduce a classifier’s ability to learn from the data. The most common data irregularities include class imbalance, missing features, absent features, class skew and small disjuncts [82]. Class imbalance occurs when all classes present in a dataset do not have equal training instances. For example, datasets for classifying disaster-related social media posts have resulted in most non-related posts. Data-level methods such as under-sampling techniques (e.g., Random Under-Sampling (RUS) [173]) and over-sampling techniques (e.g., Generative Adversarial Minority Oversampling (GAMO) [264] and Major-to-minor Translation (M2m) [192]) have been explored to mitigate the effects of class imbalance. Although researchers assume fully observed instances, practical datasets, however, contain missing features. Data imputation methods, model-based methods and more recently, DL methods have been proposed to handle missing features. A complete guide to methods that enable tackling these data irregularities is provided by Das et al. [82]. Even though methods to handle irregularities have been largely explored, more research is required as the velocity and variability of data generation accelerate.

Another key area is the variety characteristics of disasters that limit the reusability and generalizability of already trained DL algorithms. This means the variations of input data representations extracted during different disasters. Recent DL studies have focused on domain adaptation during learning where the distribution of the training data differs from the distribution of the test data [157]. Future research focus requires developing domain adaptation techniques for the DR domain.



According to the current trends, people will increasingly use social media platforms for disaster data acquisition, and dissemination, challenging the traditional media sources [133, 359, 335]. Therefore, crowd-sourced data will be more prominent in providing first-hand experiences of disaster scenes. However, responding organizations have concerns regarding the trustworthiness of user-generated content, a problem which is largely unsolved [62]. For example, fake news, misinformation, rumours, digital manipulation of images (e.g., deepfake [399]) and re-posting contents from previous events are a few challenges that future researchers will face to improve the integrity of SM content.

Another challenge in the DR domain is that previous research has largely explored the most common tasks such as *Disaster damage assessment*, *Disaster event detection* and *Location reference identification*. However, there are other important DR tasks, including evacuation management, health and safety assurance, and critical infrastructure service, as illustrated in the Guidance of Emergency Response and Recovery [97]. These tasks have not yet been analyzed using DL approaches. Some possible reasons could be insufficient training datasets, lack of computational resources to store, manage, and process data, and inadequate accuracy of existing DL architectures. These underrepresented topics need further attention by DL researchers to better support DR tasks. Moreover, the accuracy of the output produced by DL algorithms is determined by a number of factors, including the optimization algorithm and the loss function used. Thus, further research is important in this area to find the correct combination of data, DL architecture, optimization algorithm, and loss function.

## 2.10 Results of the Association Rule Mining

**Table 2.6** Some association rules extracted from the analysed papers

Item	Support	Item	Confidence
Supervised	0.94	Damage Assessment→Remote Sensing	1.0
CNN	0.70	Remote Sensing→Image	1.0
Twitter	0.48	Multimodal, CrisisMMD→Twitter	1.0
image	0.45	Remote Sensing→CNN	1.0
Item	Lift		
Multimodal, Twitter →CrisisMMD	4.50		
Multimodal →CrisisMMD	3.46		
Remote Sensing →Image	2.24		
Remote Sensing, CNN →Image	2.24		

This section discusses the interesting relationships discovered through our association rule mining task. We introduced the association rule mining process in Section 2.3.5.1. Our goal is to identify hidden relationships between the values extracted from the articles for the attributes in the extraction form. The most highly scoring rules are listed in Table 2.6. We discuss the patterns that resulted in having higher “Support”, “Confidence” and “Lift” values. However, all the associations are illustrated in our online appendix [23]. Our analysis highlights that CNN, Supervised, Image and Twitter have higher support values ( $> 0.45$ ). This result indicates that the majority of studies discussed Image as data type, CNN as DL architecture, Supervised as learning type and Twitter as their data source.

*Disaster Damage Assessment*  $\rightarrow$  *Remote Sensing*; *Remote Sensing*  $\rightarrow$  *Image*; *Multimodal*, *CrisisMMD*  $\rightarrow$  *Twitter* and *Remote Sensing*, *CNN*  $\rightarrow$  *Image* are some of the association rules having a confidence score of 1.0. This means that, for example, rule *Disaster Damage Assessment*  $\rightarrow$  *Remote Sensing* implies that the pattern appears in 100% of the analysed articles. Similarly, all the research that used Remote Sensing as the data extraction method analysed Image as their data source.

The highest lift score of 4.5 resulted for the *multimodal*, *Twitter*  $\rightarrow$  *CrisisMMD* rule. This means that when researchers used multimodal as their data type and Twitter as the Data Source, CrisisMMD has commonly been the data extraction method. Furthermore, *multimodal*  $\rightarrow$  *CrisisMMD*, *Twitter*; *Remote Sensing*  $\rightarrow$  *Disaster Damage Assessment*, *Image*; *Image*  $\rightarrow$  *Remote Sensing*, *CNN* rules were among the other high lift values. Interestingly, we observed rules such as *Twitter*  $\rightarrow$  *CNN*; *CNN*  $\rightarrow$  *text* and *text*, *Twitter*  $\rightarrow$  *CNN* having a “Lift” score of less than 1. This indicates a negative relationship between the parameter values. For example, it is very unlikely that research that used Text as the Data Type and CNN as the DL architecture. All these association rules provide future researchers a guide to select parameters in a DL-based project such as data sources, learning algorithms, and learning type.

## 2.11 Flowchart and guidelines for applying DL in future DR research

In this section, we provide a flowchart and guidelines for conducting future work using DL for DR tasks based on the findings of our SLR. Fig. 2.13 shows how we have mapped the components of learning into RQs and then as the steps in the flowchart. The extracted flowchart is a general one based on the 83 analyzed papers. However, more specific details can be added to it based on the DR task to be solved.

After identifying the DR problem to be addressed, researchers should consider whether DL is a suitable approach. That decision can be made partly based on whether it is possible to obtain or create the necessary data. If enough data can be obtained, the researcher can select either supervised, unsupervised and semi-supervised learning methods. We discussed these methods in the Section 2.6.2. If the identified problem can be better solved using a supervised approach, the next step is to decide where the annotated datasets can be obtained, or whether raw data must be annotated. Data annotation is generally labour intensive and time-consuming, and therefore, the researcher can hire paid workers or arrange volunteers based on budget and availability. We have discussed the annotated data sources and annotation methods in the Sections 2.5.2 and 2.5.3. Once the dataset is ready, the researcher should conduct an exploratory analysis to identify the nature of this raw data. This analysis provides the researcher with an overview including the size, distribution, and characteristics of the data. Proper understanding of raw data provides guidelines for the design of the preprocessing steps, which have to be well reported to enable replication. This includes outlining all the steps involved, including the normalization processes and data augmentation strategies.

After the data filtering and cleaning steps, the researcher should identify the learning algorithm, and DL architecture. The researcher should report the details of the DL architecture, including the type of layer (e.g., embedding, dropout and soft-max), number of layers, filters, and learning rate. Furthermore, all necessary details regarding optimizers, loss function and hyper-parameter tuning, have to be reported to enable replication. The information regarding training such as number

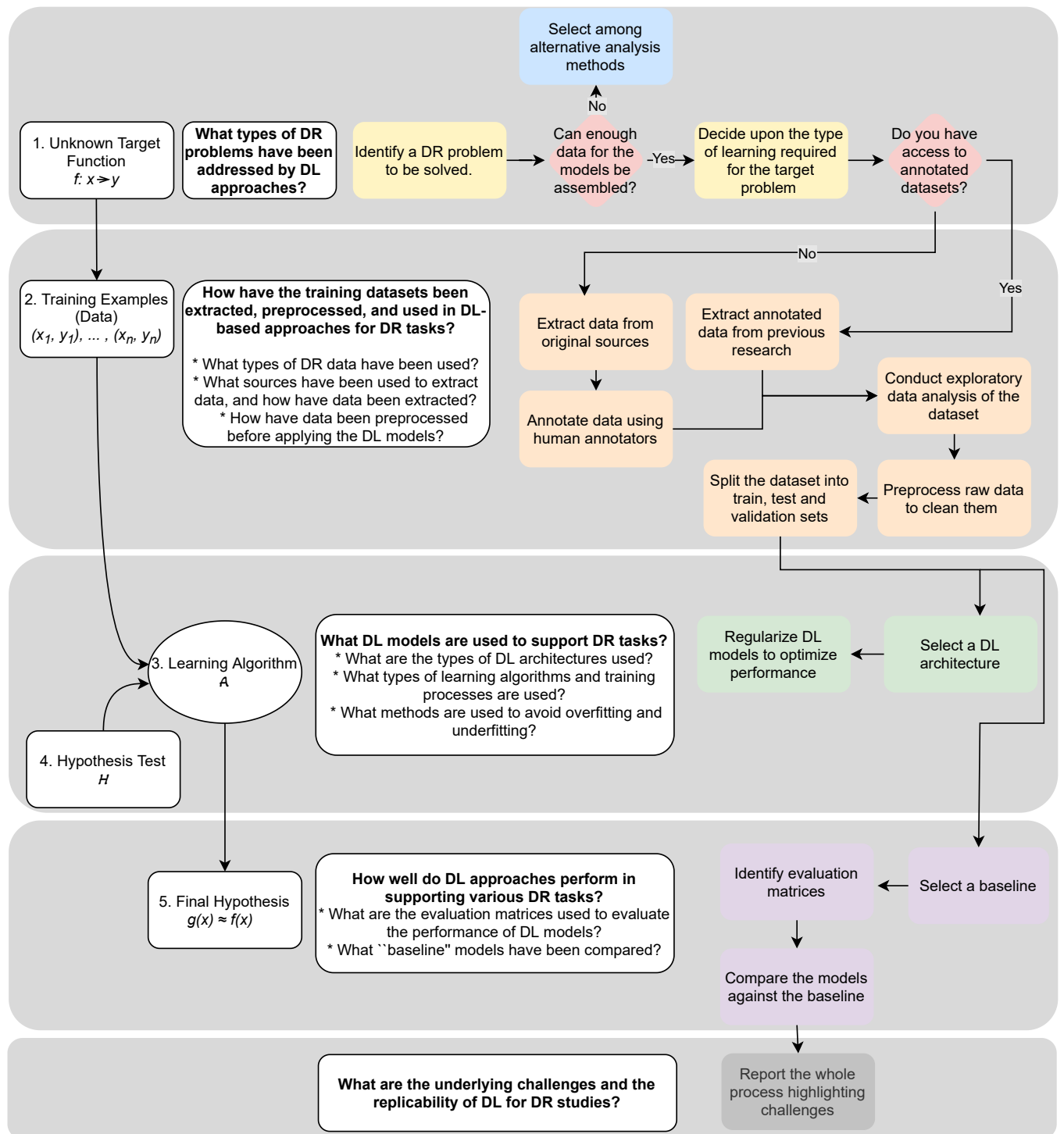


Figure 2.13 Flowchart for conducting DL for DR research

of iterations (epochs), strategies combating overfitting and underfitting, training time, computing environment, special computing resources (e.g., GPUs, high performance computing) and platforms used (e.g., Google Colaboratory) should also be explained (see Section 2.6).

Finally, the researcher should report the results compared to the selected “baseline model”. If the researchers used their own dataset, they must first implement the baseline against their data to compare the results. Any limitations and challenges encountered while applying DL models should also be discussed to provide guidance for future researchers in designing DL-based approaches for DR tasks. Furthermore, researchers can support the quality and the future of the DR research field by making publicly available the datasets, annotations, and DL architectures.

## 2.12 Conclusions

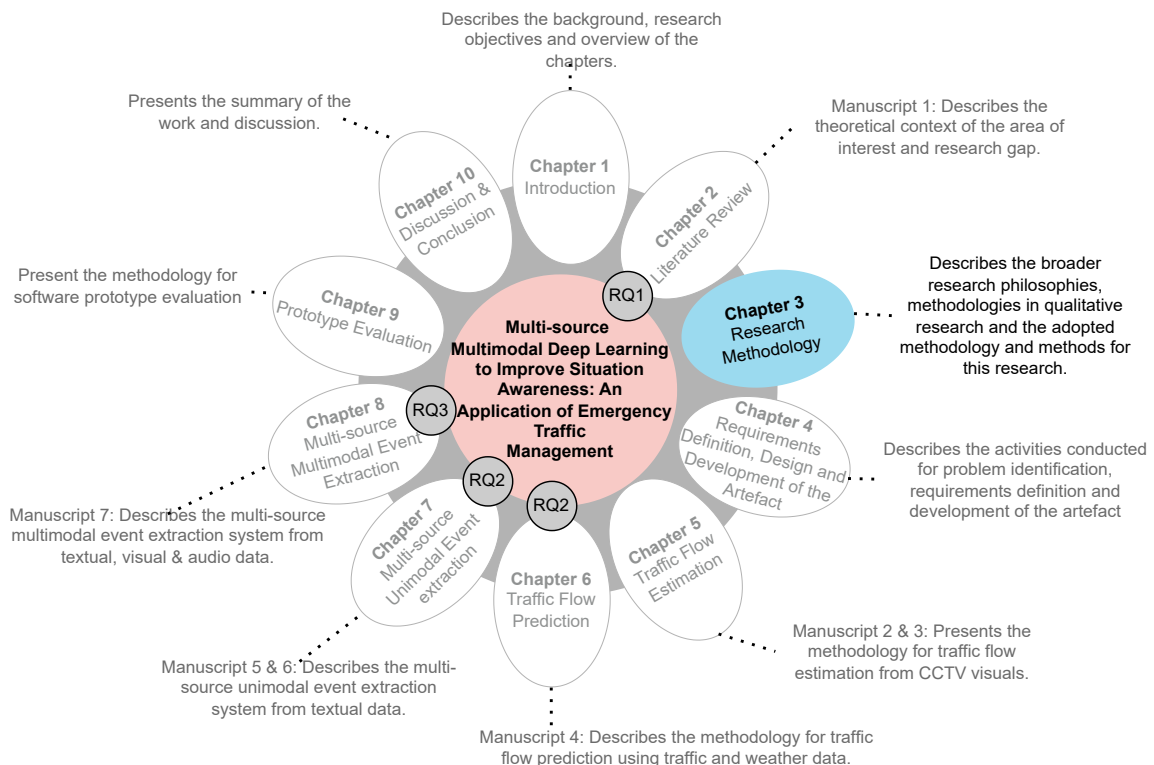
This study has presented a systematic literature review of DL in DR research. We started by identifying RQs for the analysis according to the components of learning described by Abu Moftha [366]. Then, a data extraction form with 15 attributes was created to extract answers for the questions from the selected articles. Finally, we used the KDD process to identify relationships among different attributes of the extracted data. The answers to the research questions indicate that, while some DR tasks have received much investigation, others have received less attention. Furthermore, there are multiple challenges while collecting, annotating, and preprocessing datasets for DL tasks. However, researchers have achieved better performance than traditional methods when using DL methods for DR tasks despite these challenges.

This research has identified opportunities, future research challenges, and many directions for further investigation. For example, multiple DR tasks are yet to be studied using DL approaches, such as *evacuation management* and *critical infrastructure services*. Moreover, we highlighted the need for new annotated multimodal datasets targeted at DR concerns. Some of the future research challenges are handling data irregularities, improving the integrity of social media data, and developing generalizable DL approaches across multiple disasters. Additionally, data preprocessing, DL architecture selection, word embeddings and hyperparameter tuning are areas of further exploration. Finally, we emphasized the importance of comprehensive reporting and making implemented DL methodologies publicly available for the advancement of the DL in the DR area.

The next chapter discusses the relevant methodologies and methods that allowed the thesis to explore how to help disaster responders’ SA by processing multimodal data using DL techniques.

# Chapter 3

## Research Methodology and Design



This chapter introduces the philosophical approach and the research framework adopted for this study. Section 3.1 of the chapter discusses the broader philosophical viewpoint and rationale behind the chosen approach. Section 3.2 of the chapter explains Design Science and discusses the method framework followed for designing, developing, and evaluating the artefact in the succeeding sections.

### 3.1 Research Philosophy

Research philosophy is a belief of how data about a phenomenon should be gathered, analyzed, and used [338]. Generally, a researcher makes a number of assumptions through every step of the research [57]. These include assumptions about human knowledge (epistemological assumptions) and assumptions regarding the nature of the reality encountered by the researcher while conducting research (ontological assumptions) [338, 57, 78]. These assumptions inevitably shape how the researcher understands the research questions, the methods to use, and how to interpret the findings.

The research philosophy leads to the methodological choice, research strategy, and data collection techniques and analysis procedures [78].

Due to the scope and nature of the problem driving this Doctoral thesis study, it is considered to fall under “computer science” research. Peter Wegner argues that the practices of computer scientists are effectively committed to one of three research paradigms: rationalist, technocratic, and scientific [398].

The rationalist paradigm, which was common among theoretical computer scientists, defines the discipline as a branch of mathematics [398]. Therefore, writing programs is considered a mathematical activity, and the only accepted method of investigating programs is deductive reasoning, which involves drawing conclusions based on premises that are generally assumed to be true [106].

The technocratic paradigm, promulgated mainly by software engineers, defines computer science as an engineering discipline that is concerned primarily with manufacturing reliable computing systems. The quality of computing systems is determined by established engineering methods such as reliability testing and obtained through a regimented development and testing process. Therefore, the technocratic paradigm suggests that it is impractical to specify formally or to prove deductively the ‘correctness’ of a complete program, and the reliability can be only proven by means of testing [398, 95].

The scientific paradigm prevalent in artificial intelligence defines computer science as a natural (empirical) science comparable to astronomy, geology, and economics [269]. The only significant difference between various topic matters in computer science is the limitations of scientific theories [95]. Moreover, according to the scientific paradigm, computer programmes are equivalent to mental activities. Therefore, computer science approaches combine both deduction and empirical validation in order to explain, describe, comprehend, and predict the behavior of computer programmes [398].

This PhD research aimed to develop learning algorithms that can validate SM content by integrating different modalities of data to support disaster response. This aim was achieved by developing multiple components utilizing deep learning/ machine learning techniques. Therefore, it may be argued that in order to achieve the research objectives, this study required the testing of claims to cover the scope of the research problem and successfully achieve the objectives. Therefore, after considering the nature of the research problem and determining the research questions, aims, and objectives, it is argued that *the scientific paradigm* is the most suitable epistemological approach for this research.

The objectives of this research reiterated from Chapter 1 are:

Objective 1: To identify deep learning algorithms that can be used to analyze different modalities of data extracted from different sources for disaster response.

Objective 2: To develop algorithms that are capable of integrating multi-source unimodal data.

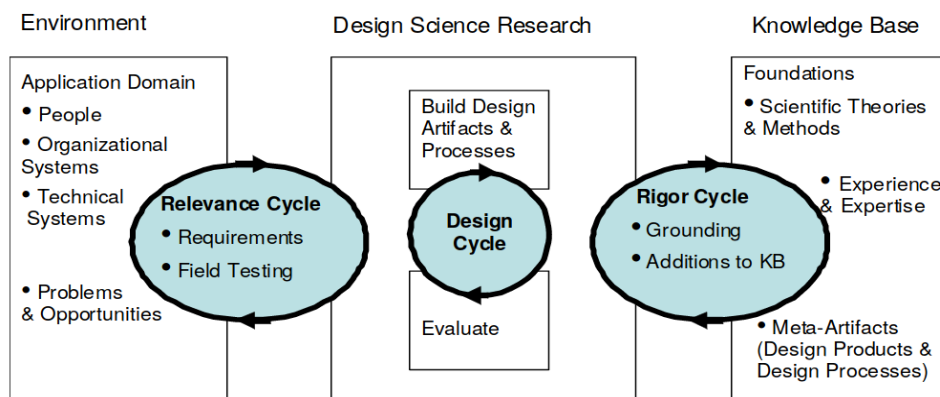
Objective 3: To develop algorithms that are capable of integrating multi-source multimodal data.

Achieving the second and third objectives required engaging with stakeholders to understand their requirements. In addition, the research outcome was determined to have a practical benefit

to the environment from which the requirements have been identified. Therefore, a broader research framework that supports socio-technical research was necessary to guide this research. The suitability of design science research to undertake the study is discussed in the following section.

## 3.2 Design Science Research (DSR)

Johannesson et al. [175, p. 8] define design science as “the scientific study and creation of artefacts as they are developed and used by people with the goal of solving practical problems of general interest”. For example, in design science, researchers take an intentional stance in the sense that they consider an artefact as something that should help people in a practice. A design artefact can be a physical entity, a drawing, a set of guidelines, or an Information and Communication Technology (ICT) solution. Hevner et al. [142] describe DSR as having three overlapping research cycles: the relevance cycle, the design cycle, and the rigor cycle (see Figure 3.1). The relevance cycle bridges the contextual environment of the research project with the design science activities, and the design cycle iterates between the core activities of building and evaluating the design artefacts and the processes of the research. The rigor cycle connects the design science activities with the knowledge base of scientific foundations, experience, and expertise that informs the research project [142, 143].



**Figure 3.1** Design science research cycles adapted from Hevner et al. [142]

The following subsections describe the relevance, design, and rigor cycles.

### 3.2.1 The relevance cycle

The relevance cycle provides the requirements to design studies and defines acceptance criteria for the final study outcomes assessment [385]. Therefore, it enables the researcher to understand the answers to the question: “*Does the design artefact improve the environment, and how can this improvement be measured?*”. The environment consists of the people, organizational systems, and technical systems that interact with work toward a goal [385]. Generally, the output from the DSR must be returned to the environment for evaluation in the environment (field testing). As a result, the design science researcher can determine whether additional iterations of the relevance cycle are needed for the particular project in two instances:

- the new artefact having deficiencies in functionality, or in its inherent qualities that may limit its utility in practice, or

- the requirements for the design science research were incorrect or incomplete, with the resulting artefact satisfying the requirements but still inadequate.

Another iteration of the relevance cycle can occur based on the feedback of the field testing [385, 142].

### 3.2.2 The rigor cycle

The rigor cycle provides prior knowledge to the research project to ensure its innovation. Design science builds on a large knowledge base of scientific theories and techniques, providing the foundation for rigorous design science research. The knowledge base also provides two additions as follows:

1. the experiences and expertise that define the state-of-the-art in the application domain of the research, and
2. the existing artefacts and processes found in the application domain.

It is up to the researcher to thoroughly review the knowledge base to make sure that their designs are research contributions [385]. Additions to the knowledge base may result from the design science research, in the forms of extensions to original theories, methods, design products and processes, and all experiences gained from performing the research. The extensions would be attracted by the academic audience as research contributions, and the practitioner audience as contributions to the environment [142].

### 3.2.3 The design cycle

The design cycle is the heart of the DSR, which iterates between the construction of artefacts and their evaluation. This cycle rapidly iterates, generating design alternatives, and evaluating the alternatives against requirements until a satisfactory design is achieved [348]. The design cycle depends on the relevance cycle for the user requirements and the rigor cycle for the theories and methods. However, it maintains relative independence during actual execution. Hevner et al. [142] advise the researcher to maintain a balance between the efforts spent in constructing and evaluating the evolving design artefact. Furthermore, he states the importance of rigorous and thorough testing of artefacts in laboratory and experimental situations before releasing them into field testing in the relevance cycle.

According to Johannesson et al. [175], a project must satisfy three conditions in order to be considered as design science, in accordance with Hevner's three-cycle model.

- Firstly, the project has to apply an overarching research strategy to investigate the context of the problem and elicit stakeholder requirements. This strategy includes research methods for data collection and analysis. The project also needs to evaluate the artefact produced using appropriate research strategies and methods, which can be different from those used for problem analysis and requirements elicitation.
- Secondly, the project has to relate the results to existing knowledge within various subareas of the problem domain. This knowledge comprises relevant artefacts as well as established theories and models. It is only possible to evaluate the originality and validity of a project's findings by comparing them to existing knowledge.



- Thirdly, the project must disseminate its results to researchers and professionals by publishing them in academic journals and conferences and presenting them at professional conferences and other similar events.

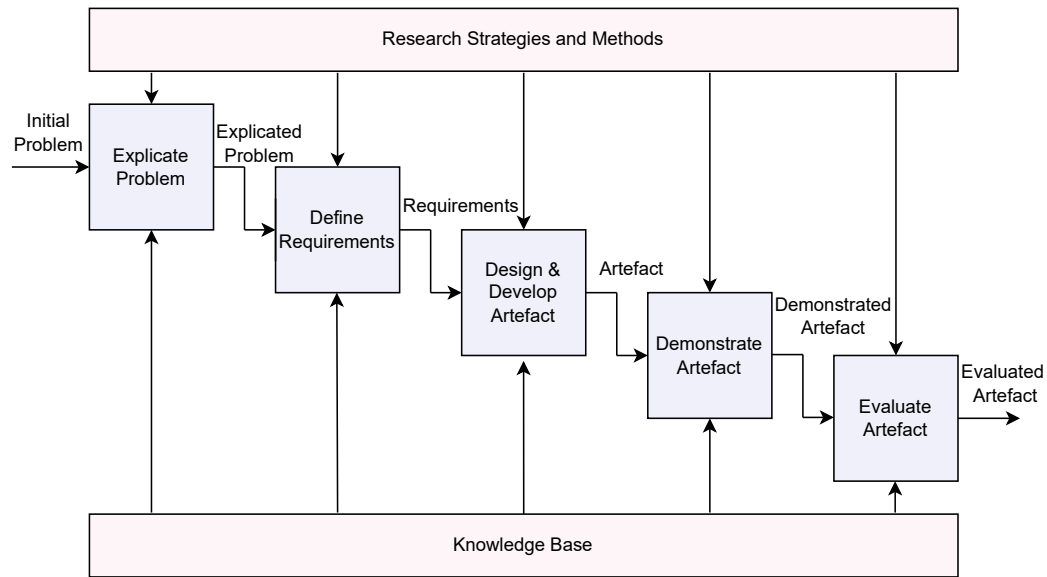
This research aimed to develop an artefact that uses deep learning algorithms for validating SM content to support the SA of disaster responders. The requirements for the artefact development were required to be identified from the user environment. Moreover, the artefact must be exhibited to the users who provided the requirements to determine if it met their expectations. Therefore, appropriate research strategies and methods were selected to explore and evaluate the problem. Therefore, these steps satisfy the first condition. A comprehensive literature search was carried out in this research to select the most appropriate techniques to develop and evaluate the artefact. In addition, new algorithms or extensions to the existing algorithms needed to be developed and evaluated against the existing foundations of the knowledge base throughout the development process of the artefact. Hence, the second condition is satisfied. Finally, the results were planned to be disseminated in multiple research venues, including journal publications, conference publications and presentations, lightning talks and posters satisfying the third condition. Therefore, it can be argued that these steps are clearly aligned with the three conditions proposed by Johannesson et al. [175], and so, this Doctoral research fits under design science research.

### 3.2.4 Method Framework for Design Science Research

Johannesson et al. [175] introduced a method framework for design science research that any design science project can utilise following Hevner's three cycles model [142]. The framework consists of several logically related activities with well-defined inputs and outputs, guidelines for conducting the activities, guidelines for selecting research strategies and methods to be used in the activities, and guidelines for relating the research to existing knowledge. Johannesson's et al. [175] summarises these activities into the following five-phase activity framework.

- **Explicate Problem**—This activity involves investigating and analysing a real-world problem. Once the problem has been recognised, it must be defined and justified by demonstrating its relevance to some practice. Furthermore, the root causes of the problem must be recognised and investigated.
- **Define Requirements**—This activity describes a solution to the described problem as an artefact and elicits requirements, which can be seen as a transformation of the problem into requirements for the proposed artefact. The requirements for functionality, structure, and environment will be specified.
- **Design and Develop Artefact**—This activity creates an artefact that addresses the described problem and meets the specified requirements. Designing an artefact involves determining both its functionality and structure.
- **Demonstrate Artefact**—This activity shows the viability of the developed artefact by applying it to an example or real-world scenario, sometimes known as a “proof of concept”. Additionally, the presentation will demonstrate how the artefact can resolve a specific instance of the issue.
- **Evaluate Artefact**—This activity assesses how well the artefact meets the requirements and how effectively it can address or alleviate the practical problem that motivated the research.

This method framework is illustrated in Figure 3.2.



**Figure 3.2** Method framework for DSR introduced by Johannesson et al. [175]

Having identified as a better-suited design science-driven research framework, the 5-stage design science process framework proposed by Johannesson et al. [175] was used as the guide for creating, developing, and evaluating the artefact proposed in this thesis. Chapter 4 discusses the Explicate Problem and Define Requirements activities. The artefact’s design is given in Chapter 4, and its development is covered in detail in the following four chapters (Chapter 5, Chapter 6, Chapter 7 and Chapter 8). Finally, the Demonstrate Artefact and Evaluate Artefact activities are discussed in Chapter 9.

Johannesson et al. [175] argue that design science projects can use several research methodologies and methods because different design science activities may require different approaches. Therefore, sections 3.3 and 3.4 provide an overview of research methodologies and methods in general and the rationale behind the selected methodologies and methods for this doctoral research.

### 3.3 Research Methodology

Research in design science seeks not just to produce artefacts but also to answer questions about them and their contexts. Research methodologies or strategies and methods are essential for ensuring that the answers reflect reliable knowledge [142, 175]. Saunders et al. [246] define research methodologies as derived from many factors including the research questions, objectives, existing literature, research resources, approaches, and the researcher’s philosophical views. These methodologies have been broadly categorised into qualitative, quantitative, and mixed-methods [75].

According to Creswell et al. [75], experimental and non-experimental designs are the two key quantitative methodologies. Locke et al. [231] defined the three key qualitative methodologies as ethnography, case study, and action research, with grounded theory overlapping all three. Mixed-method researchers have adopted convergent, explanatory sequential, and exploratory sequential methodologies. These methodologies are summarised in Table 3.1.

Quantitative	Qualitative	Mixed-methods
Experimental	Ethnography	Convergent
Non-experimental	Case study	Explanatory sequential
	Action research	Exploratory sequential
	Grounded theory	

**Table 3.1** Alternative research methodologies.

**Quantitative methodologies:** Experimental and non-experimental are the two key quantitative methodologies. Experimental research assesses the impact of a particular treatment by providing it to one group, withholding the other group, and comparing the scores. In comparison, non-experimental research provides numerical descriptions by analyzing a sample of a population [75].

**Qualitative methodologies:** Ethnography, case study, action research, and grounded theory are the methodologies categorized under qualitative methodologies. Among them, ethnography is the study of societies and customs conducted to understand the culture of the group being researched [72, 337]. One main focus of ethnography research is in-depth participatory observation. In contrast, the grounded theory methodology is considered the most direct form of an inductive research approach, aiming to use the defined theory to develop recommendations [72]. As defined by Robson et al. [328], the case study was proposed as a strategy for doing research that involves an empirical investigation of a particular contemporary phenomenon within its context using multiple sources of evidence. Thus, it is recommended that this methodology is well suited for research where the context is an integral part of the study. Furthermore, it provides a researcher with a methodology to observe and explain contemporary events over which the researcher has little or no control [328]. In comparison, there is no single, commonly accepted definition for action research. However, it composes the dual outcomes of action and research or (change and understanding) at the same time. Action research includes the researcher taking genuine action, which comprises a series of successive research cycles comprising planning, action, and reflection [88]. In addition, action research encourages system development efforts to be carried out with the in-depth collaboration of the end-user [101].

**Mixed-methods methodologies:** Mixed methods involve combining or integrating qualitative and quantitative research and data in a research study. According to Cresswell et al. [75] there are three methodologies under mixed-methods, namely, convergent, explanatory sequential, and exploratory sequential (see Table 3.1). The convergent strategy combines both qualitative and quantitative data to provide a comprehensive analysis of the research problem. However, the initial quantitative data results are explained further with the qualitative data in *explanatory* sequential mixed methods [75]. As opposed to explanatory strategies, the reverse sequence is followed in *exploratory* sequential methods. The researcher begins with qualitative research, analyses data, and then findings from that data are used to build a second, quantitative phase.

The research questions of this study reiterated from Chapter 1 are:

Research Question 1: How have different deep learning algorithms been applied to data from various sources to support disaster response tasks?

Research Question 2: How can data from multiple sources be fused to support disaster response?

Research Question 3: How can the integration of multi-source multimodal data effectively

support disaster response by cross-validating social media data?

Investigating various disciplines, such as social science, technology, and engineering, was necessary to find the answers to these questions. Moreover, the final artefact, which answers the research questions, was expected to be developed as components. The development of these components required both quantitative and qualitative methodologies. Thus, *Mixed-methods - convergent strategy* was selected as the most appropriate methodology for this research.

## 3.4 Research Methods

Identifying data collection and analysis methods is important to achieve the research objectives. The following sections describe different research methods and their applicability to this doctoral study.

### 3.4.1 Data collection methods

Data collection on the phenomenon under study is crucial in any empirical research project. Data can be numeric (quantitative) or other kinds such as text, audio, and visuals (qualitative). Data collection methods are used to collect data regardless of the type. This research required data for requirement capturing, artefact evaluation, and machine learning and deep learning algorithm training and testing. There are different methods for data collection such as observation, questionnaire, and interview [246]. There are multiple methods to collect data for algorithm training and testing including web scraping, through Application Programming Interface (API)s, social media, and from third party data sellers [1, 20]. Among these methods, the suitability of the data collection depends on the nature and the scope of the project.

The *observation method* requires the researcher to participate fully in the activities of subjects and thus become a member of their group, organization, or community. This allows the researcher to investigate and feel the real-world scenario of the environment under investigation [337].

The *questionnaire* is a data collection method in which each participant is asked to respond to the same questions. It is a quick and easy approach to getting responses from a bigger group of people. However, it can be challenging to create a high-quality questionnaire that can collect the precise and sufficient data needed to answer research questions and achieve study objectives [285, 338].

An *interview* is a purposeful discussion between two or more people to gather valid and reliable data relevant to research questions and objectives. Interviews allow the researcher to gather rich information while also enabling them to establish personal contacts [338].

Interviews can be conducted in a variety of modes including *structured*, *semi-structured* and *unstructured* [314, 87]. In structured interviews, the interviewer asks a series of pre-established questions, allowing only a limited number of response categories. In contrast, unstructured interviews are set to make the interviewee feel relaxed by being more informal. Semi-structured interviews lie in between by allowing the “interviewer to modify the style, pace, and order of questions to evoke the fullest responses from the interviewee. For example, semi-structured interviews help develop an understanding of how managers make sense of and create meanings about their jobs and their environment” [314, p. 246]. Table 3.2 presents a comparison of interviewing, observation and questionnaires.

**Table 3.2** Comparison of data collection methods in qualitative research. Sources: [3, 174, 337]

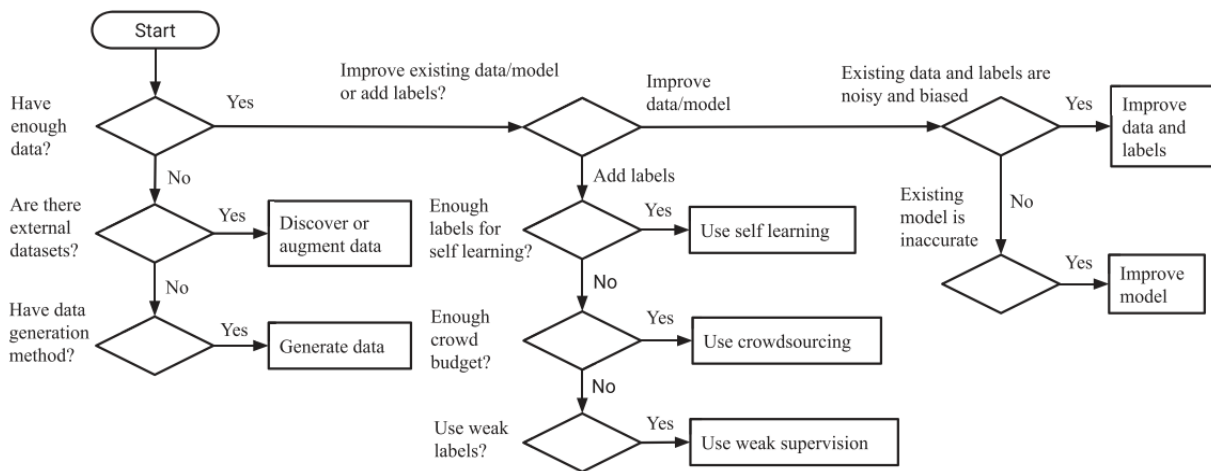
Factors to consider	Interviewing	Observation	Questionnaires
Sensitivity to subject matter	If the subject matter is sensitive, the responders will not temper responses or withhold information	If the subject matter is sensitive, the responders may alter their behaviour	Responders will provide more details in anonymous questionnaires
Depth of individual responses	A greater depth of individual responses are desirable with complex subject matters and very knowledgeable responders	The depth of the information can not be observed for a longer period as responders may experience fatigue	The depth of responses depends on the design of the questionnaire and therefore the researcher has to carefully design them
Extent of issues to be covered	A greater volume of issues can be covered	Only non-verbal issues can be covered. No scope for probing	A limited amount of issues can be covered
Continuity of information	Provide scope to understand how attitudes and behaviors link together on an individual basis	Provide scope to understand the responder's behavior for a limited time	Provide less opportunity to understand the linking of attitudes and behaviors
Emotional and cognitive aspects	Ability to experience the affective and cognitive aspects of responses	Limited scope for understanding emotional and cognitive behaviour	Difficult to understand the emotional and cognitive aspects of responses
Quality of data	Rich and in-depth	Rich and mostly non-verbal data	Collection of rich information depends on many factors such as the design of the questionnaire and response rate

According to the comparison in Table 3.2 interviews and questionnaires were selected as more suitable for this research as the aim of the qualitative data collection is to identify the current challenges, requirements, and the level of acceptability of the final outcome. Furthermore, among different types of interviewing, the semi-structured method was considered the most suitable because it allows follow-up questions to be asked based on interviewees' responses. As a result, the essential requirements and evaluation feedback for the interviewer can be collected.

Various sampling techniques can be used to obtain a representative sample for interviews. For

example, purposive sampling is used to select interview participants with first-hand knowledge and experiences [76]. There are two other broad sampling techniques such as *probability* and *convenience* sampling [364]. Probability sampling is used to obtain a larger number of units of a population in a random manner, mostly in quantitative research. In contrast, convenience sampling involves drawing samples that are both easily accessible and prepared to participate in a study [364, 76]. Purposive sampling was identified as more useful in this research as it allows the researcher to select interviewees based on specific purposes associated with answering the research questions.

Data collection is critical in a successful machine learning/ deep learning project because the algorithms heavily depend on data [210]. At a higher level, data collection for machine learning tasks can include acquiring a newer dataset, labeling data, or improving existing data [331]. Roh et al. [331] developed a flowchart for data collection in machine learning research.



**Figure 3.3** A decision flow chart for data collection in machine learning research (source: Roh et al. [331])

### 3.4.2 Data analysis methods

Before conclusions can be drawn from raw data, it is necessary to prepare, interpret, and present them. Data analysis aims to extract useful information from raw data in order to describe or explain the phenomenon under investigation [337, 246, 175]. Two main kinds of data analysis are quantitative and qualitative, where quantitative data analysis generally works on quantitative data (e.g., numbers), while qualitative data analysis works on qualitative data (e.g., words, images) [175].

#### 3.4.2.1 Quantitative data analysis

Quantitative data analysis includes statistical methods such as descriptive statistics and inferential statistics. Descriptive statistics use measurements like mean, median, mode, range, and standard deviation to quantitatively describe a data sample. In contrast, inferential statistics seeks to establish conclusions that go beyond a single data sample (e.g., determine if a relationship exists between two variables or if a difference exists). A common technique used for this purpose is to calculate the correlation coefficient. A correlation coefficient is a number between +1 and -1, where +1 represents a strong positive relationship, -1 represents a strong negative relationship, and 0 represents no relationship.

Apart from these measures, other quantitative data analysis methods are used specifically for research with machine learning/ deep learning algorithms. These include exploratory data analysis and performance evaluation methods such as precision, recall, accuracy, and F1-score. Exploratory data analysis helps to understand a dataset, including features and relationships among them, missing values, and outliers. Precision, recall, accuracy, and F1-score values are defined and explained in Section 2.7.1 of Chapter 3.

In this study, quantitative data analysis was necessary for data analysis and algorithm creation. For example, exploratory data analysis was considered an essential step to better understand the data before developing any machine learning/ deep learning algorithms.

### 3.4.2.2 Qualitative data analysis

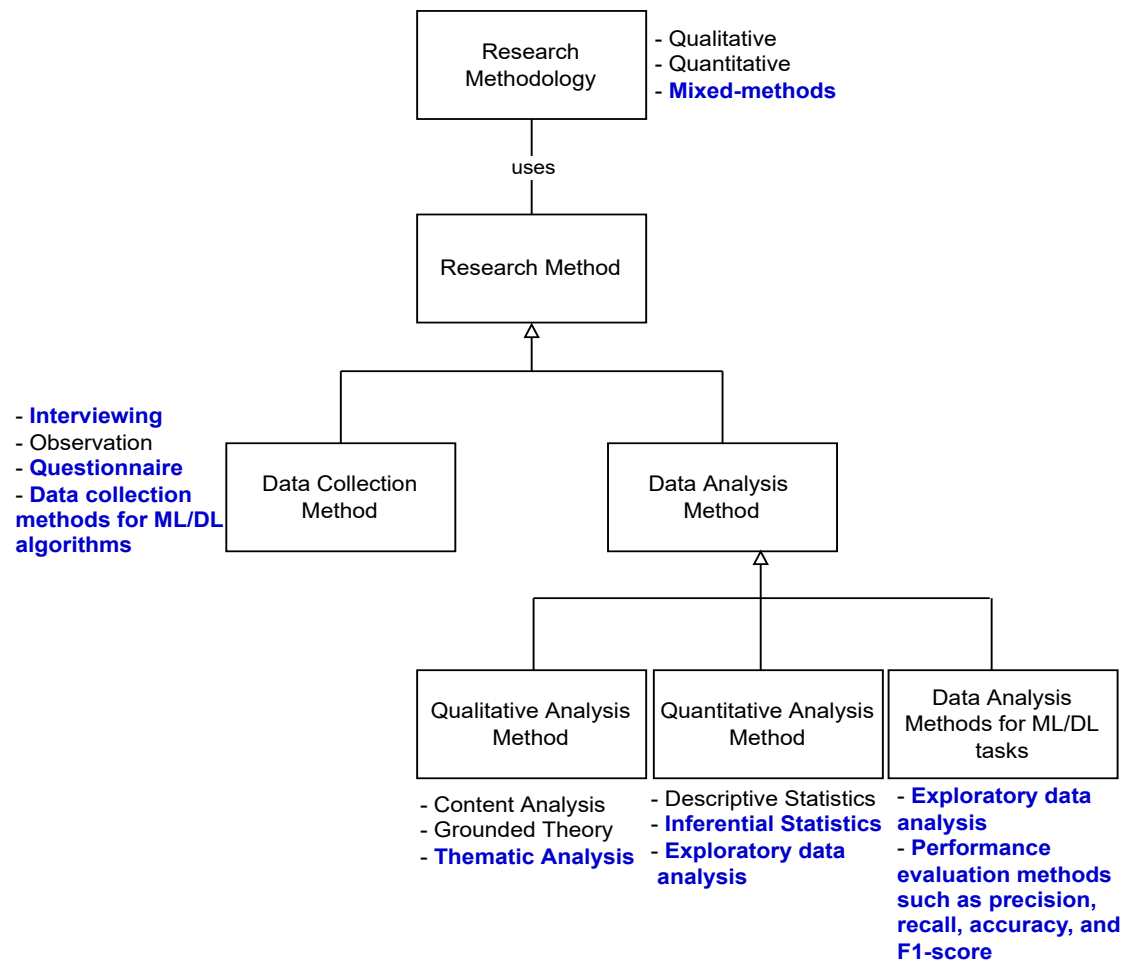
There are different methods to analyze qualitative interview data (e.g., content analysis, grounded theory method, and thematic analysis [246]). Content analysis is used to quantify the contents, which is done by categorizing elements of the text and then calculating the frequencies of the items in each category. Alternatively, in grounded theory, the researcher analyses data through coding and categorization. The researcher begins by labeling data based on their content (*open coding*), then identifies the most important codes and suggests categories into which the codes can be grouped (*axial coding*), and finally moves on to *selective coding*, where they focus on the main codes and categories and identify relationships between these [246].

Thematic analysis has been used mostly in qualitative research for identifying, analyzing, and reporting patterns of meaning (themes) within data [52]. Braun et al. [51] suggest a six-phase guide for conducting thematic analysis as follows.

- Step 1: Become familiar with the data,
- Step 2: Generate initial codes,
- Step 3: Search for themes,
- Step 4: Review themes,
- Step 5: Define and name themes,
- Step 6: Produce report

This study required qualitative interview data to gather requirements and evaluate artefacts. Therefore, there was no need to conduct a content analysis of interview data. Furthermore, the grounded theory approach was unsuitable for achieving interview data collection objectives. Instead, thematic analysis was best suited for analyzing interview data for requirements identification and artefact evaluation, as it allowed the researcher to identify main themes in the interview scripts, which can then be mapped for requirements or evaluation feedback.

The relationships between research methodologies, data collection methods, and data analysis methods discussed in previous sections are illustrated in Figure 3.4. The selected methodologies and methods to undertake this study are highlighted in blue letters.



**Figure 3.4** The relationships between research methodologies and methods. The selected methodologies and methods are highlighted in blue letters.

### 3.5 Chapter Summary

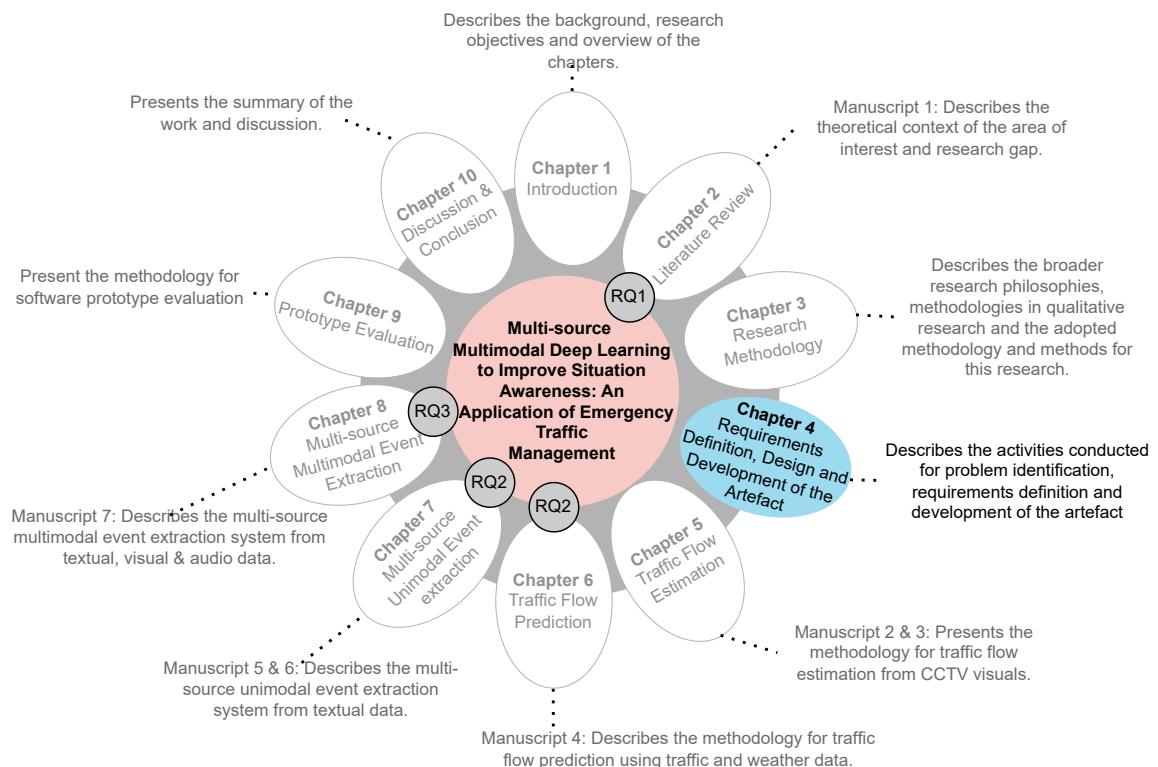
This chapter presented the philosophical view, design science research broader research framework, and a discussion on research methodologies, data collection, and analysis methods. The five-stage method framework proposed by Johannesson et al. [175] was considered to be followed through the research for problem explicating, requirement elicitation, design and development of artefact and demonstrating and evaluating artefact according to DSR. Appropriate methodological choices and research methods selection were also discussed.

The use of Johannesson's approach framework to Explicate the Problem, Define Requirements, and Design and Develop Artefact for this doctoral research is covered in the next chapter.



## Chapter 4

# Requirements Definition, Design and Development of the Artefact



Chapter 2 detailed the five-stage DSR method framework introduced by Johannesson et al. [175]. This chapter covers how the first three activities, Explicate Problem, Define Requirements, and Design and Develop Artefact was conducted in this PhD research and the results obtained in Sections 4.1 to 4.3.

### 4.1 Explicate Problem

The first activity in the artefact design and development process in DSR is to explicate the problem [385]. This task aims to clearly state the initial problem, demonstrate its importance, and investigate its underlying causes. Therefore, it addresses the question, “What is the problem experienced by some stakeholders of a practice, and why is it important?” [175]. A problem is

generally known as a gap between the desired and present states. Johannesson et al. [175] identify that there are three sub-activities in DSR to be addressed during the problem elicitation, including defining the problem precisely, positioning and justifying the problem, and finding its root causes. Initial scoping interviews were performed to gain an understanding of the problems associated with using SM data for emergency management in New Zealand. Therefore, five interviews were conducted with the most appropriate officials from city councils and emergency management offices (e.g., 2 - city council controllers, 2 - group controllers at regional emergency management offices, and 1 - emergency communications coordinator). A low-risk ethics clearance was obtained before the interviews, and the ethics clearance notification is provided in Appendix F. The interviews were transcribed and thematically analyzed. Section 3.4.2 describes the process of thematic analysis. As a result, the initial problem was formulated as “There is no proper system for utilizing social media data for real-time situation awareness of emergency management in New Zealand”.

The second sub-activity involves positioning the problem and providing justification by placing it in the practice. During the scoping interviews, it was found that the disaster response in New Zealand includes various governmental bodies covering many emergencies such as flooding, earthquakes, and landslides. Addressing such a vague and broad problem identified above was difficult and even unfeasible, and the problem needed to be narrowed down to something more manageable and testable. Therefore, a second round of scoping interviews was conducted with multiple responders from various government agencies including officers from traffic operations centers, met service, and city councils. After conducting another five interviews, referring to literature, and analyzing available documentation few areas of disaster management were identified to be explored. However, one of the main concerns for this project was the data accessibility for algorithm development. Traffic-related data were freely available from API and additional data was easily accessible for research through TOCs after signing agreements. Therefore, the scope was confined to emergency traffic management and, a new problem was formulated: “*No proper system for utilizing social media data for real-time situation-awareness while dealing with traffic emergencies in New Zealand*”. This problem is clearly significant for the responders of traffic emergencies. Moreover, the general public has a direct impact if traffic emergencies are handled efficiently.

The final sub-activity is to identify the root causes of the problem to get a more in-depth understanding. As a result, the underlying causes of the problem are identified, analyzed, and represented. Therefore, the academic literature on disaster response was reviewed, focusing on real-time event extraction and deep learning. In addition to the literature study, the scoping interviews conducted were used to identify root causes. The following are the root causes identified.

“...All the information comes in all sorts of places. For example, ground truth information from police, community hub, smartphone apps, social media, cell phones, emails, text messages etc...”

- During an emergency event, information flows from multiple channels such as call centres, SM, and other governmental bodies [12, 313, 411].

“...If there’s an event, lots of information flows in here from all social media forms, we will see that on Facebook, someone has posted it’s got the trees down in this street. What is the reality? Is that true? How can we know?...”

- Even though the responding organizations receive a significant amount of information through SM channels, they have concerns about the reliability of the information [163].

“...All this information comes in, and it should get filtered. If not, how long it will take to know which information we should focus on? How long to wait and do nothing while someone is hurt or die?...”

- The responders are unable to process this information in real-time, and they feel information-overloaded [308, 61].

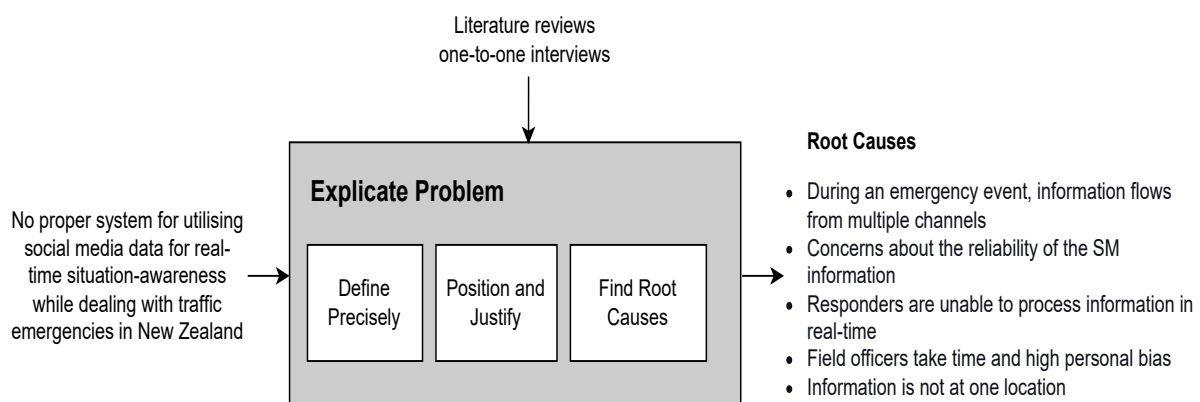
“...I sent two teams out there; one was the council report team, and the other was the general response team...volunteers, there were out there at different times, half an hour apart and gave quite different reports after 2 hours. One team said there was no way to come down, and the other team reported Oh my god, it’s going to come down, so where does the truth lie?...”

- Field officers are sent to the event locations; however, it takes a considerable amount of time for them to prepare reports and send them back for response activities and there is a high personal bias.

“...Information is here and there...inputs into these phones, text messages, and e-mails, both from outside and also within. We have an internal e-mail system, radio, smart-phone apps, and online portals. When it comes to briefings, someone has to run here and there to collect everything into one place. Currently, there is no way we can easily see the common operating picture...”

- As information is dispersed, responders struggle to collate them for decision-making and official briefings

After having considered the outcomes of the above steps, the Explicate Problem activity can be summarised as in Figure 4.1.



**Figure 4.1** Explicate Problem activity

## 4.2 Define Requirements

To bridge the gap between problem and solution, a researcher needs a clear and deep understanding of the problem being addressed that can be achieved through defining requirements [385]. A requirement is a statement made by a stakeholder of a practice that can guide a researcher in further designing and developing the solution. Moreover, the requirements can also be used to validate the quality of the developed artefact [385, 175]. Therefore the Define Requirements activity addresses the problem, “What artefact can be a solution for the explicated problem and which requirements for this artefact are important to the stakeholders?”

Johannesson et al. [175] suggest that the requirement definition has to be achieved in two stages: outlining the artefact and eliciting requirements. The first stage is to decide on the type of artefact to be designed and its basic characteristics. There are multiple types of computer applications such as stand-alone, web-based and mobile. Among them, the web-based application was considered as best suited for this project due to several factors. First, a web-based system is less dependant on hardware at operational centres. Second, a web-based system can run on multiple devices such as laptops, standard personal computers, tablets or mobile devices. Therefore, a single implementation allows the use of the system through many devices. Third, any number of users from any location can consume the system without any installations. However, the system required multiple DL algorithms to be trained and tested to automate the tasks. The training and testing of deep learning algorithms require extensive time as raw datasets have to be collected, preprocessed, labelled, and analysed. Moreover, the algorithms have to be fine-tuned, and this has to be done iteratively multiple times. The total duration for the design and development of the system was around one year, limiting the time available for development. Moreover, some components of the system required special computing facilities such as Graphical Processing Units (GPUs) that prevented the development of a live web-based system. In light of these considerations, it was decided to create a prototype of the live system capable of demonstrating its full functionality.

The relevant academic literature was thoroughly examined to elicit the requirements. Moreover, in-depth interviews were conducted with 11 experts, as presented in Table 4.1.

**Table 4.1** Background of participants selected for the interviews.

Organisation	Experience
Wellington city council	8 years experience as a city council member
Wellington Region Emergency Management Office (WREMO)	2 years experience as the group controller
WREMO	6 years experience at WREMO
Hutt City Council	30+ years experience in civil defence
Hutt City Council	14 years experience as a city council member
Hutt City Council	2 years experience as the Science & Technology Manager
Christchurch City Council	5 years experience as a city council member

Wellington Transport Operations Centre	15 years experience in NZTA
Kestrel group	30+ years experience in building, infrastructure, emergency management
Waikato Regional Council	10 years experience in the Waikato council
Ministry of Civil Defence & Emergency Management (MCDEM)	30+ years experience in civil defence

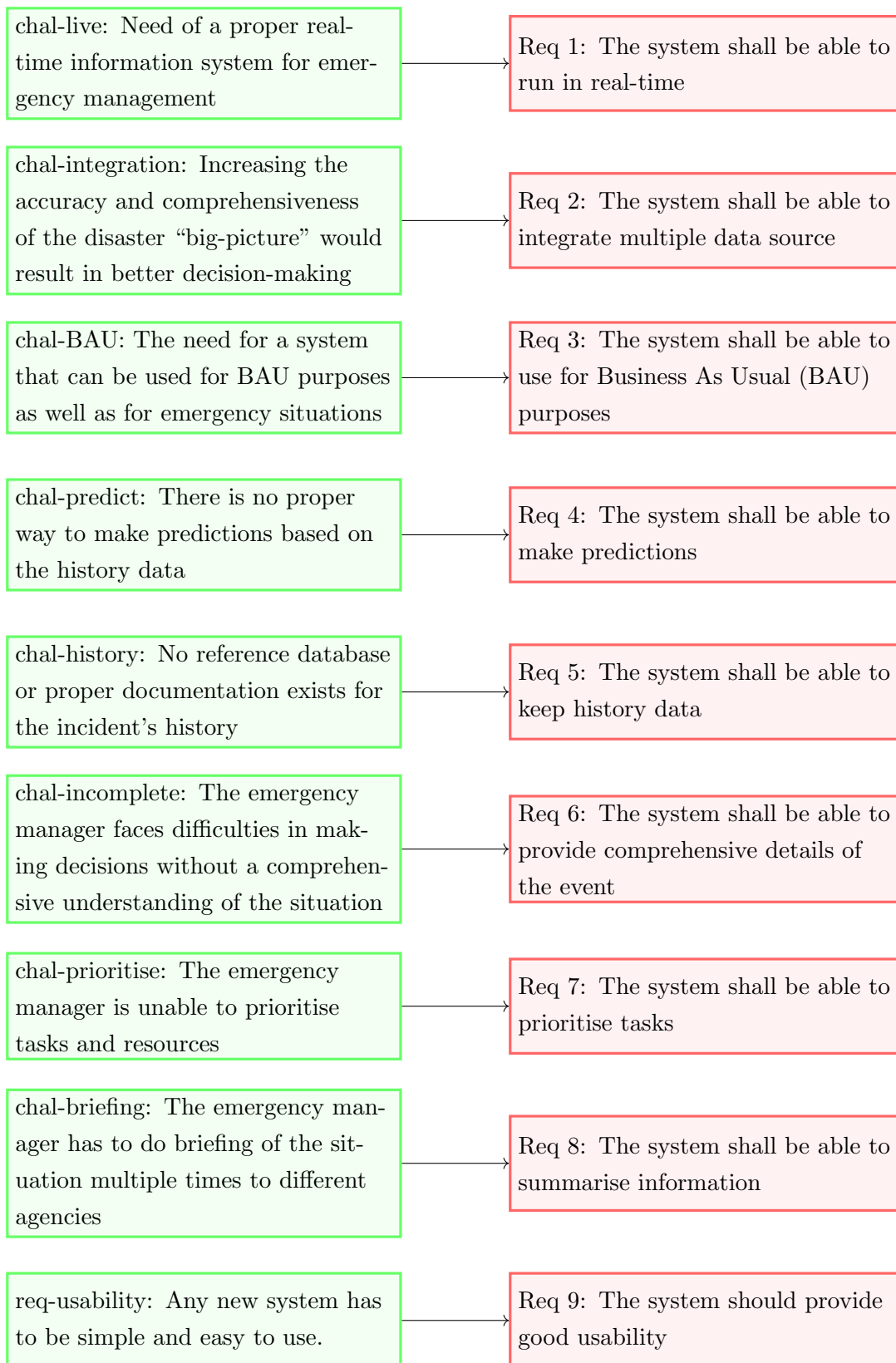
Interviews were transcribed and analysed using thematic analysis. During the thematic analysis, the interview data were transcribed and coded to identify features, meaningful themes were constructed from the coded data, and the themes were finally reviewed to determine the most important themes [51]. The results of this analysis process are illustrated in Table 4.2.

**Table 4.2** Findings from interview data.

Code	Theme	Theme Description	Example Extraction
chal-systems	Need of a proper information system for emergency management	There is no proper system specifically designed to manage an emergency, except for some common file sharing and email tools	“at the moment we actually don’t have a proper Emergency Management Information System, we are kind of relying on Email, Outlook, Excel, Word”
			“at the moment we do that by email, or else emails are not working just by phone, if that's not working by radar. So, now we have no sophisticated systems”
chal-history	Unavailability of history data	The history data is with the individuals involved in managing the incident, and there is no reference database or any proper documentation	“how we respond to an event is pretty much kept in the heads of the individuals that were involved in the last one”
			“It’s really been exhausting for the staff, but more importantly, how well we’ve kept an audit trail of our actions.”
chal-briefing	Multiple briefings to different agencies	The emergency manager has to do briefing of the situation multiple times to different agencies	“When our liaison agencies come in, like Police, Fire, Ambulance, the Downers, NZTA, they can come in and get a really good situational understanding of what’s going on without having to pull me up all the time and say right oh controller, brief me on what’s going on.”
			“We do a lot of briefings, and it means that I’m constantly being taken out of the picture”
chal-progress	Not able to track the status of tasks	The manager is unable to track the progress of tasks assigned to different individuals	“I also want to know, what’s the outcome of your actions? How does that get fed back into the bigger picture so that I don’t have to keep coming over and asking you has that been done?”
chal-prioritise	Not able to prioritise the tasks	The emergency manager is unable to prioritise tasks and resources.	“So, a message comes in two buildings have been collapsed, the people trapped, but you have resources to tackle one at a time. Which one to get through first?”

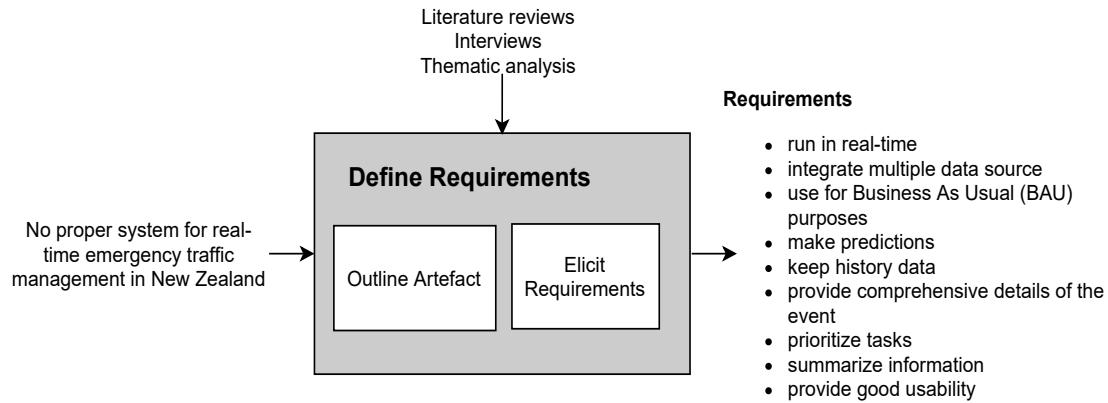
			“we’ve seen a whole range of events come in, you know, requests from the public, it would be nice to be able to map them, prioritise them, and have it actually”
chal-incomplete	Not able to see the disaster “big picture”	The emergency manager faces difficulties in making decisions without a comprehensive understanding of the situation	“So, at the end, to make a decision, whether do something based on incomplete, possibly inaccurate, non-helpful information or do nothing until better information comes in. And then, how long that is going to take? How long to wait and do nothing while someone being hurt or die?”
chal-predict	Not able to get any predictions	There is no proper way to make predictions based on the history data	“I was saying to my team, can someone tell me whether or not there is going to be a tsunami? But no one could”
chal-bias	Human bias in inference	Human cognitive bias in the preparation of incident reports	“We had a flooding event Wainuiomata, we sent two teams out there, half an hour apart and gave quite different reports one team said there is no way to come down, the other team reported Oh, it’s going to come down, so where does the truth lie?”
req-BAU	Need for a new technical solution which can be used for BAU purposes	The need for a system that can be used for BAU purposes as well as for emergency situations	“If you do something only for emergencies, that’ll be rarely used, you can target something which is useful for day to day life, and also that can be used in emergency situations.”
			“it needs to be an information system that can be used on an almost a business as a usual basis”
req-integration	Need for a new technical solution which is able to integrate data	Increasing the accuracy and comprehensiveness of the disaster “big-picture” would result in better decision-making	“The better, more accurate and more complete, the picture on which your basing decision making, then the better the decisions are made.”
			“So that any decision I make is based on what is actually happening not what we think is happening which necessary the same thing at all”
req-usability	The need for the system to be simple	Any new system has to be simple and easy to use.	“we are working with emergencies, so it has to be simple, very simple...”

The themes identified in Table 4.2 led to several specific end-user requirements, as follows.





The Define Requirements activity is summarised in Figure 4.2.



**Figure 4.2** Define Requirements activity

### 4.3 Design and Develop Artefact

The most important part of a design science project is the activity Design and Develop Artefact, which can answer the explicated problem and fulfil the specified requirements described in Section 4.2. The activity can be divided into four sub-activities, which are carried out in parallel and iteratively, such as Imagine and Brainstorm, Assess and Select, Sketch and Build and Justify and Reflect [175].

The first activity Imagine and Brainstorm is used to generate new ideas or to further enhance existing ones. Therefore, a clean sheet of paper was used to generate multiple alternative ideas or solutions to address the problem. The following are some of the ideas generated during the brainstorming.

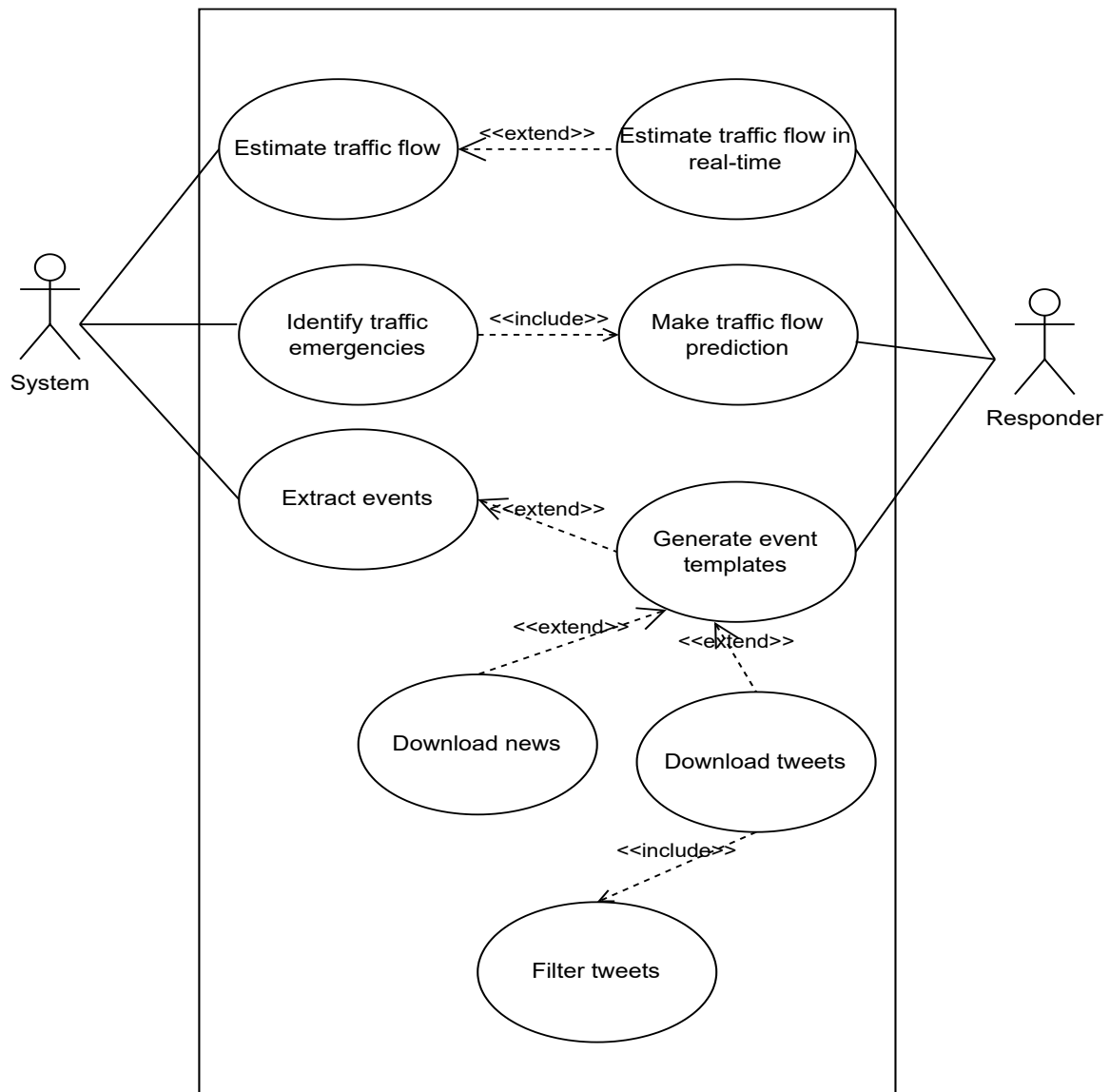
- Study traffic flow patterns in New Zealand
- Study traffic emergencies and how to identify them in real-time
- Study what is available in SM regarding traffic emergencies
- Identify ways that SM can help during traffic emergencies
- Identify ways to validate SM content
- Identify ways to provide real-time updates to responders

The next sub-activity, Assess and Select, entails selecting one or more of the brainstormed ideas as the foundation for developing the artefact. The decisions on sufficiently good ideas were chosen by evaluating them in depth. Therefore, the following ideas were considered during this step.

1. Study traffic flow patterns in New Zealand
2. Study traffic emergencies and how to identify them in real-time
3. Identify ways to validate SM content

## 4. Identify ways to provide real-time updates to responders

In the sub-activity Sketch and Build, a sketch of the artefact is made starting from the ideas selected in the previous sub-activity. Figure 4.3 illustrates the design sketch of the artefact generated based on the ideas.



**Figure 4.3** Design sketch of the artefact

Based on this sketch, the development of the outlined artefact was conducted in four stages carried out individually and iteratively, as follows.

- **Component 01: Real-time traffic flow counting from CCTV visual data:**

The development of this component was motivated by the selected brainstormed idea 1. Traffic flow data were identified as the key input for the software artefact. However, a continuous traffic flow in New Zealand could not be found from official sources. As a result, during the first stage, an algorithm was developed to estimate traffic flow from CCTV images obtained through the NZTA Traffic Cameras API. More details of the design and development of this component are discussed in Chapter 5.

Then the researcher used CCTV footage to obtain live traffic flow data based on the vehicle class and movement direction. This approach is discussed in Chapter 5.

- **Component 02: Short-term traffic flow prediction:**

In the second component, an algorithm was developed to obtain short-term traffic flow prediction. More details of the design and development of this component are discussed in Chapter 6.

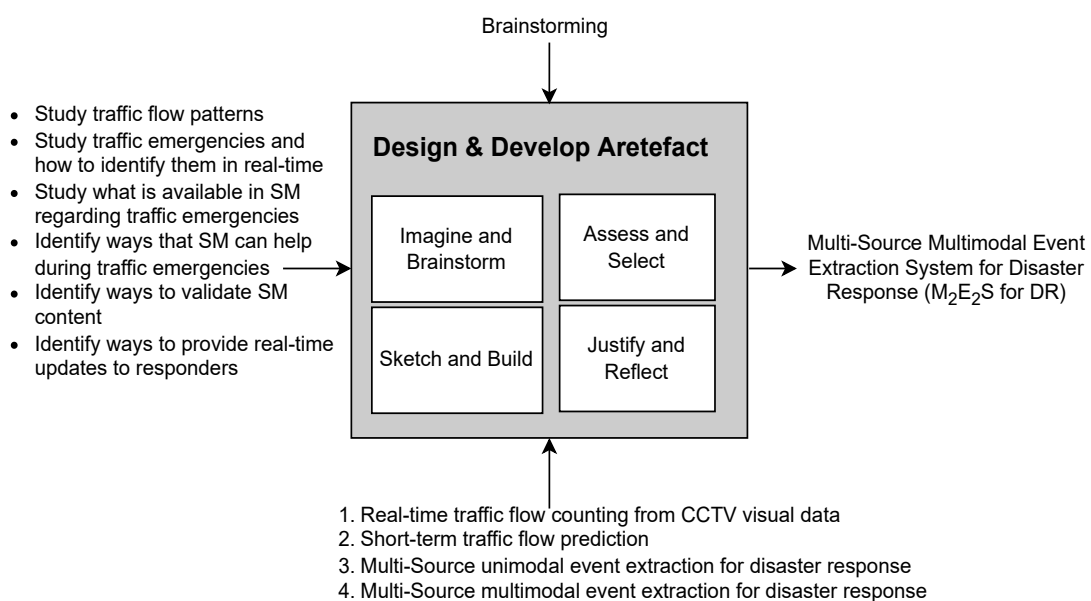
The final implemented system was expected to automatically identify event details when the system detects a traffic anomaly which was inspired by selected brainstormed idea 2. The traffic anomaly was supposed to be detected if there was a deviation between live traffic flow and traffic prediction. However, this component will be developed in future research due to the time constraints and difficulties in obtaining historical data.

- **Component 03: Multi-Source unimodal event extraction for disaster response:**

The third component was developed considering the selected brainstormed ideas 3 and 4. First, we studied the use of online news text data to validate SM content. Event templates were generated using text data extracted from SM and online news to answer what, where and when questions. The architecture and development process of this component is further discussed in Chapter 7.

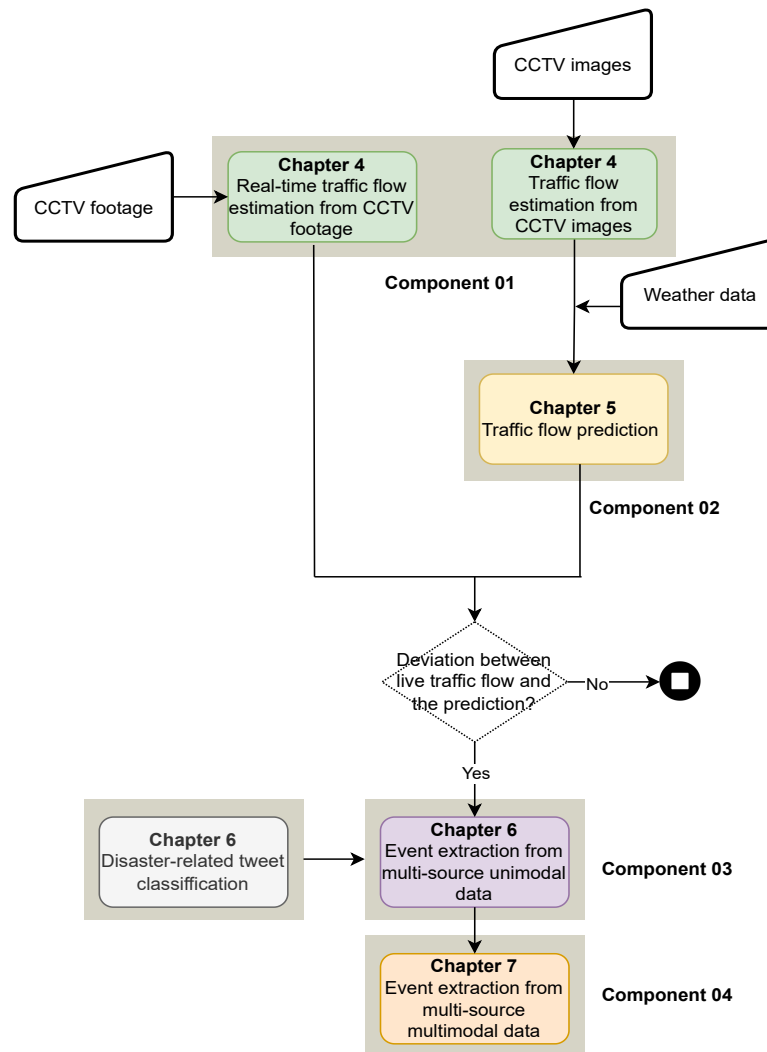
- **Component 04: Multi-Source multimodal event extraction for disaster response:**

The final component developed was to extract real-time event templates from multi-source multimodal data based on the selected brainstormed ideas 3 and 4. Therefore, the final outcome of the research was named as Multi-Source Multimodal Event Extraction System for Disaster Response ( $M_2E_2S$  for DR). The system used SM and online news and considered text, visual and audio modalities while extracting event templates. This process is discussed further in Chapter 8.



**Figure 4.4** Design and Development activity

Figure 4.4 summarises the Design and Development activity and Figure 4.5 shows the relationships between each component of the artefact.



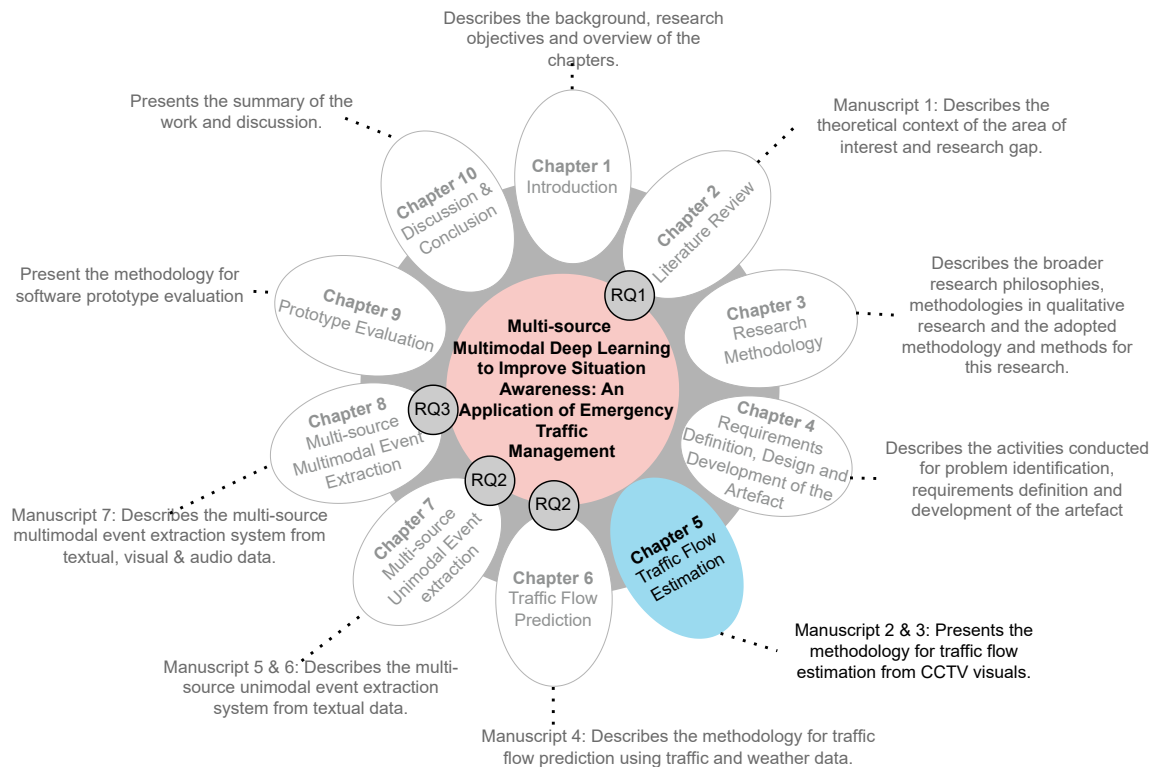
**Figure 4.5** The relationship between algorithms developed in Chapters 5 - 8

## 4.4 Chapter Summary

This chapter discussed the application of the first three stages of the design science method framework proposed by Johannesson et al. [175]. The first activity, Explicate Problem was used to define the broader problem, position the problem and provide justification by placing it in practice and identifying the root causes. During the second activity, the artefact was outlined, and requirements were identified. For this purpose, multiple interviews were conducted with experts from city councils, traffic operation centres and emergency management groups in New Zealand. The third activity was to design and develop the artefact. Therefore, ideas were brainstormed, selected best ideas were made, and created a design sketch. The development of the artefact was conducted in four stages, and this process is discussed in the following four chapters (Chapter 5, 6, 7 and 8).

## Chapter 5

# Traffic Flow Estimation from CCTV Data



This chapter describes the design, development, and validation of the first component of the software artefact, which is the estimation of traffic flow using CCTV data. The process is outlined in two conference papers, the first in Section 5 describing traffic flow estimation from CCTV images and the second in Section 5.5 describing traffic flow estimation from CCTV footage.

## Manuscript 2: Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images

The following article is published as: Nilani Algiriyage, Raj Prasanna, Emma E H Doyle, Kristin Stock, & David Johnston. (2020). Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images. In Amanda Hughes, Fiona McNeill, & Christopher W.

Zobel (Eds.), ISCRAM 2020 Conference Proceedings – 17th International Conference on Information Systems for Crisis Response and Management (pp. 100–109). Blacksburg, VA (USA): Virginia Tech.

## Abstract

Emergency Traffic Management (ETM) is one of the main problems in smart urban cities. This paper focuses on selecting an appropriate object detection model for identifying and counting vehicles from closed-circuit television (CCTV) images and then estimating traffic flow as the first step in a broader project. Therefore, a case is selected at one of the busiest roads in Christchurch, New Zealand. Two experiments were conducted in this research; 1) to evaluate the accuracy and speed of three famous object detection models namely faster R-CNN, mask R-CNN and YOLOv3 for the data set, 2) to estimate the traffic flow by counting the number of vehicles in each of the four classes such as car, bus, truck and motorcycle. A simple Region of Interest (ROI) heuristic algorithm is used to classify vehicle movement direction such as “left-lane” and “right-lane”. This paper presents the early results and discusses the next steps.

## 5.1 Introduction

Traffic flow estimation is important for urban planning and management of road traffic infrastructure. It is also essential to have a good understanding of the flow of traffic in order to manage emergencies (e.g., to re-route traffic through alternative routes) [413]. Researchers and developers have recently focused on deep neural networks for traffic prediction, particularly in smart urban cities [236, 183]. Emergency Traffic Management (ETM) can be described as a specific case of traffic management requiring extensive planning to ensure secure and effective egress. The causes of traffic emergencies can be small-scale (e.g., vehicle crash) or large-scale (e.g., earthquake or tsunami). They can also be planned (e.g., scheduled maintenance, noticed evacuation before a disaster) or unplanned. Wrong decisions made without a clear picture of the situation have led to multiple unfortunate incidents resulting in dozens of human casualties during mass evacuations [60, 135].

Identification of traffic flow is the first step in consolidated planning of managing traffic emergencies. Today, closed-circuit television (CCTV) systems are extremely common and mounted in many public areas to support real-time monitoring. As they are operated continuously, they generate a massive amount of data contributing to big data. Among the other types of sources, CCTV data can be used as the foundation for accurate traffic flow estimation [104, 296].

The majority of recent research use state-of-the-art deep learning based object detection frameworks such as Faster R-CNN [322], YOLO [321] and mask-RCNN [138] for vehicle detection and tracking [209, 420, 416, 419] from image data sets. Faster R-CNN and mask-RCNN belong to the R-CNN family networks that use regions to locate the objects within an image. In comparison, the YOLO algorithm divides the entire image into cells and predicts the bounding boxes and class probabilities. However, traffic flow estimation using these algorithms for surveillance camera data sets is still in very early development. Difficulties in moving, storing and developing efficient, intelligent algorithms for processing and analyzing CCTV big data have been identified as major challenges. [104].

The study presented in this paper seeks to answer the questions; 1) What object detection algorithm is best suited to the CCTV image data set for vehicle detection? 2) Can traffic flow be estimated by counting the number of vehicles in CCTV images using an object detection algorithm?. Therefore, we collect real-time CCTV imagery from traffic cameras through the New Zealand Transport Agency's (NZTA) traffic cameras Application Programming Interface (API)<sup>1</sup>. During the first experiment, we compare the performance and accuracy of faster R-CNN, Mask R-CNN and YOLOv3 algorithms in vehicle detection for the CCTV image data set. Then, as a case study, we focus on one of the busiest roads in Christchurch Central Business District (CBD) to estimate the traffic flow. However, broader research extending this work would use the estimated flow of traffic to predict the short-term traffic flow. The results of this research, along with the prediction system can be used by city authorities to understand traffic flow patterns, predict traffic flow at a given time, understand traffic anomalies, and make management decisions.

The rest of our paper is outlined as follows. The next section reviews the existing work. Then we illustrate our methodology of the study and discuss the results. Finally, we present concluding remarks and future research steps.

## 5.2 Related Work

One of the first steps in traffic flow estimation is vehicle identification. Object detection differs from classification as it attempts to draw a bounding box around the object of interest in order to locate it within the image. In computer vision research, there are three primary object detectors [321, 322]:

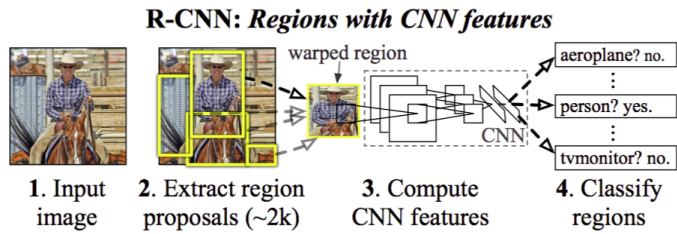
- R-CNN and their variants, including the original R-CNN, Fast R-CNN, Faster R-CNN and Mask R-CNN
- Single Shot Detector (SSDs)
- YOLO

These Convolutional Neural Networks (CNN) based object detectors can be roughly divided into two main categories: single-stage detectors and two-stage detectors. The single-stage detectors are generally fast and predict object bounding boxes together with classes within a single network pass (e.g., SSDs and YOLO) [321, 227]. Comparatively, two-staged detectors such as R-CNN family networks detection happens in two-stages; 1) the model proposes a set of regions of interests by selective search [374] or using Regional Proposal Network (RPN) 2) a classifier only processes the region candidates to identify the objects [116, 115, 322, 138]. Therefore, two-stage detection tends to be slow. Huang et al. [151] provide a thorough review of the key advantages and disadvantages of single and two-stage detectors.

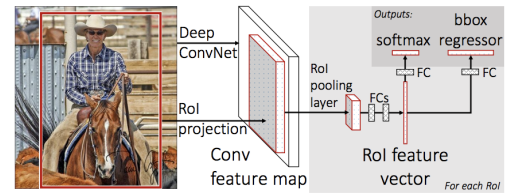
The use of CNNs to identify objects in all regions of an image has been computationally inefficient [423]. Therefore, Girshick et al. [116] proposed the first Regions with CNN features (R-CNN) algorithm to use selective search [374] to extract just 2000 regions from images to propose candidate bounding boxes that could contain objects. The identified regions were then passed into a CNN for classification, which eventually led to one of the first deep learning-based object detectors.

---

<sup>1</sup>NZTA Traffic Cameras API . Retrieved January 5, 2020, from <https://www.nzta.govt.nz/traffic-and-travel-information/infoconnect-section-page/about-the-apis/traffic-cameras/>



**Figure 5.1** R-CNN architecture [116].



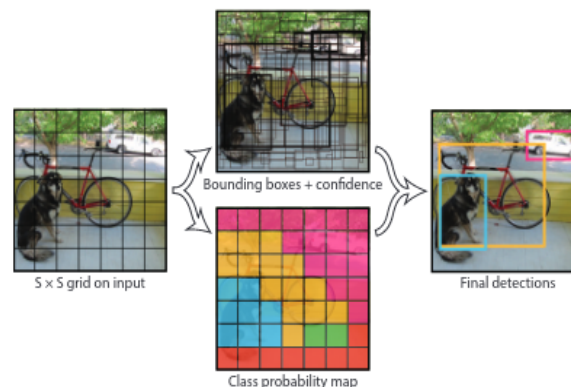
**Figure 5.2** Fast R-CNN architecture [115].

However, the R-CNN algorithm was painfully slow and could not be used for real-time detection. Therefore, Girshick et al. [115] improved R-CNN and published a second paper in 2015, entitled Fast R-CNN. The Fast R-CNN algorithm made significant improvements to the original R-CNN, namely by increasing accuracy and reducing the time it took to perform the forward pass. However, the Fast-RCNN model still relied on an external region proposal algorithm (see Figures 5.1 and 5.2).

In 2015, the follow-up paper by Ren et al. [322] introduced Faster R-CNN as a true end-to-end detector of deep learning artefacts. It has been improved by removing the selective search requirement and instead relying on RPN that is fully convolutional and can predict the object bounding boxes and “objectness” scores (e.g., a score that quantifies the probability of an image region may contain an object). The outputs from the RPNs were then passed into the R-CNN component for final classification and labelling.

Kaiming et al. [138] extended faster R-CNN by proposing instance segmentation rather than drawing bounding boxes, which resulted in proposing mask R-CNN. It is considered as a flexible and efficient framework compared to other R-CNN family networks [423].

The most significant problem with the R-CNN family of networks was their speed as they were incredibly slow, even faster R-CNN obtaining only 5 FPS (Frame Per Second) on a Graphical Processing Unit (GPU) [322]. Both SSDs and YOLO use a one-stage detector strategy to help increase the speed of deep learning object detectors. One-stage detectors treat object detection as a regression problem, taking a given input image and simultaneously learning bounding box coordinates and corresponding class label probabilities. Generally, single-stage detectors tend to be less accurate than two-stage detectors, but are significantly faster [151].



**Figure 5.3** YOLO object detection [321].

You Only Look Once (YOLO) was first introduced in 2015 by Redmon et al. [321] as an object detector capable of real-time object detection, obtaining 45 FPS on a GPU (see Figure 5.3). YOLO



has gone through several different iterations until they introduced YOLO9000 [319]. They were able to achieve such a large number of object detection by performing joint training for both object detection and classification. The authors simultaneously trained YOLO9000 on both the ImageNet classification data set and the COCO detection data set using joint training. However, as the performance was not satisfactory, they recently introduced YOLOv3 [320], which is significantly larger than previous models and with greater accuracy. Table 5.1 provides the details of popular object detection models, their objectives and links to source codes.

Algorithm	Author	Objective	Code
R-CNN	[116]	Object detection	<a href="https://github.com/rbgirshick/rcnn">https://github.com/rbgirshick/rcnn</a>
fast R-CNN	[115]	Object detection	<a href="https://github.com/rbgirshick/fast-rcnn">https://github.com/rbgirshick/fast-rcnn</a>
faster CNN	R- [322]	Real-time Object detection	<a href="https://github.com/rbgirshick/py-faster-rcnn">https://github.com/rbgirshick/py-faster-rcnn</a>
mask CNN	R- [138]	Image segmentation	<a href="https://github.com/facebookresearch/Detectron">https://github.com/facebookresearch/Detectron</a>
YOLO, YOLO9000, YOLOv3	[321, 319, 320]	Real-time object detection	<a href="https://pjreddie.com/darknet/yolo/">https://pjreddie.com/darknet/yolo/</a>

**Table 5.1** Popular object detection models and their objectives

Multiple open-source visual data sets with manually labelled features have contributed to the advancement in computer vision research [111]. We use YOLOv3 and R-CNN networks trained on two such data sets, namely Common Objects in Context (COCO) [225] and Computational Learning Visual Object Classes (PASCAL VOC) [102]. COCO is an image data set introduced by Microsoft, consisting of 80 common objects in their natural context [225]. Training and testing data sets for PASCAL VOC consisted of 27,450 detection objects in 11,530 images of 20 different classes. Also, the test and training data sets of PASCAL VOC segmentation consist of 6929 segmented objects in 11,530 images [111, 102].

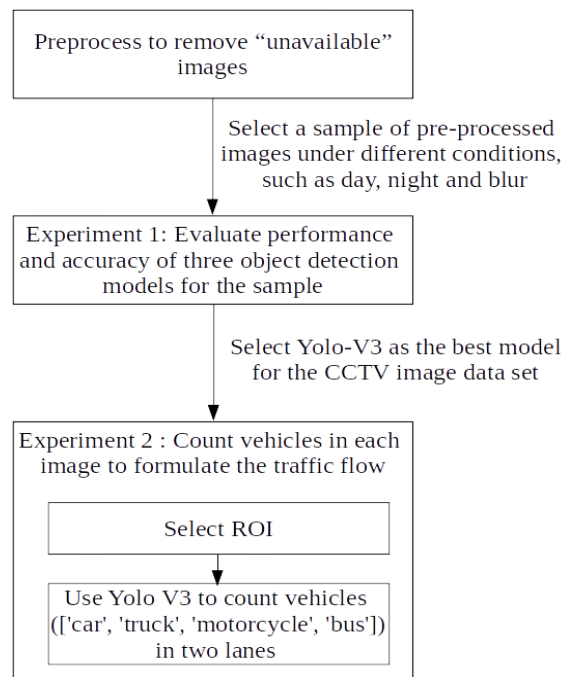
The *traffic flow estimation* is identifying the number of vehicles during the  $t^{th}$  time interval at the  $i^{th}$  observation location in a transportation network which can be denoted as  $X_i^t$ . Therefore, the *traffic flow prediction* problem can be stated as follows: Given  $X_i^t$  denote the observed traffic flow during the  $t^{th}$  time interval at the  $i^{th}$  observation location,  $t = 1, 2, \dots, T$  and  $i = 1, 2, \dots, m$ , the problem is to predict the traffic flow at time interval  $(t + \Delta)$  for some prediction horizon  $\Delta$  [236]. Most of the traffic-related research using Deep Learning techniques have focused mainly on the problem of traffic flow prediction [407, 303, 205]. However, these studies are different from the scope of this paper. The main problem we address is the traffic flow estimation from CCTV image series as discussed by Fedorov et al. [104], a subject area in very early development. Therefore, there are a minimal number of researches which address the same problem.

Previously, inductive loop detectors, pneumatic road tubes, and temporary manual counts were the primary methods for estimating traffic flow [42]. However, depending on the costs and difficulties of installation, these methods can not be used in large areas. As highlighted in the introduction, CCTV monitoring is currently very common and these large networks are barely used except for the investigation of incidents and anti-social behaviour [42]. Previously, due to privacy concerns, only the police and city councils used to access these data. However, the current trend of most

city councils is to put their CCTV data sets for open access. Fedorov et al. [104] have used Faster R-CNN for a video data set to identify traffic flow. However, only a short video clip containing 982 frames has been considered for their research. A similar study has evaluated the accuracy of two deep learning algorithms, namely MobileNet, and faster R-CNN trained on COCO data set [296]. They show that the accuracy of faster R-CNN is more for their CCTV image data set. However, the direction of vehicle movement is not considered during the flow estimation process. Taking advantage of the open access CCTV image series and considering the research gap, we focused on developing a method for estimating traffic flow.

### 5.3 Methodology

The study presented in this paper evaluates the performance of three Deep Learning algorithms and estimates the flow of traffic from CCTV images. Therefore, in the first experiment, we evaluate the performance and accuracy of three of the most popular object detection algorithms, such as faster R-CNN, mask R-CNN and YOLOv3. We used a faster R-CNN model trained on PASCAL VOC data set with ResNet-50 backbone [137] and mask R-CNN model trained on COCO data set with ResNet-50 backbone, which is implemented in GluonCV<sup>2</sup>. Also, YOLOv3 implementation in cvlib python library<sup>3</sup> trained on COCO data set. During the second experiment, we apply YOLOv3 to formulate the traffic flow by counting the number of vehicles in each minute. The flow of methodology is shown in Figure 5.4.



**Figure 5.4** Methodology.

<sup>2</sup>GluonCV. Retrieved January 5, 2020, from <https://gluon-cv.mxnet.io/index.html>

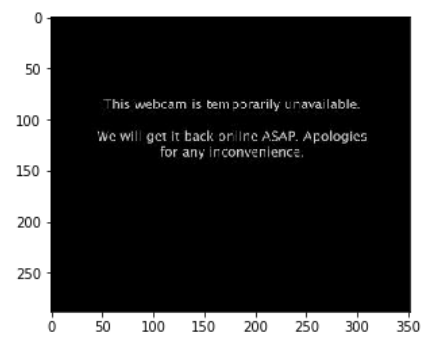
<sup>3</sup>cvlib. Retrieved January 5, 2020, from <https://www.cvlib.net/>

### 5.3.1 Data set

A CCTV image data set was formulated by collecting traffic camera images in real-time through the NZTA traffic cameras API from 10<sup>th</sup> of October to 31<sup>st</sup> of October 2019 in Christchurch CBD. There are 83 cameras operated in the CBD, and for this experiment, we selected a busiest road namely "West along Yaldhurst Rd from Curletts Rd" (latitude -43.53074, longitude 172.56812). The total size of the data set we selected for the experiments is 1.6 GB. The images are low resolution, taken from different angles, in different illumination levels and also under different weather conditions. Each image has a height of 600 pixels and a width of 800 pixels. A sample image is shown in Figure 5.5. Occasionally, cameras create an "unavailable" image with a message due to the technical faults (see Figure 5.6).



**Figure 5.5** A sample CCTV image



**Figure 5.6** Camera "unavailable" image

#### 5.3.1.1 Data processing

We found 6.3% of the data set considered for the experiment as "unavailable images" (see Table 5.2). Therefore, they were filtered out using a simple python script, considering the total pixel value.

Total number of images before pre-processing	24, 085
Total number of "unavailable images"	1519
Total number of images after pre-processing	22, 566

**Table 5.2** Data set before and after pre-processing

### 5.3.2 Experiments

#### 5.3.2.1 Experiment 1

We selected a sample 10% of the of pre-processed images under different conditions, such as day, night and blur. Then we evaluated the accuracy and performance of YOLOv3, mask R-CNN and faster R-CNN in the detection of vehicles in each of the classes such as ['car', 'truck', 'motorcycle', 'bus']. The performance was measured in terms of time taken to detect objects. According to Table 5.3, YOLOv3 has always achieved the highest performance by having the lowest time detect vehicles.

For the same sample, we evaluated the accuracy in terms of precision and recall (see Table 5.3). First, we manually counted the number of vehicles in each image and then used the faster

Model	Performance/ mean time taken to detect vehicles (seconds)	Recall	Precision
YOLOv3	0.86	0.79	0.96
faster R-CNN	8.37	0.50	0.96
mask R-CNN	55.6	0.69	0.77

**Table 5.3** Performance and accuracy of the three models for our CCTV data set

R-CNN, mask R-CNN and YOLOv3 models to count the number of vehicles (see Figure 5.7). Then precision and recall values were obtained. Precision is the number of True Positives (TP) over the number of predicted positives (PP), and recall is the number of true positives over the number of actual positives (AP).  $PP = TP + FalsePositives (FP)$  and  $AP = TP + FalseNegatives (FN)$  and therefore, Precision =  $\frac{TP}{PP} = \frac{TP}{TP+FP}$  and Recall =  $\frac{TP}{AP} = \frac{TP}{TP+FN}$ .

### 5.3.2.2 Experiment 2

Our next experiment was to estimate the traffic flow by counting the number of vehicles in each image. Traffic flow estimation has two tasks; 1) determining the direction of movement of the vehicle 2) counting the number of vehicles in each image by the direction of the vehicle movement. Therefore, to identify the “left-lane” and the “right-lane”, each image must be divided into two. We also wanted to avoid the parking area, and hence, the most appropriate region of interest was identified as a trapezium. Two trapeziums were selected having the sizes points  $[[100, 600],[250, 199],[450, 199],[800, 600],[100, 600]]$  as the “left-lane” and  $[[750,600],[450,200],[800,200],[800,600],[100,600]]$  as the “right-lane” (see Figure 5.8 and 5.9). The algorithm for selecting the trapezium is as follows:

---

#### Algorithm 1 ROI selection as a trapezium

---

```

1: for  $i \in I$  do
     $y\_size, x\_size = i.shape[: 2]$ 
     $vert\_coef = 0.3333$ 
     $hor\_coef = 0.312$ 
     $v\_coef = vert\_coef$ 
     $up\_left\_coef = hor\_coef$ 
     $up\_right\_coef = 1 - up\_left\_coef$ 
     $low\_left\_point = [0, y\_size]$ 
     $low\_right\_point = [x\_size, y\_size]$ 
     $up\_left\_point = [x\_size * up\_left\_coef, y\_size * v\_coef]$ 
     $up\_right\_point = [x\_size * up\_right\_coef, y\_size * v\_coef]$ 
2: end for

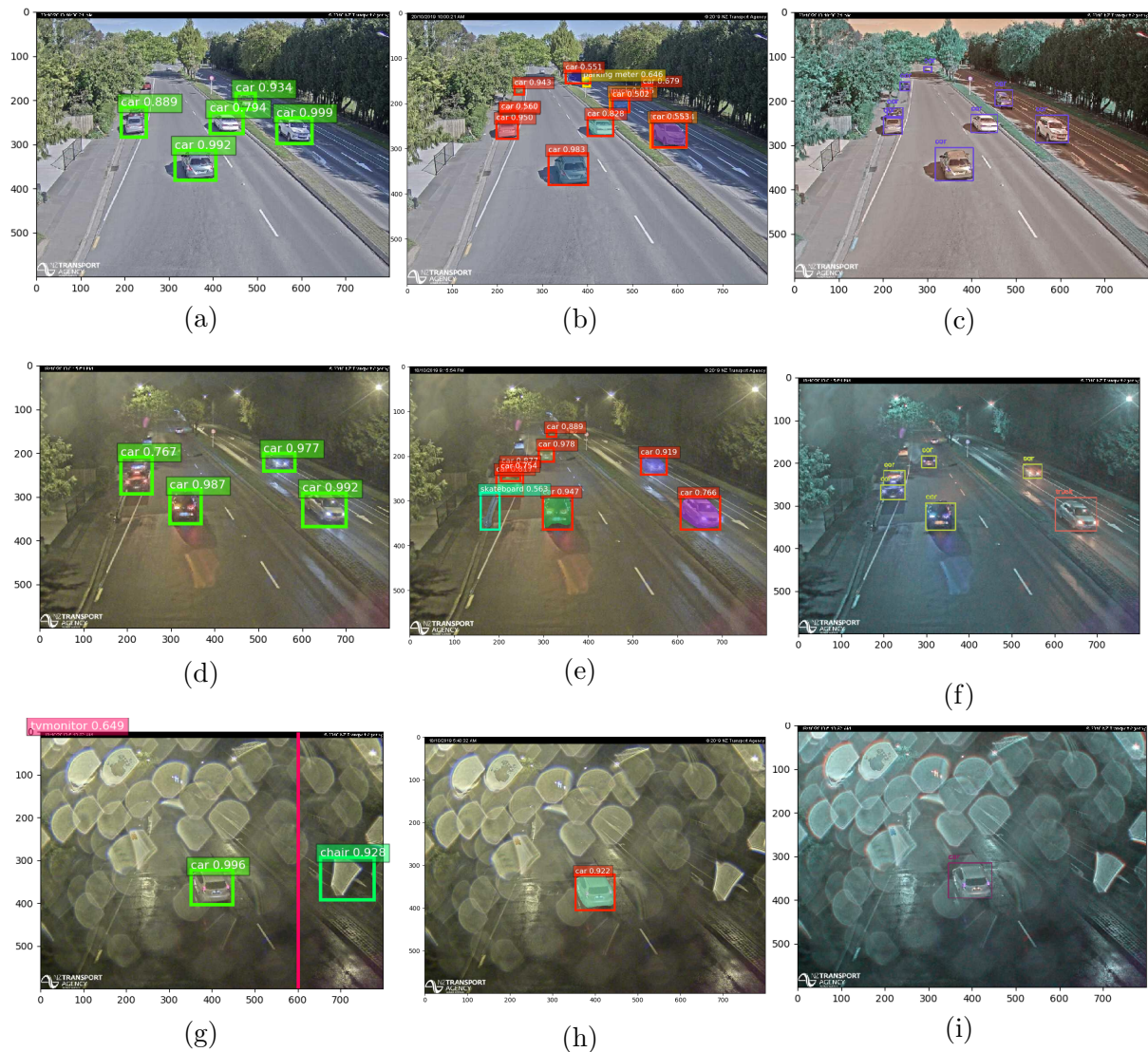
```

---

We used the YOLOv3 implementation in Python cvlib to obtain the traffic flow count in every single minute.

## 5.4 Results and Discussion

According to Table 5.3, YOLOv3 has high recall and precision values for the selected data set. High precision relates to a low false-positive rate, and high recall relates to a low false-negative rate. High scores for both show that the model is returning accurate results. Comparatively, faster R-CNN has high precision but low recall, which means that very few results are returned, but most of its identified objects are correct. furthermore, mask R-CNN returns a low recall and precision



**Figure 5.7** Vehicle Detection (a) faster R-CNN-Day (b) mask R-CNN-Day (c) YOLOv3 R-CNN-Day (d) faster R-CNN-Night (e) mask R-CNN-Night (f) YOLOv3-Night (g) faster R-CNN-Blur (h) mask R-CNN-Blur (i) YOLOv3-Blur

values compared to YOLOv3. Therefore, based on both performance and accuracy values, we have chosen YOLOv3 as the most appropriate algorithm for this project to estimate traffic flow. All experiments were carried on Ubuntu 18.04.3 with Nvidia Geforce graphics, 8 CPU cores (Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz) and 8 GB RAM. Figure 5.10 and 5.11 show a sample of the vehicle counts generated by YOLOv3 for the data set. However, these results are not validated. Therefore, we will use a CCTV video recording at the same location to get the traffic flow by manually counting as the ground truth. Then, the generated traffic flow will be evaluated against the ground true flow.

The contribution of the paper can be summarized as follows:

- We have constructed a new, challenging data set by collecting CCTV images at each minute through the NZTA traffic cameras API, which includes a total of 24,085 images for the experiments discussed in this paper. To the best of our knowledge, this is the first time that such a large CCTV data set has been used to formulate traffic flow using Deep Learning.



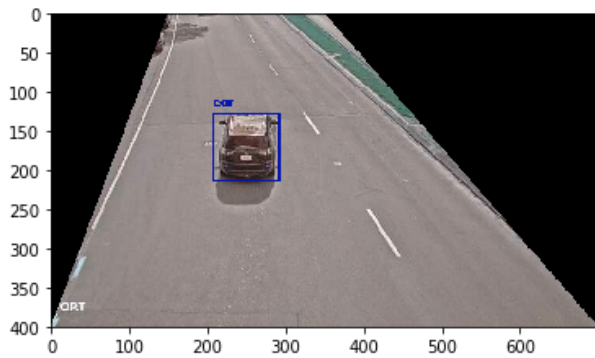


Figure 5.8 Left Lane.

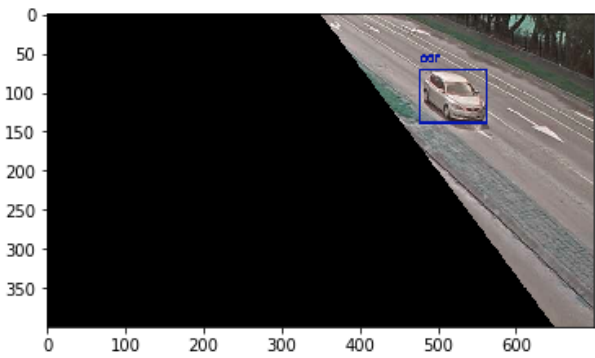


Figure 5.9 Right Lane.

Date	Time	LeftLane	RightLane
2019-10-15	14-44-00	1	0
2019-10-15	14-45-00	1	0
2019-10-15	14-46-00	0	4
2019-10-15	14-47-00	0	4
2019-10-15	14-48-00	0	4
2019-10-15	14-49-00	0	0
2019-10-15	14-50-00	0	0
2019-10-15	14-51-00	1	10
2019-10-15	14-52-00	1	10
2019-10-15	14-53-00	1	5

Figure 5.10 A sample of the obtained vehicle flow counts

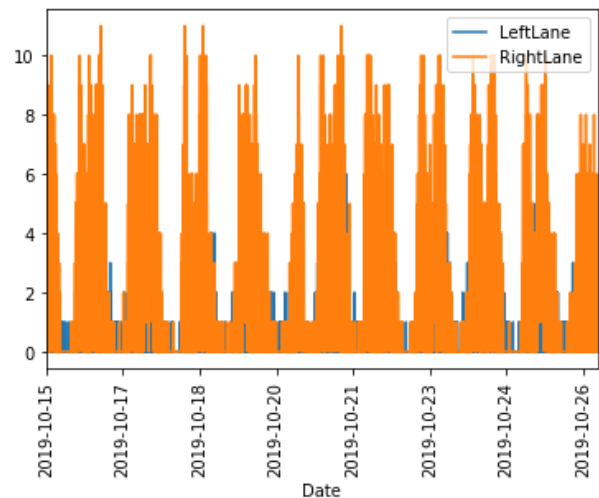


Figure 5.11 A plot showing a sample of traffic flow

- We have evaluated the performance and accuracy of YOLOv3, faster R-CNN and mask R-CNN in counting vehicles for the CCTV images. Then, we obtained the traffic flow counts for the selected road at Christchurch CBD.
- We have developed a simple ROI algorithm to identify “left-lane” and “right-lane” in the CCTV images to identify the direction of vehicle movement.

## 5.5 Conclusion

In this paper, we discussed the problem of traffic flow estimation as the first step of a broader project for emergency traffic management. As a case study, we considered one of the busiest roads in Christchurch, New Zealand. This subject area is still in the early stages of development, and there are only a few works that aim to process the CCTV image series automatically for traffic analysis. To address this issue, we started by evaluating the performance accuracy of three popular object detection models, namely faster R-CNN, mask R-CNN and YOLOv3. Our experiment 1 results showed that YOLOv3 was very fast to detect objects compared to the other two models. Also, in the same experiment, we demonstrated that YOLOv3 had the highest accuracy. During the experiment 2, we introduced a simple ROI selection heuristic algorithm to select “left-lane”

and “right-lane” of each image. We applied the YOLOv3 model to count the number of vehicles in such as car, bus, truck and motorcycle in each minute to formulate the traffic flow.

Future work for this project beyond this paper will be extended to the identification of traffic flow to the entire city of Christchurch using multi-cameras. Then, the short-term traffic flow will be predicted at any location in the city. During the final step of the project, the prediction of traffic flow will be compared with the real-time traffic flow to identify traffic anomalies. This will allow emergency management personnel to decide whether to re-route, change traffic signals or make any other decisions within a few seconds of an incident. Also, during an incident, emergency managers can use the prediction to estimate the traffic flow attempting to evacuate from different routes in the city.

### **Manuscript 3: Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage**

The following article is published as: Nilani Algiriyage, Raj Prasanna, Kristin Stock, Emma Hudson-Doyle, David Johnston, Minura Punchihewa, et al. (2021). Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage. In Anouck Adrot, Rob Grace, Kathleen Moore, & Christopher W. Zobel (Eds.), ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management (pp. 40–48). Blacksburg, VA (USA): Virginia Tech.

#### **Abstract**

Traffic emergencies and resulting delays cause a significant impact on the economy and society. Traffic flow estimation is one of the early steps in urban planning and managing traffic infrastructure. Traditionally, traffic flow rates were commonly measured using underground inductive loops, pneumatic road tubes, and temporary manual counts. However, these approaches can not be used in large areas due to high costs, road surface degradation and implementation difficulties. Recent advancement of computer vision techniques in combination with freely available closed-circuit television (CCTV) datasets has provided opportunities for vehicle detection and classification. This study addresses the problem of estimating traffic flow using low-quality video data from a surveillance camera. Therefore, we have trained the novel YOLOv4 algorithm for five object classes (car, truck, van, bike, and bus). Also, we introduce an algorithm to count the vehicles using the SORT tracker based on movement direction such as “northbound” and “southbound” to obtain the traffic flow rates. The experimental results, for a CCTV footage in Christchurch, New Zealand shows the effectiveness of the proposed approach. In future research, we expect to train on large and more diverse datasets that cover various weather and lighting conditions.

## **5.6 Introduction**

Today, with the high rate of urbanization, the number of vehicles in an urban road network has increased significantly. According to statistics released by the Ministry of Transport, there were 11,449 accidents in New Zealand, including 2,449 that caused serious and fatal injuries in 2019<sup>4</sup>.

---

<sup>4</sup><https://www.transport.govt.nz/statistics-and-insights/safety-annual-statistics/>

The resulting congestion and related issues after crash incidents cause substantial economic loss and disrupt the community’s everyday life. Furthermore, other natural and man-made disasters such as flooding, landslides and terrorist attacks causing traffic emergencies are inevitable. During such emergencies, the road network becomes congested, making evacuation impossible and rescue personnel and supplies unable to be transported [21, 260]. Therefore, traffic emergencies must be addressed in an intelligent transport system to ensure secure, responsive and efficient transportation for everyone [296, 104].

Understanding road traffic behaviour is a key component of an emergency traffic response plan. Traffic flow estimation is the first step for identifying the road traffic patterns, contributing to traffic modelling, urban planning and design processes for all aspects of a road network [104]. Traffic data acquisition is typically performed using underground inductive-loops, pneumatic road tubes, and manual counts. However, these methods are labour intensive, expensive, difficult to install and can be inaccurate. Also, they could damage the road surface and reduce the quality and life of the road and thus can not be used in large areas [28].

Closed-circuit television (CCTV) systems are now increasingly popular and are installed in many public places to enable real-time surveillance. As these systems are continuously operated, they generate a vast amount of data that contribute to big data. Recent developments in computer vision research have heightened the need for using CCTV images to tackle practical problems such as traffic congestion detection [207], automatic licence plate recognition [167, 209], emergency vehicle detection [332] and accident detection [154, 377]. However, traffic flow estimation using computer vision algorithms for surveillance camera datasets is still in very early development. Difficulties in moving, storing, processing and developing efficient algorithms to analyse CCTV data have been identified as significant challenges [104].

This study aims to answer the research question: 1) Can traffic flow be estimated from low-quality CCTV video footage in real-time?. As a case study, we focus on a multi-lane road in Christchurch Central Business District (CBD). We obtain the traffic flow based on vehicle movement direction such as “northbound” and “southbound”. Furthermore, vehicle counts are obtained for five vehicle classes, such as *car*, *bus*, *van* and *truck* and *bike*. We train You Only Look Once (YOLOv4) algorithm [48] for vehicle detection and classification and Simple Online and Real-time Tracking (SORT) [47] algorithm for vehicle tracking. Last year, we introduced our algorithm as a conference poster<sup>5</sup>. However, it was an early in-progress work that we used YOLOv3, trained on Common Objects in Context (COCO) dataset for four vehicle classes. The algorithm discussed in this paper is improved by custom training YOLOv4. Authorities can use our algorithm for traffic flow monitoring, traffic anomaly identification, and the development of emergency rescue plans. Also, responders are able to make management decisions such as detour allocation and changing traffic light timing length during emergencies by using real-time traffic flow.

The contributions of the paper are as follows:

- We have trained YOLOv4 with our own vehicle image dataset and publicly made available the dataset for future researchers <sup>6</sup>.
- We have introduced an algorithm to count directional traffic flow using YOLOv4 and SORT tracker.

---

<sup>5</sup><https://conference.eresearch.edu.au/2020/09/real-time-traffic-flow-estimation-based-on-deep-learning-using->

<sup>6</sup>Annotated vehicle dataset, Traffic\_Flow\_Estimation

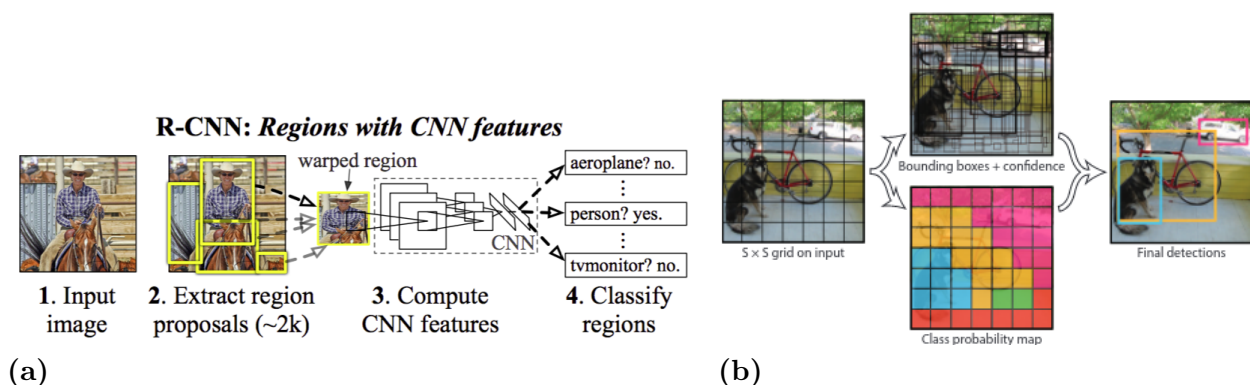


- We show that the custom trained YOLOv4 performs well having a F1-score of more than 0.95 for car class during day, evening and night times using a low-frame-rate footage.

The rest of our paper is outlined as follows. The section Related Work reviews the existing work. Then in the Methodology section, we illustrate the architecture and algorithms implemented. The Results section describes our research findings. Finally, we present concluding remarks and future research steps in section the Discussion.

## 5.7 Related work

Recently, visual datasets obtained from surveillance cameras and aerial vehicles have been explored for many traffic monitoring applications [155, 417, 187, 6]. Convolutional Neural Networks (CNN) based object detectors have been widely adopted for such visual datasets in computer vision research. These algorithms can generally be divided into two major groups, namely, single-stage detectors and two-stage detectors. Single-stage detectors such as Single Shot Detector (SSD) and YOLO are generally fast and predict object bounding boxes together with classes within a single network pass [321, 227]. In contrast, two-staged detection happens in two stages. First, the model proposes a set of regions of interests by selective search or using Regional Proposal Network (RPN). Then a classifier only processes the region candidates to identify the objects [374, 116, 115, 322, 138] (see Fig. 5.12). As a result, two-stage detection tends to be slow (e.g., R-CNN family networks including the original R-CNN, Fast R-CNN, Faster R-CNN and Mask R-CNN).



**Figure 5.12** Two-stage detection vs single-stage detection (a) R-CNN architecture [116] (b) YOLO object detection [321].

Vehicle object detection, classification and tracking are the three main tasks involved while processing video datasets for traffic flow estimation [104, 282]. Object detection deals with drawing bounding boxes around the objects of interest to locate it within the image. Classification helps to categorise objects into different classes such as “car, bus, truck”. In 2015, Redmon et al. introduced You Only Look Once (YOLO) as a fast, accurate and real-time object detection system. It went through several modifications of the architecture until it produced YOLOv3 in 2018 [319, 320]. Chakraborty et al. (2018) evaluated the performance of deep convolution neural network (DCNN), support vector machine (SVM) and basic YOLO algorithm for classifying traffic congestion from CCTV images. They show that YOLO algorithm obtaining the highest accuracy of 91.4 for the task. In a similar study by Algiriyage et al. (2020) uses a CCTV image dataset to obtain traffic flow and compare the performance of YOLOv3, faster R-CNN and mask-RCNN for object detection. They

show that among them, YOLOv3 showed the best performance in terms of speed and accuracy for their image dataset having a precision value of 0.96. Corovic et al. [73] train YOLOv3 to detect five classes of objects, namely, cars, trucks, pedestrians, traffic signs and traffic lights under different lighting conditions. Though they used a small dataset having 300 images, they could get an F1-score of 0.59. YOLOv3-tiny version pre-trained on COCO dataset was used by Oltean et al. (2019) for real-time traffic counting. They show that for the few frames they considered for the experiment out of the total 27, 26 vehicles were correctly detected. In 2020, YOLOv4 was introduced as a faster and more accurate detector than the all available CNN based detectors [386]. However, far too little attention has been paid to research on using YOLOv4 for vehicle object detection.

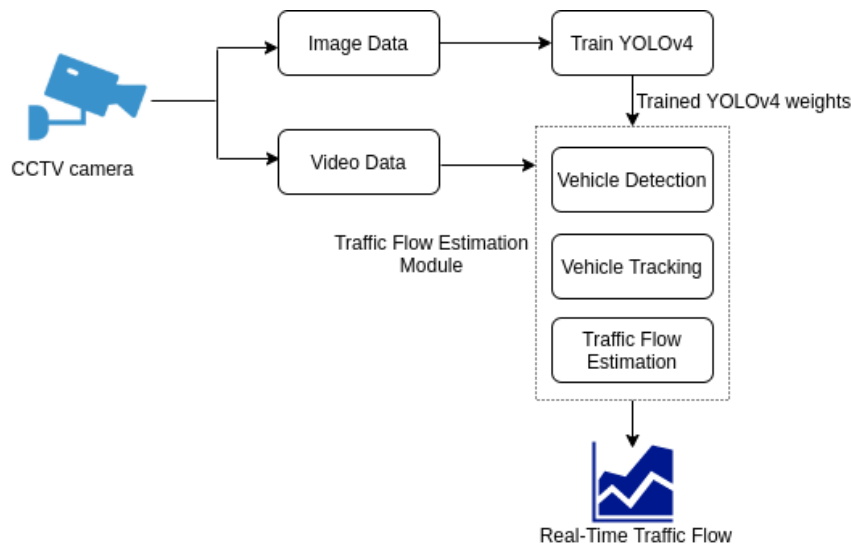
Object tracking analyses the movement path of an object across different frames. Depending on the tracking target, there are two categories of tracking algorithms such as Single object tracking (SOT) and Multiple object tracking (MOT). In SOT, a single object is tracked from the beginning, while in MOT, several objects are detected and tracked from one frame to the other [266]. Two well-known examples of SOT algorithms include Kalman Filtering and Particle Filtering, whereas SORT and DeepSORT are two state-of-the-art MOT algorithms [53]. Several studies have investigated vehicle tracking and counting from CCTV videos [69, 234, 64, 343]. For example studies by, Bui et al. (2020) and Nam Bui et al. (2020) use DeepSORT for vehicle tracking and virtual lines for traffic counting. However, the direction of vehicle movement is not considered while obtaining the traffic flow. Closer to our objective is the traffic counting system introduced by Fedorov et al. [104]. They use Faster-RCNN object detector and SORT tracker. However, they have carried out experiments for 982 video frames and do not obtain the traffic flow by vehicle class. Apart from Fedorov (2019), there is a general lack of research in investigating the real-time traffic flow estimation from surveillance video, while also considering movement direction and vehicle class. Thus, this study was set out to explore traffic flow estimation in real-time from CCTV video considering these gaps.

## 5.8 Methodology

This study investigates real-time traffic flow estimation from low-quality surveillance video data. Also, we classify vehicles and obtain the flow rate based on their movement directions. In order to achieve this objective, we train the novel YOLOv4 algorithm with a custom image dataset collected from the same camera to detect five object classes namely, car, bus, truck, van and bike. Then the trained weights are used for the traffic flow estimation module (see Figure 5.13). The traffic flow estimation module counts vehicles based on the direction of movement and the class of the vehicle from CCTV video data. Therefore, this module is divided into three sub-tasks: vehicle detection, vehicle tracking, and traffic flow estimation. The vehicle detection module draws a bounding box around vehicle objects in order to locate it within a frame, while the vehicle tracking module tracks the movement of a vehicle object between different frames. Our algorithm can be easily applied to any similar location with very few modifications and extended to complex locations with changes based on the degree of complexity.

### 5.8.1 Dataset

We obtained CCTV image and video datasets from the New Zealand Transport Agency (NZTA), Christchurch, New Zealand. As a case study, we selected a busy road namely “West along Yaldhurst



**Figure 5.13** Methodology for Real-time traffic flow estimation.

Rd from Curletts Rd” in Christchurch CBD. The image datasets were used to train YOLOv4 while the footage datasets were used to validate the real-time traffic flow counting algorithm. The camera at the selected location generates video with a frequency of  $\approx 10$  frames per second (fps) and resolution of  $1280 * 720$  (*width \* height*). The three video footage that we analysed for this research was recorded during the day, evening and night times, in February 2020. Table 5.4 and Table 5.5 summarise the details of the image and video dataset respectively.

**Table 5.4** Details of the image dataset used to train YOLOv4

Vehicle Class	Total Instances
Car	13,627
Bus	141
Van	779
Truck	1,273
Bike	280

### 5.8.2 Vehicle detection

The foundation of our detection module is the novel single-stage YOLOv4 detector [48]. This model was trained on the image dataset obtained from the NZTA as described in Table 5.4, using the Darknet 1<sup>7</sup> implementation of the YOLOv4 algorithm. The images were annotated using LabelImg tool<sup>8</sup> prior to carrying out the training process. The training was carried out using the Mahuika High-Performance Computing (HPC) cluster of the New Zealand eScience Infrastructure (NeSI) for a total of 10,000 epochs. The total amount of time taken for the training was around 15 hours on 2 GPU cores.

<sup>7</sup>Darknet, open source neural network framework, <https://github.com/pjreddie/darknet>

<sup>8</sup>LabelImg, graphical image annotation tool, <https://github.com/tzutalin/labelImg>

**Table 5.5** Details of the analysed CCTV videos (hr: hours, mins: minutes and secs: seconds)

Description	Start Time	Finish Time	Duration	No of frames
Video 01	10:00:00 (UTC + 12:00)	11:42:33 (UTC + 12:00)	1 hr, 42 mins & 33 secs	69, 676
Video 02	18:06:17 (UTC + 12:00)	19:06:56 (UTC + 12:00)	1 hr, 0 mins & 39 secs	86, 987
Video 03	20:26:12 (UTC + 12:00)	21:56:30 (UTC + 12:00)	1 hr, 30 mins & 18 secs	55, 794

### 5.8.3 Vehicle tracking

Vehicle tracking deals with identifying the vehicle movement from one frame to the other. To handle this, we adopted the SORT tracker [47] as it is both powerful and fast [104].

### 5.8.4 Vehicle movement direction estimation and traffic flow counting

Figure. 5.14 shows a drawing of the location we analysed which is a “multi-lane” road where there are two lanes for each direction. The *traffic flow rate* can be defined as the number of vehicles during the  $t^{th}$  time interval at the  $i^{th}$  observation location in a transportation network which is given by Eq. 1 [204].

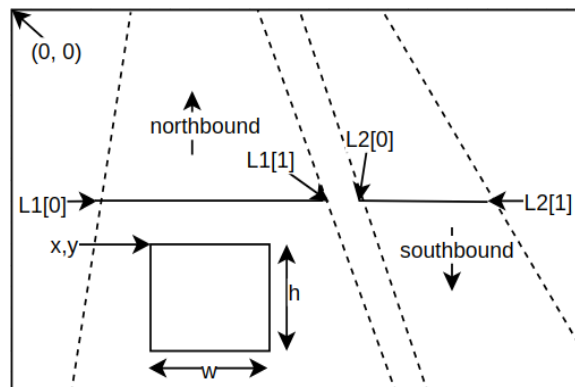
$$X_i^t = n/t \quad (5.1)$$

where:

$X_i^t$  = traffic flow rate

$n$  = number of vehicles

$t$  = time duration



**Figure 5.14** Drawing of the location analysed - Line coordinates (L1[0], L1[1], L2[0], L2[1]) and bounding box properties of a vehicle object (x, y, width (w), height (h)).

The width and height of a single frame in the analysed video is 1280 \* 780. We define two lines with coordinates (400, 300) – L1[0], (750, 300) – L1[1], (820, 300) – L2[0] and (1160, 300) – L2[1] to identify the movement direction of a vehicle such as “northbound” and “southbound”. We analyse each frame ( $i$ ) in the set of frames ( $I$ ) of the footage. If a vehicle enters a particular line, it is detected, classified and tracked over different frames. A simple mathematical calculation is applied

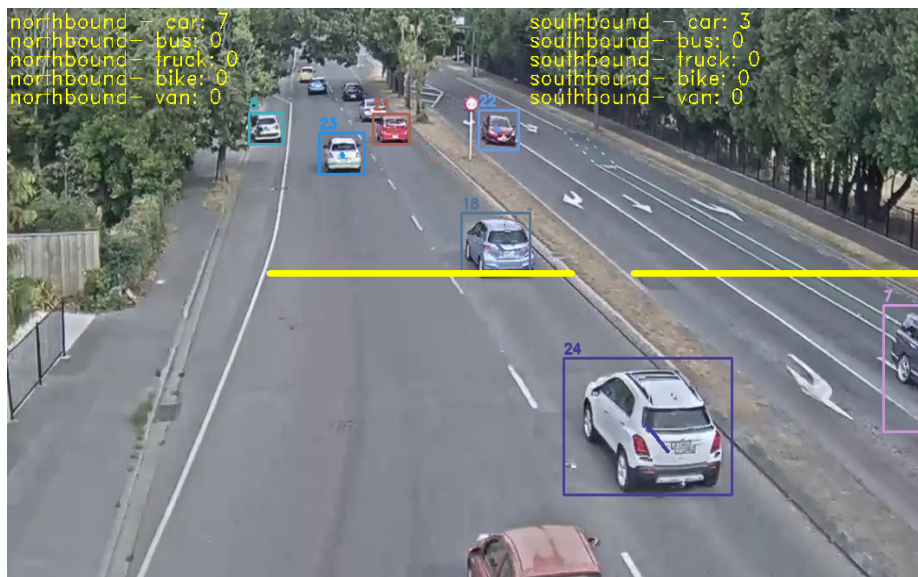
to count the intersections between the vehicles' previous and current frame positions using the defined lines. This is performed using the center of the bounding boxes ( $cnt0, cnt1$ ) in the current and previous positions and also using the line coordinates. Then, when an intersection is found, our algorithm checks the YOLOv4 class label to increase the car ( $n\_car\_count$ ), bus ( $n\_bus\_count$ ), van ( $n\_van\_count$ ), truck ( $n\_truck\_count$ ) and bike ( $n\_bike\_count$ ) count in each movement direction. Our algorithm writes the real-time traffic counts into a text file. We use python pandas library<sup>9</sup> and matplotlib FuncAnimation<sup>10</sup> to live plot the traffic flow. The pseudo-code for the traffic estimation is proposed in Algorithms<sup>11</sup> and Algorithm<sup>12</sup>.

## 5.9 Results

The overall Mean Average Precision (mAP) of the YOLOv4 model was 92.35%, and the performance on each class as per the validation dataset is summarized in Table 5.6.

**Table 5.6** Mean Average Precision (mAP) of vehicle detector classes

Class	Average Precision (AP)
Car	96.94%
Bus	93.64%
Van	90.22%
Truck	90.24%
Bike	90.72%



**Figure 5.15** Traffic flow estimation from video footage.

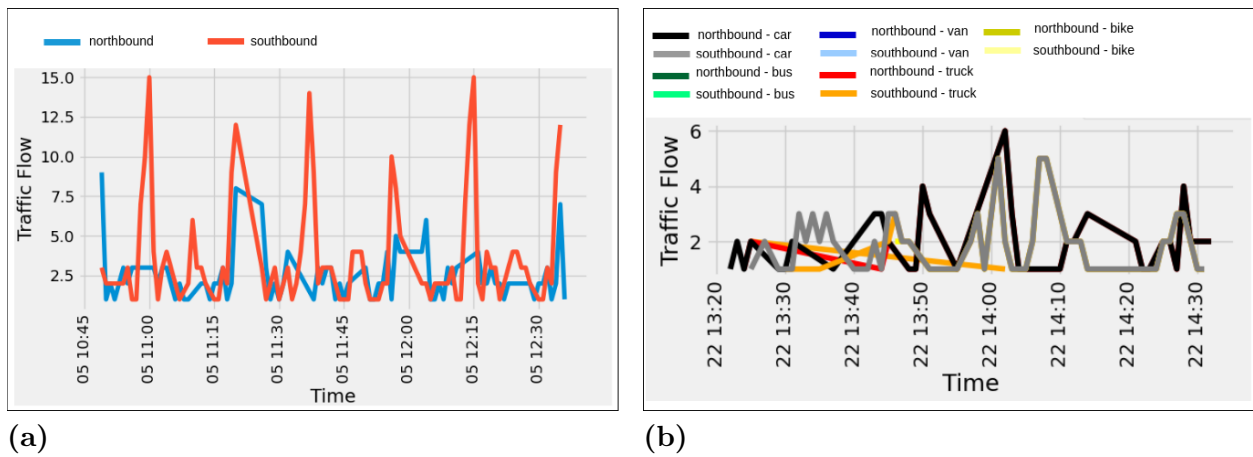
The AP value has dropped proportionately to the number of instances we used in the training dataset. For example, we used a higher number of images for the car class and hence it has got an

<sup>9</sup>pandas software library, <https://pandas.pydata.org/>

<sup>10</sup>Real-time plotting library, [https://matplotlib.org/stable/api/\\_as\\_gen/matplotlib.animation.FuncAnimation.html](https://matplotlib.org/stable/api/_as_gen/matplotlib.animation.FuncAnimation.html)

<sup>11</sup>Traffic flow estimation algorithm, Traffic-Flow-from-Footage/blob/master/Traffic\_Flow\_Estimation.png

<sup>12</sup>Intersection detection algorithm, Traffic-Flow-from-Footage/blob/master/Intersect.png



**Figure 5.16** Live plots of traffic flow (a) Directional traffic flow (b) Traffic flow by vehicle class.

AP value of 96.94%. Figure 5.15 illustrates the real-time traffic flow estimation system based on custom trained YOLOv4. Two live plots are generated to show the directional traffic flow, and the flow counts by vehicle class, as presented in Figure 5.16.

To measure the detection accuracy, we manually counted the number of vehicles in each class for the video footage analysed and used as ground-truth values. Then the accuracy value is measured using the equation 5.2.

$$\text{Accuracy} = \frac{\text{No of correct detections}}{\text{No of ground-truth detections}} \quad (5.2)$$

**Table 5.7** Number of vehicles counted by humans (ground-truth), automatically by our algorithm and the accuracy for video 01 (day), video 02 (evening) and video 03 (night).

Video Dataset	Vehicle Class	Ground-truth		Number Detections		Accuracy	
		northbound	southbound	northbound	southbound	northbound	southbound
Video 01	car	1832	1308	1770	1189	0.9662	0.9090
	bus	8	4	5	2	0.6250	0.5000
	van	40	58	28	34	0.7000	0.5862
	truck	343	389	220	248	0.6414	0.6375
	bike	2	2	1	0	0.5000	0.0000
Video 02	car	1396	1057	1368	1022	0.9799	0.9669
	bus	3	2	2	1	0.6667	0.5000
	van	20	18	14	12	0.7000	0.6667
	truck	34	26	24	18	0.7059	0.6923
	bike	2	0	1	0	0.5000	1.0000
Video 03	car	798	802	774	774	0.9699	0.9651
	bus	3	2	1	0	0.3333	0.0000
	van	22	17	18	12	0.8182	0.7059
	truck	68	46	58	38	0.8529	0.8261
	bike	2	1	1	0	0.5000	0.0000

Table 5.7 illustrates the accuracy scores for all three videos considered for our experiments. The mean accuracy for obtaining “northbound” traffic flow is 0.7114 while the “southbound” is 0.6397 for all vehicle classes. A possible explanation for this might be that the camera was located

close to the “northbound” lane. Therefore, our detection module could identify vehicles in the “northbound” lane more accurately. Furthermore, the flow count for the car object class is more accurate; obtaining a mean accuracy score of 0.9595. For the location we considered, the vast majority of the vehicles consisted of car class. The higher accuracy for car class indicates that our system performs well finding the traffic flow from video footage. However, the detection accuracy for bus class is lower (mean accuracy score : 0.1667). An implication of this is the possibility that the lower number of bus image instances in the training dataset. Furthermore, it is interesting to note that there is no significant difference between the accuracy scores of the three footage considered during different times of the day. Finally, several limitations need to be considered. First, we didn’t incorporate the vehicle re-identification problem. For instance, the same vehicle can be counted many times with the current approach. This can affect the traffic flow count as duplicated entries. Second, our CCTV footage was captured during the summertime in New Zealand. The lighting conditions might vary during other times of the year. Future work needs to explore the detection accuracy during night times. Third, we lose tracking vehicle objects in some frames due to occasional poor quality visuals generated from the cameras. As a result, some vehicle objects are missed by the flow counting algorithm.

## 5.10 Conclusion

In this study, we focused on obtaining real-time traffic flow using low-quality CCTV footage. As a case study, we selected one of the busiest multi-lane roads in Christchurch CBD, New Zealand. We trained the YOLOv4 model to detect five vehicle object classes: car, bus, van, truck and bike. The test results of this study show that we could obtain a high accuracy for the car class (mean accuracy score : 0.9595) while obtaining the traffic flow based on the movement direction. However, as our training image dataset was unbalanced, we obtained a lower accuracy score for the bus class. Therefore, further work needs to train YOLOv4 with a higher number of vehicles for low accurate classes. Therefore, in future work, we hope to train YOLOv4 algorithm with a large and diverse dataset. In addition, we hope to apply this work to more complex crossroads and consider the traffic count per lane.

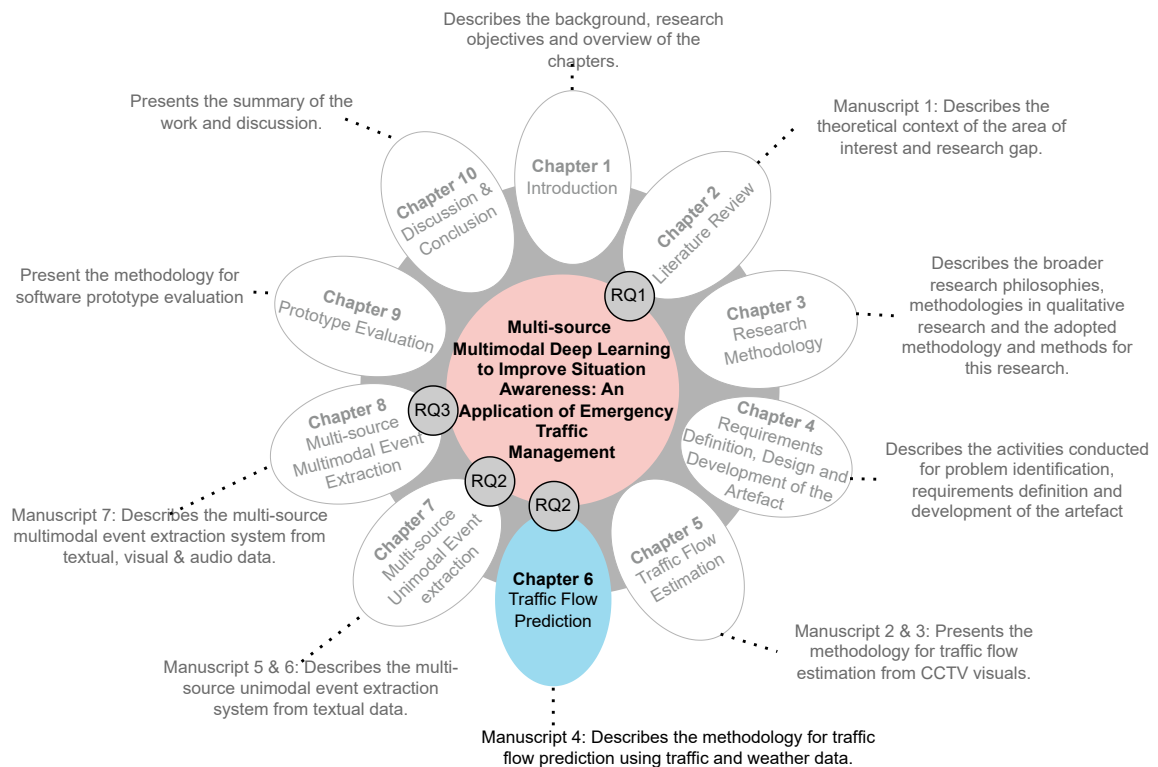
## 5.11 Summary

This chapter presented two conference papers describing the software artefact’s first component. First, section 5 described the traffic flow estimation from CCTV images. The performance and accuracy of different DL algorithms for this purpose were evaluated. YOLOv3 model was selected and used to count the number of vehicles by class, such as car, bus, truck and motorcycle, in each minute to formulate the traffic flow. Second, in section 5.5, the methodology for traffic flow estimation from CCTV footage was presented. During this project, YOLOv4 model was trained for a New Zealand-based vehicle dataset for traffic flow, counting for five vehicle object classes: car, bus, van, truck and bike.



## Chapter 6

# Deep Learning-based Short Term Traffic Flow Prediction with Weather Data



This chapter presents the fourth manuscript that describes the second component of the software artefact. This chapter intends to answer the second research question, “How can data from multiple sources be fused to support disaster response?”. The chapter describes a novel DL architecture that predicts short-term traffic flow using both traffic flow and weather data.



## Abstract

Short-term traffic flow prediction is essential in intelligent transport management and control. This problem has been extensively explored in research using statistical methods. However, traffic flow prediction has recently become too complex for these statistical procedures, due to the increased volume of data that has become available from various sources such as inductive loops, Bluetooth sensors, and traffic cameras. Statistical parametric methods underperform due to the highly non-linear and stochastic characteristics of transportation systems. Recently, multiple Deep Learning (DL) methods have been explored for traffic flow prediction. However, existing models for traffic flow prediction give little consideration to the influence of weather parameters. This paper proposes a new DL architecture for traffic flow prediction and comparatively evaluates the performance of the model considering traffic data as a single source and a fusion model using both traffic and weather data. We propose a univariate time series model based on stacked Bidirectional LSTM networks (Bi-LSTM) for traffic-only data, and a fused model that combines stacked Bi-LSTMs and Dense Networks for traffic and weather data. The effectiveness of the proposed model is verified using multiple experiments.

## 6.1 Introduction

With increased urbanization, traffic flow plays an integral part in everyday life. A proper understanding of traffic flow patterns provides multiple ways to relieve heavy traffic and is essentially helpful for both individual travelers and government agencies. Short-term traffic flow prediction has recently received considerable attention due to the emergence of Intelligent Transportation Systems (ITS) [415, 183]. For example, traffic flow prediction systems can foresee potential congestion and inform people of detour routes, enabling redistribution of traffic to avoid congestion in city environments. As a result, travel time is reduced, and carbon emissions are minimized. Additionally, traffic flow prediction systems can help authorities with their daily decision-making tasks (e.g., traffic signal optimization) in order to improve traffic operation efficiency [384].

Vlahogianni et al. [383] define short-term traffic forecasting as the process of estimating traffic conditions directly at a future time, based on continuous short-term feedback of traffic information. Moreover, according to Lee et al. [212] the predictions can vary from “few seconds to possibly a few hours based on current and past traffic information”. The short-term traffic flow prediction problem has been widely explored with many parametric and non-parametric methods since the 1980s. However, recently traffic data is being captured from various sensors and devices, such as Bluetooth sensors, Global Positioning System (GPS) sensors embedded in mobile devices, social media, and CCTV cameras in addition to traditional methods such as inductive loops and radars [236, 212]. As more traffic data of different kinds become available, Machine Learning (ML)/Deep Learning (DL)-based traffic flow prediction has attracted a lot of academic and industrial interest. These different approaches for traffic prediction underlining the complexities of design and development have been reviewed by multiple authors [383, 384, 212].

In reality, traffic flow is affected by many external factors such as weather, school holidays, special events (e.g., protests), planned/unplanned maintenance, vehicle crashes, and other disasters (e.g., landslides and flooding). Among them, weather factors affect the traffic flow throughout the year. The existing research has mainly focused on traffic flow data, with only a few studies

incorporating weather data for the short-term traffic flow prediction problem [425, 147]. However, with the existing very limited research, it is difficult to say that one method is clearly superior to other methods in any situation. One reason for this is that the proposed models are developed with a small amount of separate specific traffic data, and the accuracy of the traffic flow prediction model is dependent on the traffic flow features embedded in the collected spatiotemporal traffic data. Moreover, the correlation between different weather parameters and traffic data depends heavily on the location. Thus there is much potential for further exploration of deep learning-based traffic flow prediction using weather data.

Siami et al. [347] show that Bi-directional Long Short Term Memory Networks (Bi-LSTMs) perform better than regular Long Short Term Memory Networks (LSTMs) for time series prediction tasks. Bi-LSTMs accomplish this performance improvement by processing the input sequence twice, resulting in improved learning of long-term relationships and, as a result, improved model accuracy. According to the recent survey done by Lee et al. [212], very little research has explored the performance of Bi-LSTM for the short-term traffic flow prediction problem. Based on the deficiencies of existing methods, we propose a novel DL architecture using stacked Bi-LSTM networks for short-term traffic flow prediction using traffic-only data and fusing both traffic and weather data to explore the problem further. We compare our results with the proposed approaches by Hou et al. [147] and Zheng et al. [425].

The contributions of this paper are as follows:

- We propose a novel DL model to predict traffic flow given weather parameters based on stacked Bi-LSTM networks.
- We evaluate the performance of traffic flow estimation when using traffic-only data and when utilizing both traffic data and weather data:

Two experiments are conducted to evaluate the performance of the proposed DL architecture; a univariate time series model (traffic-only data) and a decision-level fused model (fusing both traffic and weather data). Furthermore, we conduct a correlation analysis using Pearson coefficient value and expert meteorologists and transport experts' feedback while choosing weather parameters for the fusion model.

- We compare the performance of the proposed Bi-LSTM model against LSTM, GRU, and AE models and the recently proposed architecture by Hou et al. [147].

The remainder of this paper is organized as follows. Section 6.2 presents relevant related work. Section 6.3 introduces the methodology followed by the proposed DL architectures. In Section 6.4 experimental results are discussed. Finally, Section 6.5 provides concluding remarks.

## 6.2 Related Work

Traditionally, traffic flow has been predicted using parametric approaches such as autoregressive models and the Kalman filter. For example, Auto-Regressive Integrated Moving Average (ARIMA) [214] is one of the most frequently used parametric regression models. It assumes that traffic condition is a stationary process, where the mean-variance and auto-correlation are unchanged. Later, multiple variants to the ARIMA model were proposed such as Kohonen ARIMA (KARIMA) [375], subset ARIMA [213], ARMA [182] and seasonal ARIMA (SARIMA) [402]. Other studies,

for example, [96, 256, 92], use Kalman filters to predict traffic flow since they are easy to adjust, fast to compute and robust against external disturbances and errors of the model. However, one of the main disadvantages of parametric models is that they need to be readjusted whenever the parameters that define the model need to be changed [250].

In comparison to parametric approaches, non-parametric approaches such as k-nearest neighbors, Bayesian Networks or Support Vector Regression models (SVR) show stronger function fitting ability in complex and nonlinear traffic flow prediction problems [230, 146]. The essential idea of this kind of method is to transform low-dimensional and linearly inseparable traffic data into high-dimensional and linearly separable expressions through the application of a kernel function. Among the studies of these methods, Hong et al. [146] proposed an SVR traffic flow prediction model employing the hybrid genetic algorithm to determine the suitable combination of parameters. In another study, Hu et al. [149] used particle swarm optimization (PSO) to determine optimal parameters for SVR for higher precision in short-term traffic flow forecasting problems. Ling et al. [226] proposed an addition to the work by Hu et al. [149] by introducing a multi-kernel SVM and using adaptive particle swarm optimization (APSO). A novel SVM with an adaptive multi-kernel (AMSVM) was introduced by Feng et al. [105] for the traffic flow prediction problem. However, all of these non-parametric models are unable to incorporate the complexities of large-scale traffic data, limiting the models' performance [250, 230, 422].

Recently, with more traffic data becoming available with ITS/BigData, the short-term traffic flow prediction problem has become more complex, requiring models with greater data modeling capabilities. For example, GPS tracking devices embedded in mobile devices, Bluetooth sensors, CCTV monitoring, and social media generate a massive amount of traffic data in addition to the traditional methods such as inductive loops, pneumatic tubes, piezoelectric cables, bending plate detectors, and radar technology [372]. DL algorithms such as Convolutional Neural Networks (CNNs), LSTMs, and recurrent neural networks (RNNs) have been used for time series prediction in ITS [391, 422, 405, 407]. The ability of Deep Neural networks to learn features automatically and handle data over-fitting problems makes them suitable for highly complex non-linear functions in traffic prediction problems. Among them, LSTM networks have shown improved performance for traffic flow prediction problems. This improved performance is due to the ability of these models to learn information with a long time span [425, 422]. Yi et al. [407] applied an LSTM network for short-term traffic prediction at an expressway in Korea. In a similar study, Zaho et al. [422] used LSTM for traffic flow prediction and showed improved results over multiple parametric and non-parametric approaches. Some research has proposed additions to the general LSTM network for improved prediction results. For example, a study by Liu et al. [230] proposed Conv-LSTM to extract spatial-temporal traffic flow information of traffic flow information.

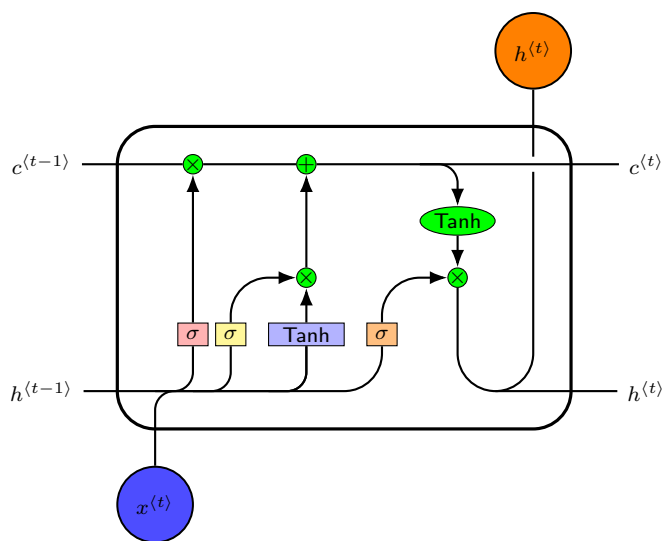
Traffic systems are dynamic and susceptible to a variety of external factors, including weather conditions, road maintenance activities, vehicle crash incidents, and other disaster events. As a result, traffic flow prediction models also have to consider such parameters to make a more accurate forecast. However, in the literature, very few studies have explored weather conditions for the traffic flow prediction problem with DL models. For example, a study by Koesdwiady et al. [194] used weather parameters such as temperature, humidity, visibility, wind speed, wind gust, dew point, and cloud layer height for a Deep Belief Network (DBN)-based traffic prediction model. They used decision-level fusion to combine traffic and weather data. However, they considered a very small dataset, having only three months of data. Zhang et al. [415] proposed a Recurrent Neural Network

(RNN)-based model using both weather (average wind speed, precipitation, maximal temperature, minimal temperature) and traffic data. The correlation between traffic flow and weather conditions plays an important role while developing DL models for traffic prediction [147], but Zhang et al. [415] did not consider the correlation patterns for the traffic prediction problem. In a recent study, Hou et al. [147] used a Stacked Autoencoder (SAE) and Radial Basis Function (RBF) neural network to predict traffic flow and showed improved results for the traffic prediction problem when combined with weather data.

## 6.3 Bi-LSTM for Traffic Flow Prediction

### 6.3.1 LSTM Networks

An LSTM network is a special kind of Recurrent Neural Network (RNN). Although RNNs have been successful in many sequence prediction tasks, they have issues learning long-term dependencies due to the vanishing gradient problem. This problem results from the gradient propagation of the recurrent network over many layers in deep architectures [369]. LSTM networks have been proposed to overcome these drawbacks. LSTMs were designed to incorporate memory units, and the network learns when to forget previous memories and update memories [183, 369]. Therefore, LSTM networks can cope with the correlation within time series in both the short and long term and are suitable for time series prediction problems such as traffic flow [422, 183]. A typical LSTM is composed of one input layer, one recurrent hidden layer, and one output layer. The memory block contains memory cells with self-connections memorizing the temporal state and three adoptive, multiplicative gating units. The input, output, and forget gates are used to control the information flow in the block. The three additional gates provide an analogue of read, write, and reset operations on the block. Multiplicative gates can learn to open and close, and thus LSTM memory cells can store information over a longer period of time [79, 183, 369]. Figure. 6.1 illustrates the architecture of an LSTM network.



**Figure 6.1** Architecture of a LSTM network

For example, if the historical traffic data sequence is denoted as  $x = (x_1, x_2, x_3, \dots, x_T)$  where,  $T$  is the prediction period, the hidden state of memory block  $h = (h_1, h_2, h_3, \dots, h_T)$  and the output

predicted traffic flow sequence  $y = (y_1, y_2, y_3, \dots, y_T)$  can be calculated by using the following equations.

$$y_t = W_{hy}h_t + b_y \quad (6.1)$$

$$h_t = H(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (6.2)$$

Where  $W$  denotes the weight matrices (e.g.,  $W_{xh}$  is the input hidden matrix),  $b$  denotes bias vectors and  $H$  is the hidden layer function.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (6.3)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (6.4)$$

$$c_t = (f_t c_{t-1} + i_t g(W_{xc}x_t) + W_{hc}h_{t-1} + b_c) \quad (6.5)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1}) + W_{co}c_t + b_o \quad (6.6)$$

$$h_t = o_t h(c_t) \quad (6.7)$$

The,  $i$ ,  $f$ ,  $o$ , and  $c$  are the input gate, forget gate, output gate, and activation vectors respectively.  $\sigma(\cdot)$  is the standard logistic sigmoid function,  $g(\cdot)$  is a centered logistic sigmoid function with the range  $[-2, 2]$  and  $h(\cdot)$  is a centered logistic sigmoid function with the range  $[-1, 1]$  as defined in the following equations.

$$\sigma(x) = \frac{1}{1 + e^x} \quad (6.8)$$

$$g(x) = \frac{4}{1 + e^x} - 2 \quad (6.9)$$

$$h(x) = \frac{2}{1 + e^x} - 1 \quad (6.10)$$

Generally, the learning ability of LSTMs increase with the number of layers. However, with too many layers, over-fitting is more likely [183].

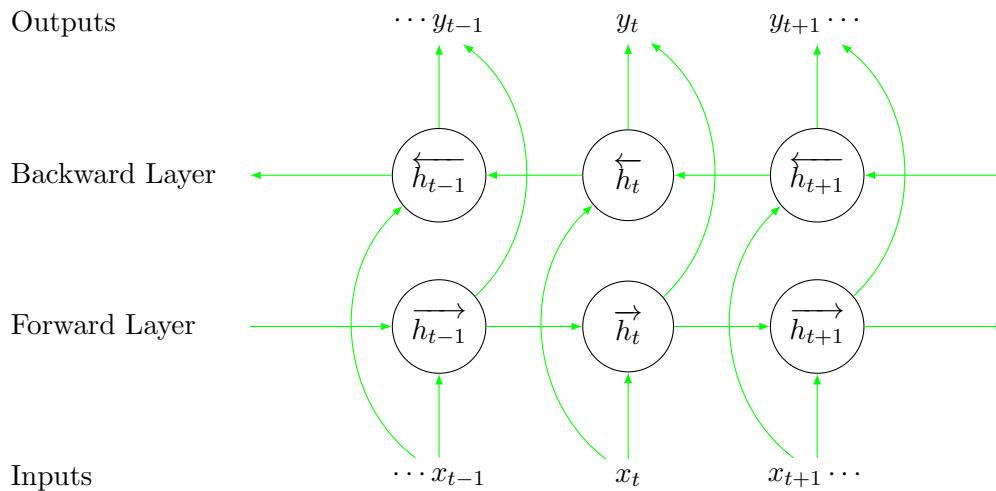
### 6.3.2 Bi-LSTM Networks

Bi-LSTMs are improved LSTMs and their structure enables two LSTMs to be trained for an input sequence. These networks process sequence data in both forward and backward directions with two separate hidden layers that are connected to the same output layer [79]. The architecture of a Bi-LSTM is shown in Figure 6.2. The forward layer output  $\vec{h}_t$  is calculated using inputs from time 1 to time  $T$ , while the backward layer output  $\overleftarrow{h}_t$  is calculated using reversed inputs from time  $T$  to time 1. The forward and backward layer outputs are calculated using the standard LSTM updating equations (see Equations (3) - (7)). The Bi-LSTM layer generates an output vector  $Y_T$

in which each element is calculated using the following equation.

$$t_y = \sigma(\vec{h}, \overleftarrow{h}) \quad (6.11)$$

where the  $\sigma$  function is used to combine two output sequences.



**Figure 6.2** The architecture of Bi-LSTM Network

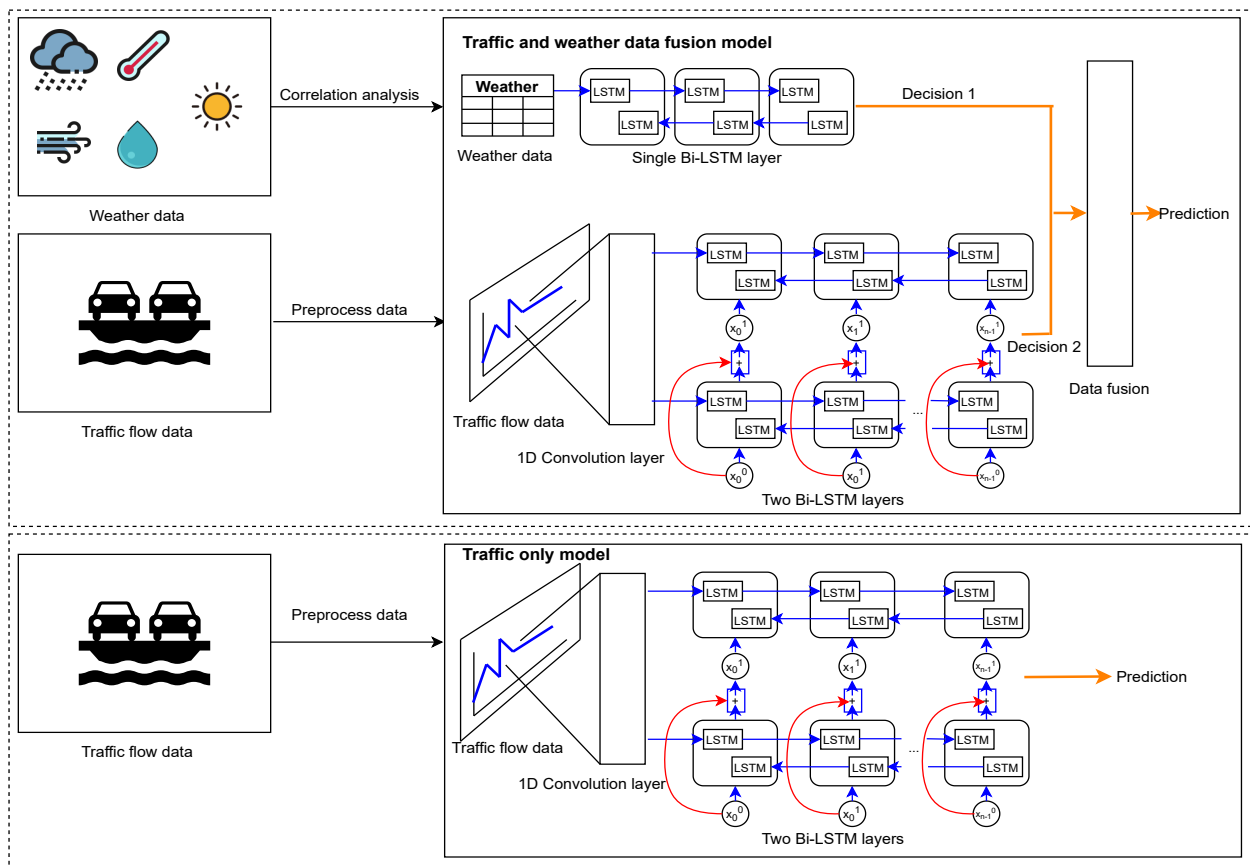
Our proposed architecture for the short-term traffic flow prediction problem in this paper has two components as follows.

1. A univariate time series model consisting of a 1-D Convolutional Neural Network Layer, stacked-Bi-LSTM layers (two), and a Dense Layer for traffic-only data and
2. A fused architecture based on a 1-D Convolutional Neural Network Layer, stacked-Bi-LSTM layers, and a Dense Layer for traffic data and one Bi-LSTM Layer and two Dense Layers for weather data, with the late fusion model using a Dense Layer.

The deep hierarchy structures of LSTM networks mostly result in the gradient vanishing problem [387]. Therefore, we limited the stacked networks in the proposed architecture to having only two layers. The fusion of data obtained from multiple sources is generally performed at two levels: feature level or early fusion, and decision level or late fusion. In the early fusion approach, the features are extracted from input data by the single analysis units and later combined. Comparatively, in late fusion approaches, each model analyzes individual features and outputs local decisions. The local decisions are then combined. Among these two approaches, Koesdwiady et al. [194] highlight that late fusion produces better results for traffic prediction tasks. Therefore, we adopted a late fusion method. Figure 6.3 illustrates the proposed architectures of the traffic-only model and weather data fused model.

### 6.3.3 Dataset

We obtained both traffic data and weather data for the duration of 03/04/2020 to 31/03/2021 for our experiments. The following sections describe the details of the datasets.



**Figure 6.3** Proposed model architecture for the traffic and weather data fused model and traffic only model.

### 6.3.3.1 Traffic Data

In New Zealand, daily traffic counts are obtained using inductive loops in the state highway network by the New Zealand Transport Agency (NZTA). However, a continuous flow at a certain location for a longer period was not available in the open data portal<sup>1</sup>. Therefore, we derived traffic flow data captured through CCTV cameras based on the methodology proposed by Algiriyage et al. [28]. Data were captured every minute using the images downloaded by NZTA Traffic Cameras API<sup>2</sup> at “SH73 Yaldhurst Rd” in Christchurch. A YOLO-based object detection model was used to detect and count the number of vehicles. The algorithm proposed by Algiriyage et al. [28] was used to identify a region of interest avoiding parking areas and footpaths (see Figure 6.4).

**Handling missing data:** We observed that there were some missing values in our traffic dataset. This occurred due to occasional missing images in the CCTV dataset as discussed by Algiriyage et al. [28]. We used the LSTM-based imputation method proposed by Tian et al. [369] to replace the missing data. Table 6.1 illustrates a sample of our preprocessed traffic data.

### 6.3.3.2 Weather Data

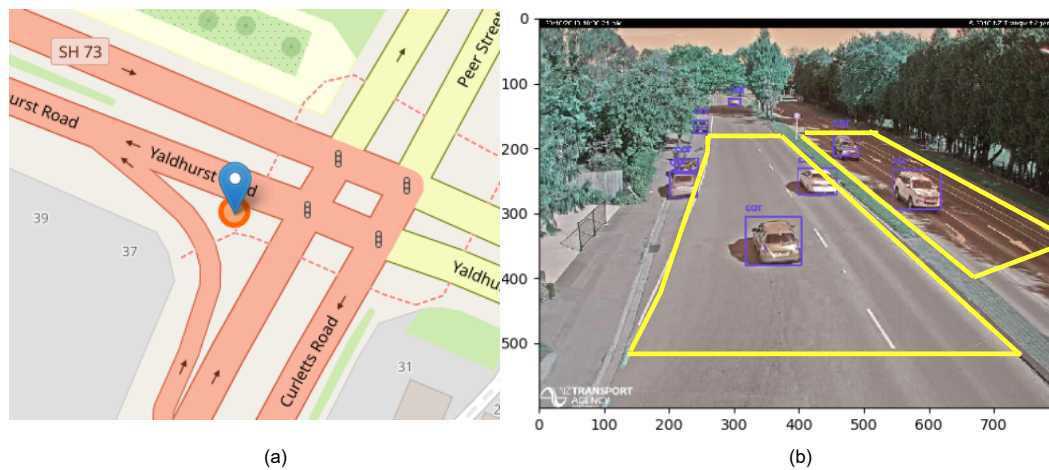
We obtained weather data from the national weather authority in New Zealand (MetService)<sup>3</sup>. A nondisclosure agreement with MetService was signed to access the 1-minute weather dataset in the

<sup>1</sup>NZTA Open Data portal,

<https://opendata-nzta.opendata.arcgis.com/datasets/tms-daily-traffic-counts-csv/about>

<sup>2</sup>NZTA Traffic Cameras API, <https://www.nzta.govt.nz/traffic-and-travel-information/infoconnect-section-page/about-the-apis/traffic-cameras/>

<sup>3</sup>MetService, <https://www.metservice.com/>



**Figure 6.4** (a) Selected location “SH73 Yaldhurst Rd” in the map. (b) Region of interest selection for the traffic flow counting using CCTV images

**Table 6.1** Traffic Flow Data **Table 6.2** Weather Data

Date Time	Flow	Date Time	AirTemp	DewTemp	...	WindLul	WindSpd
2020-04-03 01:44:00	1	2020-04-03 01:44:00	16.6	10.5	...	6.4	8.7
2020-04-03 01:45:00	5	2020-04-03 01:45:00	16.8	10.8	...	6.0	7.8
2020-04-03 01:46:00	0	2020-04-03 01:46:00	16.8	10.8	...	6.4	8.6
2020-04-03 01:47:00	0	2020-04-03 01:47:00	17.0	10.3	...	7.0	8.2
2020-04-03 01:48:00	2	2020-04-03 01:48:00	16.6	10.6	...	7.4	9.3

Christchurch region. The original weather data contained 20 parameters, including *Air temperature*, *Bright sunshine duration*, *Dew point temperature*, *Earth temperature at depth of 5cm*, *Earth temperature at depth of 10cm*, *Pressure MSL*, *Pressure QFE*, *Pressure QNH*, *Pressure sensors*, *Rainfall/precipitation amount*, *Relative humidity*, *Surface temperature*, *Average solar irradiance*, *Visibility*, *Wind counterclockwise*, *Wind clockwise*, *Wind direction*, *Wind gust*, *Wind lull* and *Wind speed*. The distribution of the different weather parameters for the considered duration is illustrated in Figure 6.5. Furthermore, Table 6.2 presents a sample of the weather data.

Not all weather parameters are directly correlated with traffic, and hence in an effort to identify the most relevant ones, we calculated the Pearson correlation score [147] using the following formula, given  $x$  and  $y$  as two target variables.

$$p(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

where:

$p(x, y)$  = correlation coefficient

$x_i$  = values of the  $x$  variable in a sample

$\bar{x}$  = mean of the values of the  $x$  variable

$y_i$  = values of the  $y$  variable in a sample

$\bar{y}$  = mean of the values of the  $y$  variable



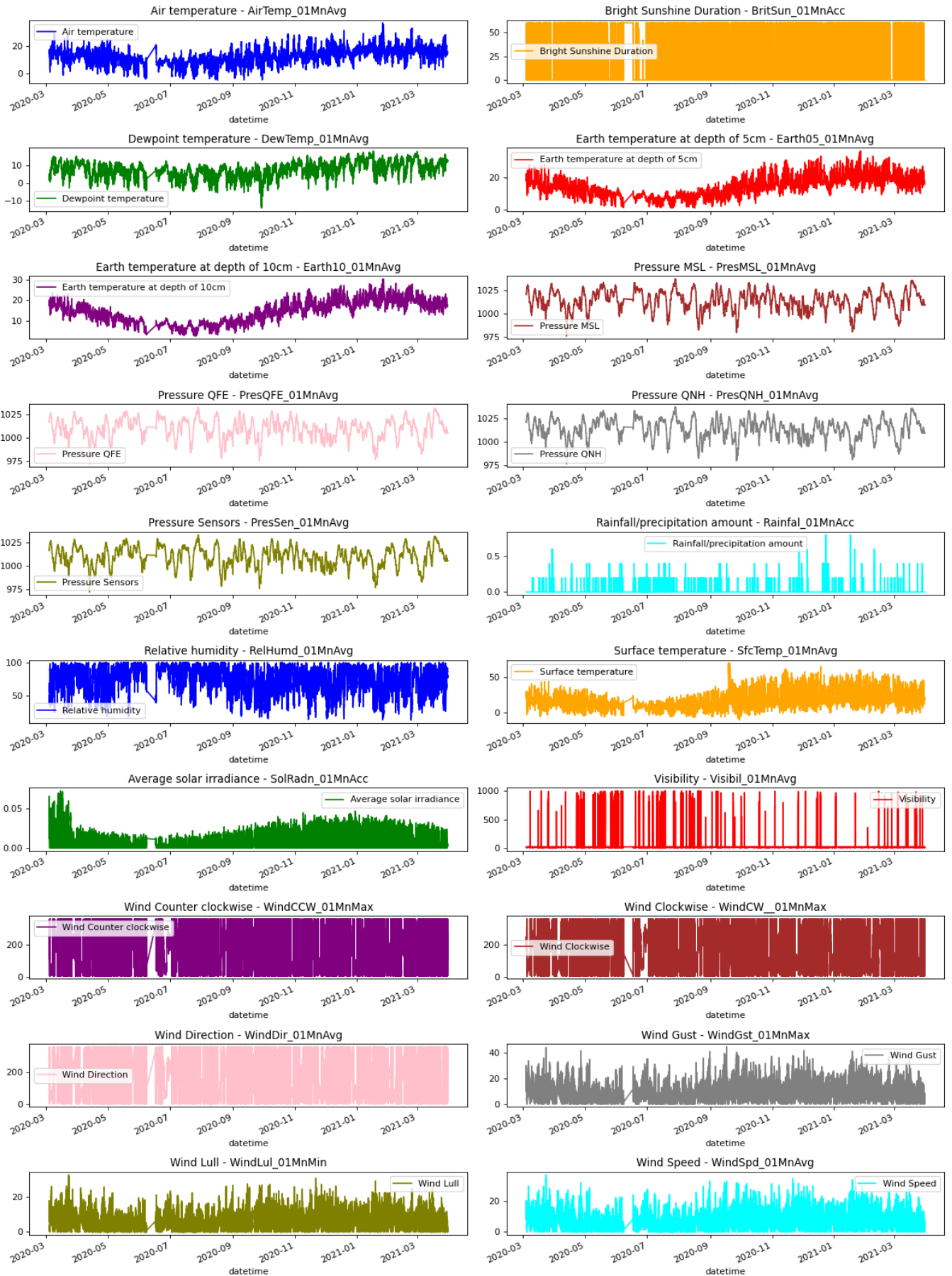


Figure 6.5 A plot of all-weather parameters for the considered duration.

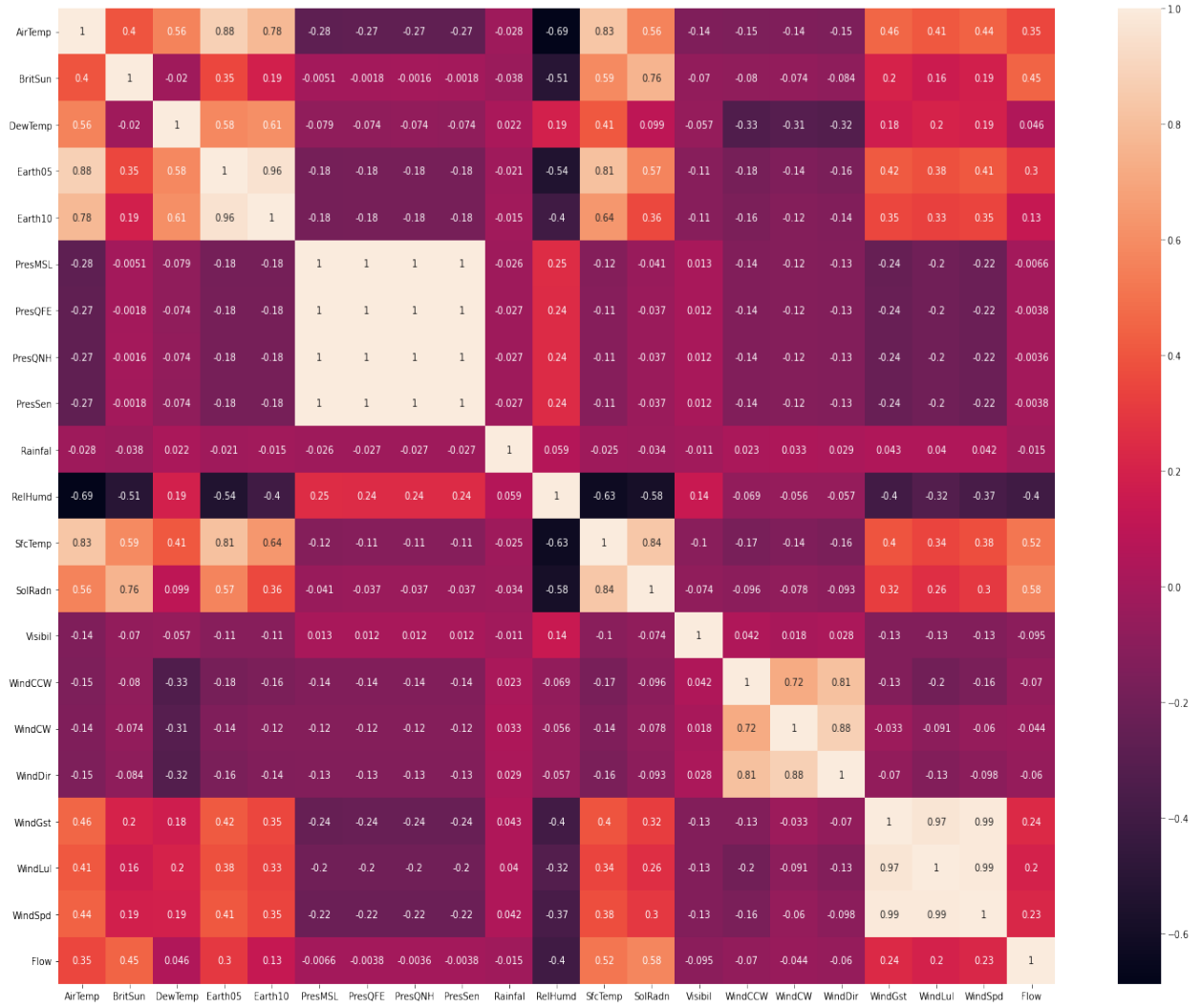


Figure 6.6 Pearson coefficient calculation.

The results of the Pearson correlation values (see Figure 6.6) were verified with meteorologists and transport experts in New Zealand. For this purpose, we created a short survey using both open-ended and closed-ended questions to verify the correlation coefficient. Based on both Pearson values and expert verification, we identified *Air temperature*, *Bright sunshine duration*, *Dew point temperature*, *Rainfall/precipitation amount*, *Relative humidity*, *Surface temperature*, *Average solar irradiance*, *Visibility* and *Wind gust* as weather parameters in our dataset that are most correlated with traffic data.

### 6.3.4 Problem Formulation

We initially model the traffic flow estimation problem as a univariate time series model with traffic-only data. Then we use the same model with weather parameters in a fusion model. The univariate traffic prediction in our research is based on the former parameters of 30 consecutive 1-minute intervals to predict the output flow in any subsequent time slice. Hence the output  $y$  of the prediction model can be expressed by the following formula:

$$y = f(x_1, x_2, x_3, \dots, x_{28}, x_{29}, x_{30}) \quad (6.12)$$

The decision-level fusion-based traffic flow prediction model considers traffic flow and weather parameters. For example,  $x_i$  represents a dataset on the time slice  $i$ , given  $n$  weather parameters. Therefore,  $x_i$  can be expressed as;

$$x_i = [x_i^{Flow}, x_i^{W_1}, x_i^{W_2}, x_i^{W_3}, \dots, x_i^{W_n}] \quad (6.13)$$

From the perspective of decision-level data fusion, the final flow prediction value is the fusion value of two decisions, so the output  $y$  of the combined model can also be expressed as follows:

$$y = f_{fusion}(y^{Flow}, y^{Weather}) \quad (6.14)$$

### 6.3.5 Experiments

Two experiments were designed to evaluate the performance of the proposed DL architectures. First, we evaluated the univariate time series model with traffic-only data for the short-term traffic flow prediction. Next, we considered the decision-level fused model that takes into account both traffic and weather data while predicting short-term traffic flow. All the models were implemented using the Python Keras library<sup>4</sup>. Both univariate and fusion models were trained for 100 iterations in the Mahuika High-Performance Computing (HPC) cluster of the New Zealand eScience Infrastructure (NeSI)<sup>5</sup>. We used 90% of the data for training and 10% for validation.

To evaluate the performance of predicted models three measures were used: Mean Absolute Error (MAE), Mean Square Error (MSE), and Root Mean Square Error (RMSE). These measures capture the gap between real values and predicted values and are defined in equation 6.15.

<sup>4</sup>Python Keras library, version 2.4.3 <https://keras.io/>

<sup>5</sup>New Zealand eScience Infrastructure, <https://www.nesi.org.nz/>

$$\begin{aligned}
MAE &= \frac{1}{n} \sum_{i=1}^n \left[ y_i^{predict} - y_i^{true} \right] \\
MSE &= \frac{1}{n} \sum_{i=1}^n \left( y_i^{predict} - y_i^{true} \right)^2 \\
RMSE &= \sqrt{\frac{1}{n} \sum_{i=1}^n \left( y_i^{predict} - y_i^{true} \right)^2}
\end{aligned} \tag{6.15}$$

We selected three baseline models including a standard LSTM network, Gated Recurrent Unit (GRU) [415] and the autoencoder (AE) model described by Hou et al. [147] to compare the results.

## 6.4 Experimental Results and Discussion

As mentioned in the literature review, external factors affecting deep learning-based traffic prediction systems have rarely been studied. As a result, this study set out to develop a DL architecture that would improve short-term traffic flow prediction and evaluate the model's performance with integrated weather parameters. Tables 6.3 and 6.4 provide the accuracy scores obtained for the univariate time series model, the fusion model, and the baseline models. There were 325,699 data points in total for the univariate time series model. According to the results in Table 6.3 the GRU network performed better than the standard LSTM model for the univariate time series model. This supports the previous research by Zhang et al. [415]. The AE model performed poorly for the univariate traffic flow prediction. However, the proposed stacked Bi-LSTM model has the best performance having 3.7155-MAE, 10.7213-MSE, and 5.8925-RMSE.

According to Table 6.4 the standard LSTM and GRU models performed poorly for the fusion model. The AE architecture introduced by Hou et al. [147] achieved 5.6795-MAE, 16.5825-MSE, and 8.8807-RMSE for our dataset (which captures weather and traffic flow every minute). In comparison, the best result they got for their own 5-min weather and the traffic data set was 9.49-MAE, 151.15-MSE, and 12.29-RMSE. This finding suggests that Hou's model performs well for short-interval datasets. Our proposed DL architecture with Bi-LSTM networks achieved the best results for the fusion model.

**Table 6.3** Traffic flow prediction results - univariate time series model

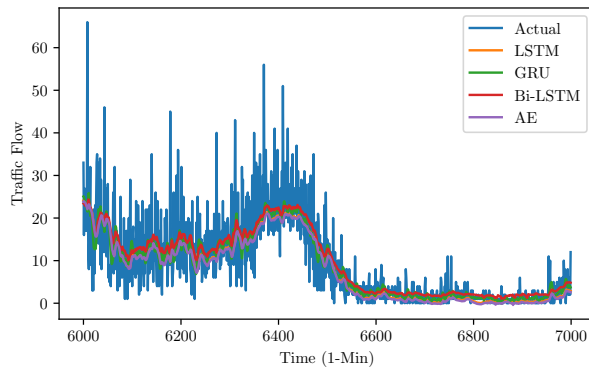
Algorithm	MAE	MSE	RMSE
LSTM	8.0974	16.2639	10.3846
GRU	7.3726	14.5863	9.7096
AE [147]	10.5814	18.7362	12.8993
Proposed Model	<b>3.7155</b>	<b>10.7213</b>	<b>5.8925</b>

**Table 6.4** Traffic flow prediction results - fusion model

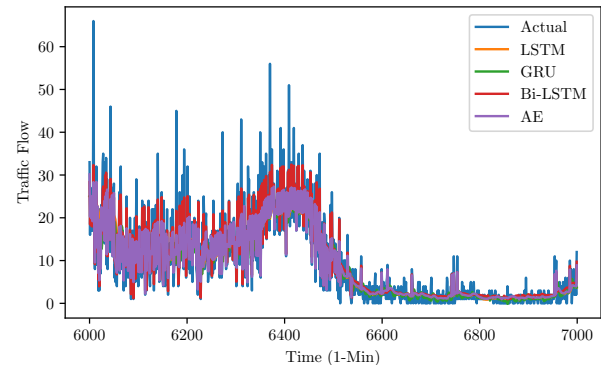
Algorithm	MAE	MSE	RMSE
LSTM	13.6807	24.7129	15.8918
GRU	14.6443	24.2898	13.8557

AE [147]	5.6795	16.5825	8.8807
Proposed Model	<b>2.1343</b>	<b>10.7612</b>	<b>3.4295</b>

The Figures 6.7a and 6.7b visualize the actual and predicted values made by our univariate time series model and the fusion model.



(a) Actual and prediction values for the univariate model



(b) Actual and prediction values for the fusion model

There are several possible explanations for these results. Firstly, the results of this study suggest that the 2-way learning capability of Bi-LSTM networks best suits the short-term traffic flow prediction problem. Secondly, stacking multiple LSTM hidden layers makes the model deeper and provides more accurate feature learning. Therefore, increasing the depth of the model has resulted in better performance than the other models. This finding demonstrates that the DL models for short-term traffic flow prediction perform well with more additional data, such as weather parameters. The evidence from this study suggests that future DL models for short-term traffic flow prediction problems must integrate external parameters for better results. However, further work is required to establish this with more data obtained from different locations.

## 6.5 Conclusions

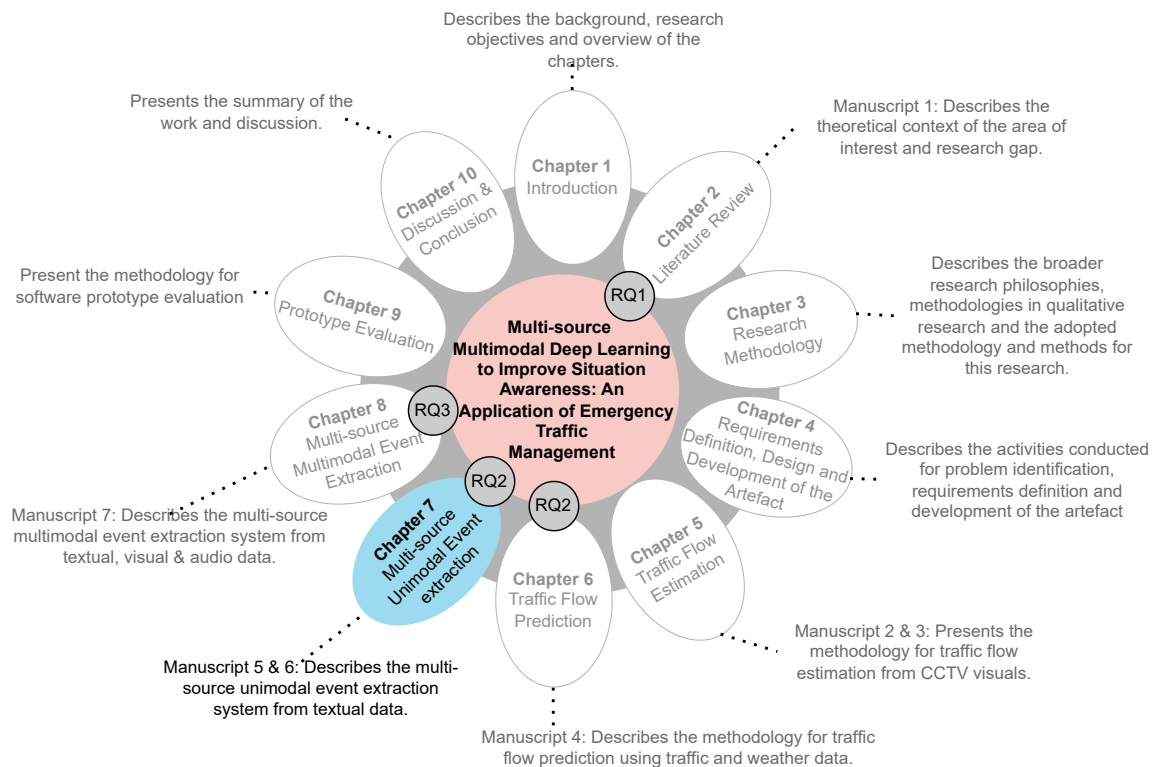
Today's Intelligent transportation systems (ITS) rely on short-term traffic flow prediction to mitigate traffic congestion. However, traffic flow prediction has become a challenging issue due to the availability of large-scale traffic data, periodic characteristics, and other external factors affecting the traffic flow. This research was undertaken to evaluate the performance of short-term traffic flow prediction tasks using a novel DL architecture. We conducted experiments using traffic-only data (a univariate time series model) and a fused model with both traffic and weather data for the traffic flow prediction problem. The results of this investigation show that the proposed stacked Bi-LSTM model predicts the traffic flow more accurately. Moreover, the accuracy of the DL-based prediction model improves when it is integrated with weather data to assist the traffic flow prediction task. The present study provides additional evidence with respect to the recent study by Hou et al. [147].

Several limitations to this study need to be acknowledged. Firstly, the experiments considered a one-year dataset. Generally, DL models work best with larger amounts of data than this. Secondly, the seasonal variations in our dataset were also limited. As a result, with more historical traffic and

weather data available, the model may be trained for more significant seasonal variations in the future. Thirdly, the proposed model has not used any regularization techniques to avoid overfitting. Overfitting occurs when the model trains well for the training data but is unable to generalize well for new unseen data. The use of proper regularisation techniques will help to avoid overfitting while improving the results. Finally, we used the LSTM-based method proposed by Tian et al [369] for missing data imputation. However, several studies in the literature provide multiple methods for imputation of missing data, thus future work needs to identify the most appropriate of these interpolation techniques. Moreover, further research needs to be done to evaluate the performance of the traffic prediction problem with other variables such as seasonal changes, extreme weather conditions, and natural disasters.

# Chapter 7

## DEES - A real-time system for event extraction from disaster-related web text



This chapter presents the fifth manuscript that describes the third component of the software artefact. The chapter aims to answer the second research question, “How can data from multiple sources be fused to support disaster response?”. The research explains how text data from various sources can be combined to validate SM content while extracting real-time event templates for disaster responders’ SA. The third component requires disaster-related tweet classification, and the conference paper (sixth manuscript) that outlines the process is detailed in Section 7.5.

## Abstract

The rapid growth of Internet-based communication technologies in the form of Social Media (SM) and associated mobile applications has enabled people to share information related to disaster events in “real-time” as they unfold. People are increasingly using SM platforms to report situational information during disasters, such as critical needs, dead or injured people, and property damage. Despite their usefulness, the majority of this pertinent data is not available to humanitarian organizations during emergencies, mainly due to several data processing and data quality issues. The proliferation of online news media has also led to the exchange of a massive amount of information during disasters, mostly validated by official sources. The integration of SM data with online news reports can provide filtered information while adding more details on the progress of an event that is already published in online news. This research project introduces *Disaster Event Extraction System (DEES)*, a real-time system for extracting disaster events from both online news and tweets. DEES is evaluated on a dataset collected during the Nepal earthquake in 2015. Our results suggest that integrating both SM and news text data improves the event extraction system’s performance compared to using SM data alone. A demonstration of DEES is available at: <https://mu-clab.github.io/>.

## 7.1 Introduction

Disasters, whether natural or man-made, often impose a significant impact on societal or organizational infrastructures and cause destruction to communities. Today, with the rapid advancement of communication technology, victims, responders, and volunteers use mobile devices to provide real-time situation updates through SM and related applications [163]. Therefore, a vast amount of data are generated within seconds of a disaster. This has provided multiple opportunities associated with the goal of quickly publicizing information for the government, professional organizations, responders, and even the general public at large [339, 185]. Furthermore, SM-based communication has challenged traditional media systems, information pathways, and hierarchies, with most real-world events in recent years have been first reported in SM [301, 181].

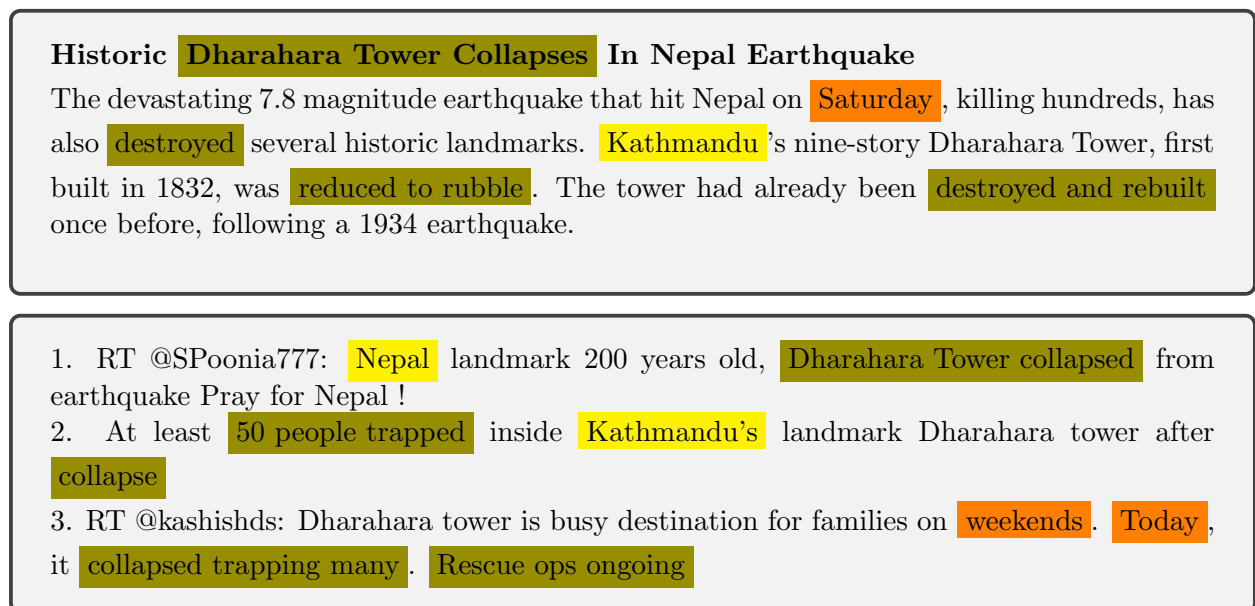
Emergency responders need actionable information associated with disaster situations in order to understand the development of a disaster and to facilitate decision-making, policy formulation, response, and resource allocation [394]. Processing large-scale crisis data in real-time can help to see the “big-picture” of a disaster, known as situation awareness [98]. People at disaster sites provide quick updates using SM platforms, resulting in a massive volume of data being generated within seconds of an event. As a result, in recent years, there has been an increasing interest in using SM data for crisis event detection [124], understanding public sentiment [130], damage assessment [164, 412] and actionable information gathering for improving situation awareness [339, 185]. Despite these benefits, timely processing and analyzing an overwhelming amount of SM data brings several challenges. The data included in SM platforms come in different modalities, such as text, image, audio, and video, and are inherently noisy. Researchers who have analyzed SM text data during disasters have found the content to be informal, mostly using colloquial language, being very brief with casual acronyms, and sometimes with non-literal language devices like sarcasm, metaphors, and double entendre. Most importantly, rumors, misinformation, and false information are very common among SM messages [163]. As a result, a significant amount of timely, valuable information



on SM platforms gets wasted without being properly utilized. Therefore, conflicting and uncertain volumes of incoming SM data during emergencies have to be effectively sorted and prioritized to be used for disaster response.

Multiple data sources can be aggregated together and leveraged to generate collective intelligence to deal with emergency events, rather than relying on SM as a single source. This would help to improve the accuracy of information extracted while also cross-validating the information provided through SM channels. For example, during a disaster, news organizations provide updates on their websites, which are typically validated by official services [297, 362, 300]. The integration of tweets with news articles provides opportunities to triangulate and validate the information.

Figure 7.1 presents a news article and three tweets extracted during the Nepal-Earthquake in 2015. The information provided in tweets can be validated from the linked online news. Furthermore, the tweets provide more contextual information than the news reported.



**Figure 7.1** The First box consists of a news article, the title (bold), and the second box consists of four tweets. Highlighted phrases in each sentence represent the answers to **what**, **when** and **where** questions

Extracting information from news and tweets has been studied mostly in static datasets [120, 9, 81]. So far, however, there has been little discussion of combining tweets with news in real-time to improve situation awareness of disaster responders, as it is challenging due to their differences in text length and reporting style [297]. Therefore, this study explores the extraction of semantic, spatial, and temporal descriptive features from SM and online news in real-time using Natural Language Processing (NLP) techniques. Therefore we develop “DEES” - an automated system for jointly extracting knowledge in a structured representation, specifically identifying the answers for semantic - *what*, temporal - *when*, and spatial - *where* questions from online news and matching tweets to support disaster response, as per the following example.

<b>What</b>	Dharahara tower collapse
<b>When</b>	Saturday (25 April 2015) *combining the reported date
<b>Where</b>	Kathmandu

To the best of the authors' knowledge, DEES is the first to extract aggregated disaster events

from news and SM feeds in real-time. We consider a streaming setting in which news articles and tweets are continuously extracted and analyzed. The architecture of our system has six main modules, namely news and tweet extraction, related tweet identification, noise filtering, clustering, candidate extraction, and candidate scoring. We introduce a new location relatedness score in identifying the geolocation of the event. Furthermore, a novel cross-media reference score is proposed for candidate scoring. Our system is also extensible (i.e. additional event types can be added with minor modifications). The methodology proposed in this study is evaluated using a news and tweet dataset related to the Nepal earthquake in 2015.

The remainder of this paper is structured as follows: Section 7.2 discusses related work. Section 7.3 presents the methodology to implement real-time event extraction system. Section 7.4 describes the evaluation process. Finally, Section 7.11 brings the concluding remarks and future work.

## 7.2 Related Work

Most of the data published online are in text format and are unstructured [390]. The main objective of an event extraction system is to create a structured representation of an event from unstructured text [297]. Previous systems that address the problem of automated extraction of news events include EventMiner [122], Giveme5W [128] and Giveme5W1H [125]. The intention of these systems is more towards answering journalistic 5W1H questions (who did what, when, where, why, and how) for news analysis, including grouping of articles about the same event, news aggregation, and news summarization [125].

There are multiple systems described in the literature that extract crisis-relevant information from SM such as Tweedr [36], CrisisTracker [330], Twitcident [2] and AIDR [158], all of which processed, classified and clustered tweets for disaster response. More recent attention has also focused on detecting and extracting large-scale (e.g., earthquake) and small-scale (e.g., vehicle crash) disaster events from SM posts [161, 160, 31, 414, 339, 124, 288, 396, 29]. Dhavase et al. [86] explored methods for extracting locations from crime and disaster-related tweets using NLP methods including Named Entity Recognition (NER), rule-based pattern matching, and gazetteer matching. Ha et al. [123] extracted location information using keyword filtering and using local name words. However, the sudden, massive burst of SM posts during an emergency brings multiple challenges such as repeated posts, informal content, fake news, and rumors that prevent them from being used directly for disaster response [163, 393].

Multiple studies have specifically explored the extraction and analysis of disaster events from news web texts [161, 168, 129]. For example, early work by Piskorski et al. [362, 300] described a system for real-time crisis event extraction from online news. Also, Valero et al. [365] described a system based on machine learning methods to improve the acquisition of disaster information from online news reports. Wang et al. [395] applied ontologies to extract spatio-temporal and semantic information on typhoons from web news reports. Stewart and Wang [354] automatically extracted spatial and temporal references from news web texts and represented the spatio-temporal characterizations of events in a dynamic mapping environment.

Methods for combining SM posts with news articles have been studied for multiple applications, such as determining interesting news articles [9], identifying trending topics [81], web text classification [172] and finding a related news wire document to a given tweet [120]. However, these studies are based on static datasets and do not address the problem of real-time event extraction,

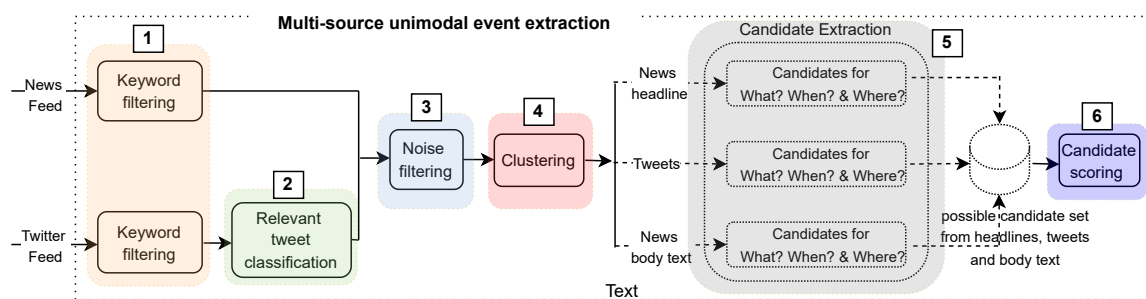
considering the dynamic nature of news. Moreover, they have only considered tweets that have news Uniform Resource Locator (URL) links in their body [81, 297]. This approach of filtering tweets having news URLs removes around 78% of tweets considered for the application [297].

So far, very few studies have focused on the influence of different web text sources and the ways in which social media data can be better filtered using news for disaster information acquisition. The structure of sentences, the information concerns, and the reporting perspectives vary among different web text sources [129]. Therefore, the characteristics of disaster information from varied web texts need to be further explored. Another challenge for the evaluation of disaster event extraction is that the evaluation data sets used in previous research are not publicly available [129, 297]. In addition, there are very few articles in the literature describing adequate details of event extraction systems that allow for re-implementation by other researchers [125].

Therefore, our overarching objective during this research is to explore methods to effectively use SM content to assist disaster response by cross-validating using online news from main online news providers in New Zealand. Closer to our objective is the work of Verma et al. [380] and Petroni et al. [297]. Verma et al. [380] analyzed a tweet and news corpora collected during the Nepal 2015 earthquake. They paired news with tweets as a supervised classification task using a Support Vector Machine (SVM) algorithm with a precision of 0.47. Petroni et al. [297] presented an online event extraction system using both news articles and tweets, trained to recognize breaking news events by co-referencing both media. Our approach is different from both studies in multiple ways. Firstly, we group related news articles with tweets dynamically using average sentence vector similarity where Petroni et al. [297] cluster tweets separately and then identify matching news articles through co-reference. Our one-time clustering approach reduces time while improving accuracy for a real-time application. Secondly, both Petroni et al. [297] and Verma et al. [380] use rule-based and topic models to filter relevant tweets, while we train a deep learning algorithm for related tweet filtering. Thirdly, Verma et al. [380] use an SVM classifier for identifying the impact factor, whereas we use dependency parsing considering the syntactic structure of the text.

### 7.3 DEES : Description of Methodos and System

DEES is a real-time event extraction system as described in Section 7.1. The extraction process is illustrated in Figure 7.2 which comprises six core modules described below.



**Figure 7.2** Real-time event extraction process. 1. News and tweet extraction, 2. Relevant tweet identification, 3. Noise filtering, 4. Clustering, 5. Candidate extraction, and 6. Candidate scoring.

1. **News and tweet extraction:** collect online news and tweets as a scheduled task and filter using key-word based heuristics.

2. **Relevant tweet identification:** identify relevant tweets using a supervised algorithm.
3. **Noise filtering:** remove noise from text such as symbols and web links.
4. **Clustering:** group news headlines with tweets.
5. **Candidate extraction:** extract words/phrases answering *what*, *when* and *where* questions from news headlines, tweets and news bodies.
6. **Candidate scoring:** select the best candidates from the possible candidate set in step 5.

This process of extracting disaster events in real-time using the six core modules is outlined in detail in the following sections.

### 7.3.1 News and tweet extraction

A Linux scheduled task (cronjob) is used to extract online news and tweets every 20 minutes. The system collects online news from three main online news providers in New Zealand namely, rnz news<sup>1</sup>, nzherald news<sup>2</sup> and stuff news<sup>3</sup> through their rss feeds.

Tweets are extracted in real-time from the Twitter streaming Application Programming Interface (API) using the python tweepy library<sup>4</sup>. We set the geographical boundary to New Zealand to collect tweets generated by the users within the area corresponding with the news feeds. Additionally, to avoid exceeding API limits, the maximum number of tweets downloaded per iteration of the extraction system is limited to 1000 tweets. A rule-based matcher is used to retrieve tweets and news headlines related to keywords. Currently, the algorithm captures event types, such as traffic and transport (e.g., crash), weather (e.g., rain, cyclone, flood, storm, snow, wind, hurricane, tornado), societal (e.g., armed conflict, terrorism), fire, earth slip, earthquake, and pandemic. The keyword-based rule matcher used in the system can be modified to limit the extraction for a single event or to add additional event types.

### 7.3.2 Noise filtering

The noise filtering module has three components to remove noise from news headlines, tweets, and news body text. Special characters are removed except for letters and numbers from news headlines. To clean up news body text, we remove special characters, videos, images, embedded tweets, advertisements, and other links.

Duplicate tweets, non-English tweets, and Re-Tweets (e.g., RT@ user:) are not considered for processing. Tweet text is cleaned by removing hashtags, links, words having a length of less than three characters, stop words, special characters, and short sentences (fewer than three words). We also expand abbreviated words (e.g., “ur” → “your”, “2morow” → “tomorrow”). Then the standard text preprocessing techniques such as sentence splitting, lemmatization, part-of-speech (POS) tagging, dependency parsing, and Named Entity Recognition (NER) are applied using the python spaCy library<sup>5</sup>.

---

<sup>1</sup>rnz news, <https://www.rnz.co.nz/>

<sup>2</sup>nzherald news, <https://www.nzherald.co.nz/>

<sup>3</sup>stuff news, <https://www.stuff.co.nz/>

<sup>4</sup>Python tweepy library version 3.9.0, <https://pypi.org/project/tweepy/>

<sup>5</sup>Python Spacy version 2.3.2, <https://spacy.io/>

### 7.3.3 Relevant tweet identification

Not all tweets are useful for analysis, and it is, therefore, important to distinguish relevant from irrelevant tweets. Thus, this module filters relevant tweets for the analysis tasks. A considerable amount of literature has been published on using ML [59, 336, 216] and DL approaches [267, 17, 401] for identifying useful twitter posts for disaster response tasks. We adopted a Bi-directional LSTM model for relevant tweet classification as described by Algiriyage et al. [24].

### 7.3.4 Clustering

The aim of this module is to group similar news headlines and tweets based on the content. Text data is converted into weighted vectors before clustering [346]. The highest accuracy for clustering was achieved while using the Word2Vec model for converting sentences to vectors, known as word embeddings. In the Word2Vec model, vectors are learned in such a way that words that have similar meanings will be located near each other in the vector space [254]. Thus, the semantic relationship between words is preserved. Density-based spatial clustering of applications with noise (DBSCAN) algorithm implemented in the python sklearn library<sup>6</sup> is used to group similar news headline and tweet vectors together [100]. This choice was made as the DBSCAN algorithm does not require the number of clusters to be provided [77]. Instead, the algorithm decides the number of clusters it can generate given the data. The DBSCAN algorithm groups together data points that are close to each other based on a distance measurement and a minimum number of points. Therefore, the algorithm requires two parameters to be specified by the user: *eps* - how close points should be to each other to be considered a part of the same cluster and *minPoints* - the minimum number of points needed to form a dense region [100, 77]. At the end of the clustering phase, a collection of clusters with news headlines and tweets or clusters with only tweets is produced.

### 7.3.5 Candidate extraction

We considered clusters with both news and tweets for the candidate extraction, and clusters with only tweets are discarded. This is done to ensure that there are news sources to validate the accuracy of social media content. Our system extracts news body text from those headlines using the python newspaper3k<sup>7</sup> package. The candidate extraction module extracts the answers for *where*, *when*, and *what* questions from news headlines, tweets, and news body text separately for each cluster.

To extract *where* candidates, we use the Named Entity Recognition (NER) implementation in spaCy<sup>8</sup>. The system extracts items tagged as Geo-Political Entities (GPEs) (e.g., countries, cities, states), Non-GPE locations (LOC) (e.g., mountain ranges, bodies of water), and FAC (e.g., buildings, airports, highways, bridges) by the NER tool.

*When* candidates, including both relative and absolute time references, are extracted from the text using both NER and regular expressions (see Table 7.1). For relative time expressions, our system includes the news reported date and time or tweet generated date and time to extract the precise temporal information (see Table 7.1).

---

<sup>6</sup>DBSCAN algorithm, <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>

<sup>7</sup>Python newspaper3k, <https://newspaper.readthedocs.io/en/latest/>

<sup>8</sup><https://spacy.io/api/annotation#named-entities>

**Table 7.1** Absolute and relative time expressions

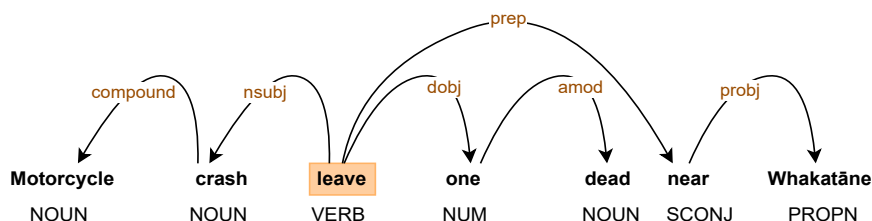
Type	Expression	Example
Absolute time	Day Month Year Hour:Minute	21 May 2019 21:27
Relative time	today, afternoon, morning	3:50 this afternoon, 10 AM in the morning

Dependency parsing is used to extract the part of the text to answer the *what* question. Dependency parsing analyzes the grammatical structure of a sentence, establishing relations among words [71]. Our algorithm starts from the ROOT node and selects multiple sub-trees, and builds a new text from that. For example, first, we check if the ROOT node has a direct object (*dobj*). In this case, subtrees are used to extract any children. Finally, the child node texts and the direct object are combined with the ROOT node as the answer to the *what* question. In the event that the ROOT node does not have a *dobj*, the ROOT node itself will be extracted. Our algorithm first splits the news body text into sentences before applying dependency parsing. Therefore, candidates are extracted for each sentence. Table 7.2 shows the dependencies of the news headline “Motorcycles crash leaves one dead near Whakatāne”.

**Table 7.2** Dependencies and POS of the news headline “Motorcycles crash leaves one dead near Whakatāne”

word	lemma	dependency	POS
Motorcycles	motorcycle	compound	NOUN
crash	crash	nsubj	NOUN
leaves	leave	ROOT	VERB
one	one	dobj	NUM
dead	dead	amod	NOUN
near	near	prep	SCONJ
Whakatāne	Whakatāne	pobj	PROPN

Figure 7.3 presents the dependency parsing diagram for the same headline. In this example, the *ROOT* is “leave” and the *dobj* is “one”. The *dobj*, “one” has a single child “dead”. Therefore, our algorithms produce “leave one dead” as the answer to *what* question.

**Figure 7.3** Dependency parsing of news headline “Motorcycles crash leaves one dead near Whakatāne”

To explain our approach to candidate extraction, we consider a cluster generated from our extraction system of news and tweets related to a vehicle crash event that occurred in the Bay of Plenty, New Zealand. Tables 7.3, 7.4 and 7.5 show candidate extraction results for the news headlines and tweets, and news body text respectively.

**Table 7.3** Candidates extracted from news headlines

News Headline	Where?	When?	What?
'One dead after two motorbikes crash in Bay of Plenty'	Bay of Plenty	-	dead
'Motorcycles crash leaves one dead near Whakatāne'	Whakatāne	-	leave one dead

**Table 7.4** Candidates extracted from tweets

Tweet	Where?	When?	What?
'I found a colleague was in the serious crash in Bay of Plenty today. Here's hoping it's a speedy recovery'	Bay of Plenty	today	found
'Farmer dies in bike crash in Bay of Plenty'	Bay of Plenty	-	die

**Table 7.5** Candidates extracted from news body text

News Body Text	Where?	When?	What?
'One person has died after two motorbikes crashed in the Bay of Plenty. The crash occurred at 3.15pm on Bell Rd in Nukuhou, south of Whakatāne, police said. Another person suffered minor injuries in the crash. WorkSafe had been advised and the Serious Crash Unit was in attendance, police said'	Bay of Plenty, Bell Rd, Nukuhou, Whakatāne	3.15pm	die, occur, suffer, minor injuries
'One person has died and another has minor injuries following a serious crash in Nukuhou, near Whakatāne in Eastern Bay of Plenty. The crash involved two motorcycles and was reported around 3.15pm., The Serious Crash Unit and WorkSafe are attending the scene'	Nukuhou, Bay of Plenty, Whakatāne, Bay of Plenty	3.15pm	die, attend the scene, involve two motorcycles, attend

### 7.3.6 Candidate scoring

Candidate scoring is concerned with determining the best candidates after extracting the candidate answers to the 3W questions. We assume that online news is reported using the inverted pyramid structure; a system of news writing that arranges facts in descending order of importance [278]. Candidates are scored based on their position, frequency, and appearance across both news and tweets (cross-media reference). Furthermore, a novel location relatedness score is introduced in identifying the best candidate for the *where* question. Table 7.6 illustrates the different scores used for candidate scoring, and each of the scores is defined below.

*Position score* ( $S_{pos}(C)$ ): A higher score is assigned if candidates are found early in the text. For occurrences in the first sentence of the body text or in the headline, a score of 1 is assigned.

**Table 7.6** Scores used to determine the best candidates among the candidate set extracted from news headlines, body, tweets, and image captions

Where	When	What
Position Score	Position Score	Position Score
Frequency Score	Frequency Score	Frequency Score
Cross-media Reference Score	Cross-media Reference Score	Cross-media Reference Score
Location Relatedness Score		

For occurrences in subsequent sentences, the score follows an exponential decay, decreasing with an increase in position,  $p$ ,  $S_{pos}(C) = e^{(-dp)}$ , with  $e$  being the exponential constant and decay coefficient,  $d > 0$  [278]. If we select logarithmic decay (i.e.  $d = \log(2)$ ), we divide the score by half whenever we move farther away from the headline or first sentence in the lead paragraph. For example, consider the location candidates in the headlines and body in the Tables 7.3 and 7.5. Let  $X = [x_1, x_2, \dots, x_n]$  be the vector of news headlines and  $Y = [y_1, y_2, \dots, y_n]$  be the vector of the news body text in the cluster. We calculate the position score for candidates presented in a headline  $S_{pos}(C_h) = \sum_{i=1}^n S_{pos}(C_{x_i})$  and the position score for candidates in body text  $S_{pos}(C_b) = \sum_{i=1}^n S_{pos}(C_{y_i})$ . Finally, the full position score is obtained using Eq. 7.1.

$$S_{pos}(C) = S_{pos}(C_h) + S_{pos}(C_b) \quad (7.1)$$

Table 7.7 shows an example of full position score calculation for *where* candidates of the example news headlines and body text in Tables 7.3 and 7.5.

**Table 7.7** Position score calculation for *where* candidates of the news in Tables 7.3 and 7.5

Locations in headline	$S_{pos}(C_h)$	Locations in body	$S_{pos}(C_b)$	$S_{pos}(C)$
Bay of Plenty	1.0	Bay of Plenty	2.0	3.0
-	-	Bell Rd	0.5	0.5
-	-	Nukuhou	1.5	1.5
Whakatāne	1.0	Whakatāne	1.5	2.5

*Frequency score* ( $S_{freq}(C)$ ): We score candidates by their frequency of occurrence in news articles and tweets. The frequency values are transformed by scaling them between 0 and 1. Table 7.8 illustrates frequency scores for *where* candidates of the news and tweet examples in Tables 7.3, 7.4 and 7.5.

**Table 7.8** Frequency score calculation for *where* candidates of the news and tweet examples in Tables 7.3, 7.4 and 7.5

Location	Frequency	$(S_{freq}(C))$
Bay of Plenty	5	1.0
Bell Rd	1	0.0
Nukuhou	2	0.25
Whakatāne	1	0.0

*Location relatedness score* ( $S_{rel}(C)$ ): One of the main problems in location identification from text is the location ambiguity where several distinct locations have the same name (e.g., Kingston in



New York and Kingston in Jamaica) [297]. Our algorithm resolves ambiguity issues based on spatial proximity clues. We introduce a new location relatedness score to identify the most accurate spatial candidate. First, the location names are transformed into coordinates using Nominatim<sup>9</sup>, which uses free data from OpenStreetMap<sup>10</sup>. Then, the geodesic distance among all location candidate pairs is calculated using Python geopy library<sup>11</sup>. Finally, we assign a higher score for the closest location candidates and lower scores for the distant location candidates following the logarithmic decay function. Table 7.9 illustrates the location relatedness scores obtained for *where* candidates.

**Table 7.9** Location relatedness score calculation for *where* candidates of the news and tweet examples in Tables 7.3, 7.4 and 7.5.

Location pairs	Pair wise distance	$S_{rel}(C)$
Bay of Plenty, Bell Rd	459.9687	0.03125
Bay of Plenty, Nukuhou	14.0604	0.50000
Bay of Plenty, Whakatāne	3.0771	1.00000
Bell Rd, Nukuhou	457.6903	0.03125
Bell Rd, Whakatāne	461.4393	0.03125
Nukuhou, Whakatāne	11.8114	0.50000

*Cross-media reference score* ( $S_{cross}(C)$ ): The system analyses the occurrences of candidates across tweets and online news and assigns a score of 1 if the candidates appeared in both media. Otherwise, the score is 0. For example, the location Bay of Plenty is in both tweets and online news and thus gets a score of 1.

Finally, we obtain the full candidate score by summing the individual score values (see Eq. 7.3) and select the highest-scoring candidate as the correct answer to the question. As can be seen from Table 7.10, the final score is designed to be between 0 and 1. Therefore, the system chooses the Bay of Plenty as the most suitable candidate for the *where* question in our example. Similarly, the scores mentioned in Table 7.6 are calculated to select the candidates for the *what* and *when* questions.

$$S_{full}(C) = S_{pos}(C) + S_{frq}(C) + S_{rel}(C) + S_{cross}(C) \quad (7.2)$$

**Table 7.10** Final *where* candidate selection of the news and tweet examples in Tables 7.3, 7.4 and 7.5.

Location	Final Score
Bay of Plenty	1.0
Bell Rd	0.0
Nukuhou	0.5
Whakatāne	0.5

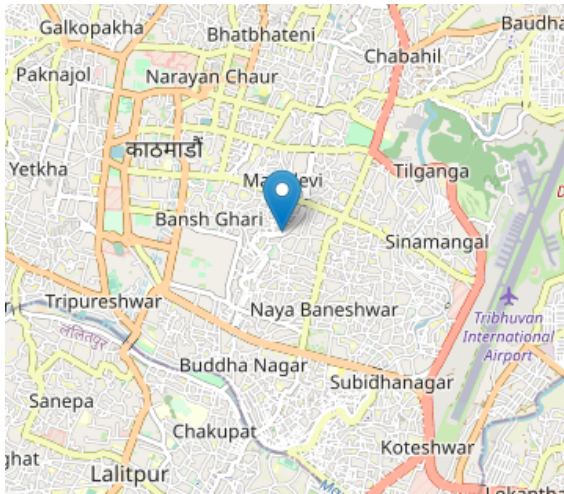
A demonstration of DEES with its functionality is available at <https://mu-clab.github.io/>. Figure 7.4 shows the first screen of DEES. The user initially sees a map with the precise location of the event marked. The user can then view the event details by clicking on the location marker, including what, when, and where information (see Figure 7.5). Finally, the user can

<sup>9</sup>Nominatim version 3.5.1, <https://github.com/osm-search/Nominatim>

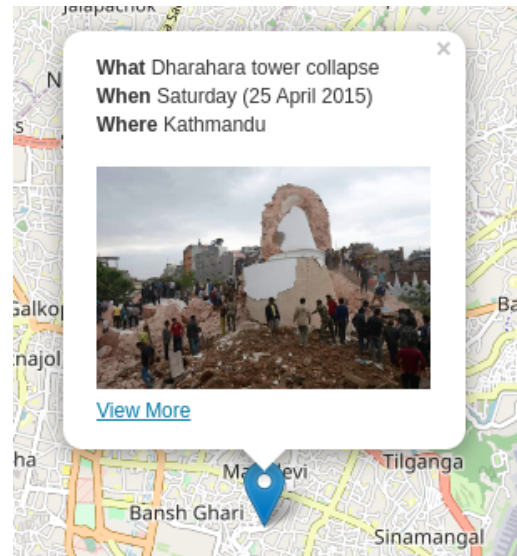
<sup>10</sup>OpenStreetMap, <https://www.openstreetmap.org/#map=2/-41.2/-6.6>

<sup>11</sup>geopy 2.2.0 <https://pypi.org/project/geopy/>

access additional information about the tweets and news articles that were used to create the event template by clicking the "view more" button as illustrated in Figure 7.6.



**Figure 7.4** Home screen showing events



**Figure 7.5** Screen showing more details of news and tweets

## 7.4 Evaluation

Evaluation experiments focused on evaluating the performance of DEES when using tweet-only data and tweet + news data. Therefore, the evaluation experiment was confined to an offline setting, and a dataset of news and tweets collected related to the 2015 Nepal Earthquake. This dataset contained online news articles published by main news agencies such as the BBC, Associated Press, Reuters, Indian Express, and the Himalayan Times between the 25th of April 2015 and the 31st of June 2015. Tweets related to the same event were collected from CrisisNLP<sup>12</sup> [15]. The details of the dataset are shown in Table 7.11.

**Table 7.11** Details of the evaluation dataset before preprocessing

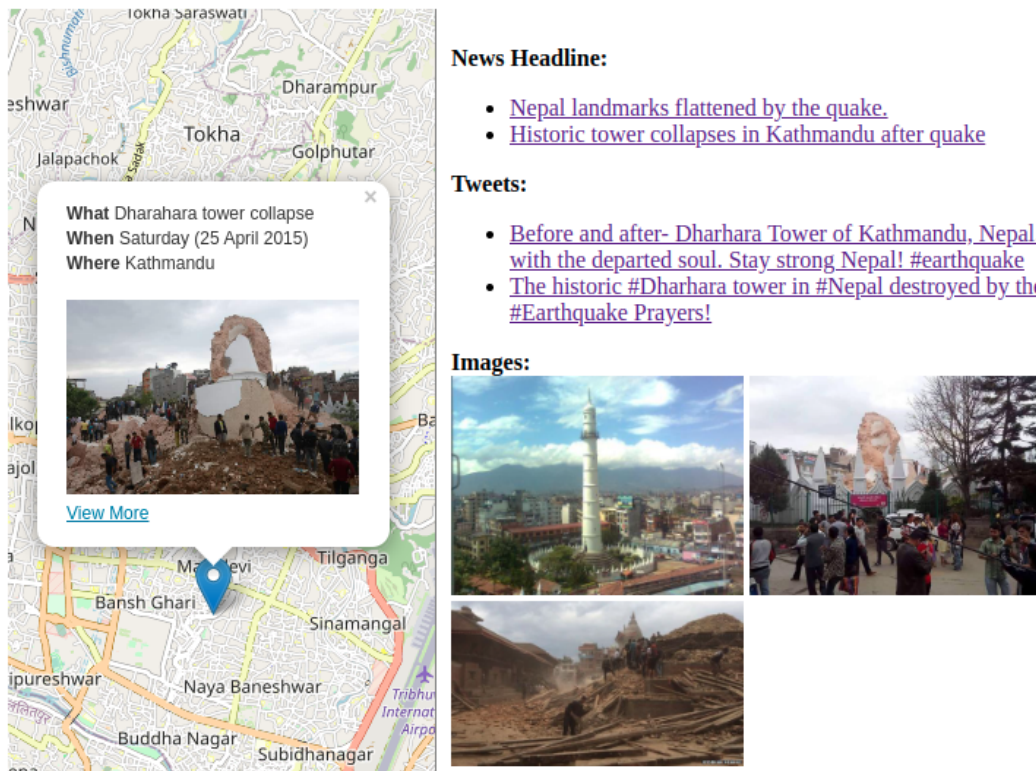
Datset type	Description	Count
Tweets	No of relevant tweets	4, 929
Online news	No of news headlines	229
	No of news body text lines	1, 374

Figures 7.7 and 7.8 present the word cloud representation of our news and tweets experimental dataset. A vast majority of high-weight words are similar in both representations (e.g., "nepal", "kathmandu", "earthquake", "quake", "help", "people"). Furthermore, Figure 7.9 presents the distribution of the length of the news headlines and tweets in terms of word counts.

As the histogram in Figure 7.9 shows, a large number of tweets and news headlines have 50-80 words. Interestingly, both news headlines and tweets have similar lengths and content based on these visualizations and can be validly integrated using text features.

DEES was operated to generate event templates using only tweet data and then using both tweet

<sup>12</sup><https://crisisnlp.qcri.org/>



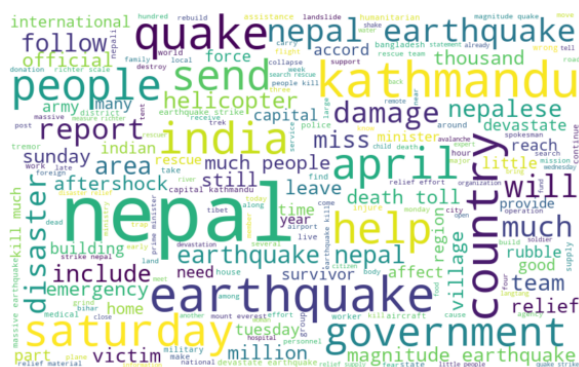
**Figure 7.6** Scores used for identifying best candidates among the candidate set extracted from news headlines, body, and tweets

and news data. Generalised precision ( $gP$ ), a score suitable for retrieval performance evaluations as described by Kekäläinen et al. [188] and used by hamberg et al. [125, 127] was used to evaluate the performance.  $gP$  is calculated using the following formula:

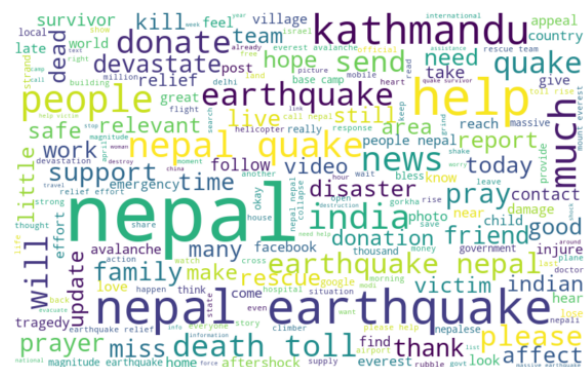
$$gP = \sum_{d \in R} r(d)/n \tag{7.3}$$

where  $R$  is the set  $n$  event templates from a database  $D = d_1, d_2, \dots, d_N$ . Let the event template  $d_i$  in the database have relevance scores of  $r(d_i)$  being real numbers ranging from 0.0 to 1.0.

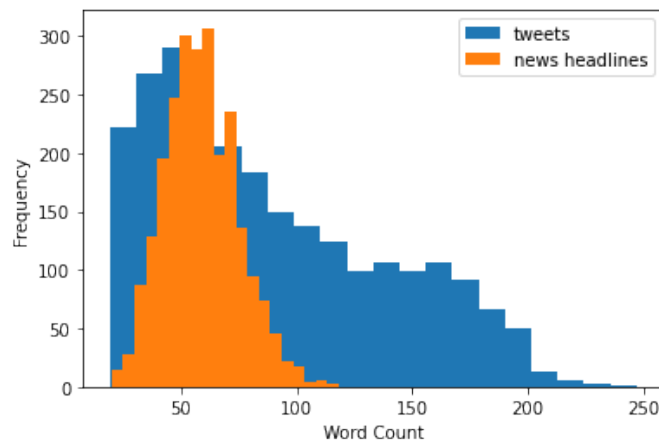
We generated event templates with DEES using only tweet text data and tweets + news text data. There were 56 event templates from news text data and 17 event templates from both news and text data. We recruited two graduate students as assessors. They were given the grouped



**Figure 7.7** Wordcloud representation of news headlines



**Figure 7.8** Wordcloud representation of tweets content



**Figure 7.9** The the distribution of length of news headlines and tweets, in terms of words

tweets and news that the system used to identify event templates. After reading each tweet or news article or in a group we provided the assessors with the 3W phrases that had been extracted by the system and asked them to judge the relevance of each answer on a 3-point scale as follows;

$$r(d_i) = \begin{cases} 0, & \text{if an answer contained not relevant information} \\ 0.5, & \text{if only part of the answer was relevant or if the information was missing} \\ 1, & \text{if the answer was completely relevant without missing information} \end{cases}$$

$$S_{cross}(C) = \begin{cases} 1, & \text{if candidates appear in news and tweets} \\ 0, & \text{Otherwise} \end{cases}$$

Table 7.12 shows the Average generalized precision ( $gP$ ) scores

**Table 7.12** Generalised precision scores of 3W

**Table 7.13** Tweets text data

	$gP$ Score
What	0.68
Where	0.74
When	0.85
Avg (3W)	0.76

**Table 7.14** News + tweets text data

	$gP$ Score
What	0.78
Where	0.92
When	0.86
Avg (3W)	0.85

## 7.5 Discussion and Conclusion

The noisy nature of tweets presents multiple challenges when used as a single source for disaster response. In contrast, online news sources provide officially validated reports regarding ongoing disaster events. Therefore, we developed DEES - a real-time event extraction system using online news and tweets in this article. Our system clusters tweets related to online news headlines, allowing responders to identify real-time updates from the crowd. The results of this investigation

show that event templates can be extracted more than 90% accurately, with more additional real-time information from tweets when using the proposed joint extraction system. An implication of this is the possibility that this system can be useful for improving the situational awareness of disaster responders during a developing disaster event. Even though it is hard to make direct comparisons, we chose Giveme5W1H [125] and the work by Norambuena et al. [278] to compare our *where*, *when* and *what* candidate extraction results. Both of these works predominantly focused on extracting 5W1H candidates only from large news corpora. Giveme5W1H used the BBC news (120 articles) while Norambuena et al. used AP news (1,529 articles). In comparison, we focused on clustered news and tweet text having around 30 text items maximum. The newly introduced location relatedness score achieved an improved accuracy score of (0.98) for *where* candidates comparatively to the accuracy scores of (0.78) in Giveme5W1H and (0.84) in Norambuena et al. Moreover, with the cross-media reference score, our *when* candidate extraction performed better than both Giveme5W1H (0.78) and Norambuena et al (0.77). The low accuracy of *what* candidate extraction (accuracy = 0.76) indicates that the differences in the structure of sentences highly affects the extraction algorithm. Further work is required to improve the *what* candidate extraction accuracy by improving the dependency parsing algorithm. Moreover, we plan to train a machine-learning algorithm to detect what candidates and compare how well it would detect the answers to *what* question over the proposed dependency parsing algorithm. For the first time, this study has demonstrated a real-time event extraction system for supporting disaster response while also providing details for all of the steps followed to support future implementation by other researchers.

Several limitations to this pilot study need to be acknowledged. First, the current study has only extracted *what*, *when*, and *where* questions. The practical application of this system could be further improved by incorporating the answers for *why*, *who*, and *how* questions. Second, we have limited the system to the English language and the geographic boundary to New Zealand. In future work, we envisage this research extended to integrate visual data from both news and tweets to provide more context while extracting candidates. Moreover, we plan to integrate other APIs available in New Zealand, such as Geonet, to extract live earthquake data, MetService to extract live weather data, and New Zealand Transportation Agency (NZTA) to extract live traffic data to improve the event extraction. A demonstration of the proposed extraction system is available at: <https://mu-clab.github.io/>. Our future work will develop this as a fully open-source system.

## Manuscript 6: Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison

The following article is published as: Nilani Algiriyage, Rangana Sampath, Raj Prasanna, Kristin Stock, Emma Hudson-Doyle, & David Johnston. (2021). Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison. In Anouck Adrot, Rob Grace, Kathleen Moore, & Christopher W. Zobel (Eds.), ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management (pp. 731–743). Blacksburg, VA (USA): Virginia Tech.

## 7.6 Abstract

Social media applications such as Twitter and Facebook are fast becoming a key instrument in gaining situational awareness (understanding the bigger picture of the situation) during disasters. This has provided multiple opportunities to gather relevant information in a timely manner to improve disaster response. In recent years, identifying crisis-related social media posts is analysed as an automatic task using machine learning (ML) or deep learning (DL) techniques. However, such supervised learning algorithms require labelled training data in the early hours of a crisis. Recently, multiple manually labelled disaster-related open-source twitter datasets have been released. In this work, we collected 192, 948 tweets by combining a number of such datasets, preprocessed, filtered and duplicate removed, which resulted in 117, 954 tweets. Then we evaluated the performance of multiple ML and DL algorithms in classifying disaster-related tweets in three settings, namely “in-disaster”, “out-disaster” and “cross-disaster”. Our results show that the Bidirectional LSTM model with Word2Vec embeddings performs well for the tweet classification task in all three settings. We also make available the preprocessing steps and trained weights for future research.

## 7.7 Introduction

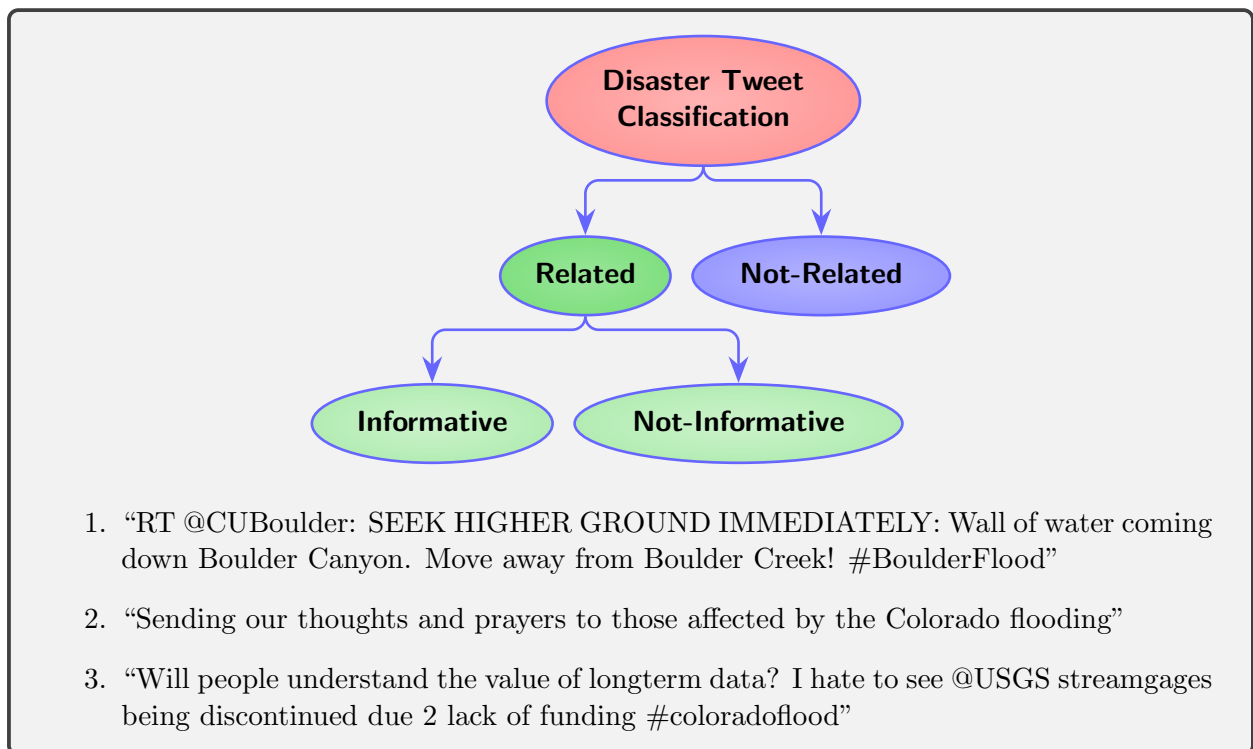
Social media (SM) platforms play an important role in providing a quick understanding of the situation as it unfolds during disasters. Research has found that the general public use SM applications during disasters to communicate information regarding urgent needs, infrastructure damage, injured or dead people, volunteering or donation efforts, and situational updates [201, 239, 30, 16]. Timely access to SM data can be leveraged for emergency response in the first few hours to significantly reduce both human loss and economic damage [16].

One of the main challenges for utilizing SM in crisis situations is the reliable detection of useful messages in a massive amount of streaming data. A straight forward method for collecting disaster-related tweets is to use disaster-keyword filtering. For example, tweets can be filtered using a dictionary with relevant keywords (e.g., “flood”, “earthquake”) or specific hashtags (e.g., “#NepalQuake”, “#boulderflood”, “#coloradoflood”). However, these descriptive keywords are diverse and ambiguous, and the hashtags chosen by individual users are often not consistent over time [401, 277]. As a result, a significant portion of the collected tweets can be irrelevant. Therefore, detecting disaster-related tweets is commonly modelled as an automatic classification task and tackled with Machine Learning (ML) algorithms and more recently with Deep Learning (DL) algorithms [401, 16, 267].

Tweet classification for disaster response is a text classification task that aims to determine if a tweet is related to a particular type of predefined informative class [30]. Olteanu et al. (2015) showed that crisis related tweets can be broadly categorised into: *related and informative*, *related but not informative*, and *not related*. For example see Figure 7.10 representing the Olteanu categorization using tweets extracted during 2013 Colorado floods.

A vast majority of the existing literature has focused on classifying tweets of the same event type and mostly used the CrisisLexT26 dataset for training classifiers [5, 110, 54, 191]. The CrisisLexT26 contains around 250K tweets posted during 26 crisis events in 2012 and 2013 [283]. Communication patterns of people might change over the years and, therefore, classification accuracy using classifiers trained on older datasets may not be high for future events [118]. Furthermore, supervised





**Figure 7.10** The Olteanu categorization for tweets, and three example tweets from the 2013 Colorado floods [283]. The first tweet is categorised as “Related and informative”, second as “Related - but not informative” and third as “Not related”

learning algorithms work well with more and complete training data covering the full spectrum of inputs that the model is supposed to handle during the classification task. Therefore, there is a timely need to test classifiers for new and more comprehensive datasets. To the best of the authors’ knowledge, to date, there exists no research for large scale ML and DL model evaluation in identifying disaster-related tweets combining multiple datasets. Therefore, during this research, we address the following research question.

- What ML or DL model has the best performance for a disaster-related tweet classification task?

To answer this question, we conduct experiments in the following three settings:

- In-disaster: training and test data belong to the same disaster type.
- Out-disaster: training and test data belong to different disaster types.
- Cross-disaster: training set consists of tweets of various disaster types.

During in-disaster experiments, we take both train and test data belonging to the same disaster type. The currently labelled datasets belong to a few disaster categories, including flood, earthquake, hurricane and biological. However, in reality, there are many more disasters (e.g., landslides, volcanic eruptions, droughts and tsunami) where people use tweets to communicate. Therefore, we wanted to explore how accurate a model can be if they are applied outside the domain. As a result, we explore a setting where train and test datasets belonging to different disaster types. We train

models for a combination of disaster types and test on individual types during the cross-disaster experiment. Altogether, we carry out 540 train, test experiments.

Our contributions can thus be summarized as follows:

1. Evaluation of a large-scale ML, DL model for disaster-related tweet classification
2. Evaluation of three state-of-the-art word embedding models
3. Publication of all the learning weights so that the response agencies can quickly adopt the trained models for an ongoing disaster<sup>13</sup>

The rest of this article is organized as the follows: The Related Work section reviews the literature related to disaster tweet classification. In the Methodology section we discuss the technical architecture and algorithms developed. The Results section provides results and critique of the findings. Finally, the Conclusion gives a brief summary and some directions for future research.

## 7.8 Related Work

In the crisis domain, useful information retrieval is an early step in processing data from SM platforms [43, 113, 424, 251]. A large and growing body of literature has investigated this as an automatic tweet classification problem [290, 371, 16, 110, 267]. These studies can be divided into three categories; related tweet classification, informative tweet classification and specific topical classifications. Related tweet classification focuses on identifying whether a tweet is related to a crisis event or not [118]. The concept of “Informativeness” is subjective, which heavily depends on the receiver of the information. However, generally “informative” tweets can be defined as tweets that provide valuable information to anyone in the scene of a disaster (e.g., a victim, supporter or responder). In comparison, “non-informative” tweets can be defined as tweets which do not convey any useful content in the scene of a disaster [267]. Research on specific topical classifications group tweets into multiple categories such as injured or dead people, sympathy and emotional support, affected people, caution and advice, missing people and donation needs [274]. A summary of the closely related work is presented in Table 7.15.

The vast majority of literature has considered classifying useful tweets using ML algorithms such as Naïve Bayes (NB) [290], Random Forest (RF) [186], Logistic Regression (LR) [275], Artificial Neural Networks (ANNs) [58] and Support Vector Machines (SVMs) [191]. More recent attention has focused on using deep neural networks such as Convolutional Neural Networks (CNN) [267, 277] and Long Short-Term Memory Networks (LSTM) [30] to address the disaster-related tweet classification task. Supervised ML or DL algorithms require labelled data to train classifiers that can be further used for classifying new data. Labelling the training data is typically carried out manually and is, therefore, a time-consuming and expensive process. This poses a major challenge when attempting to use supervised learning algorithms to assist disaster response in the event of a new disaster, as the time and effort needed to label tweets from the disaster prevent timely use of classifiers. However, recently multiple research work made manually labelled datasets such as CrisisNLP, CrisisLex and CrisisMMD freely available online [161, 16, 283, 284]. The review article by Kruspe et al. [197] summarises the details of such datasets. Furthermore, the Incident Streams of Text REtrieval Conference (TREC-IS) editions were designed to provide annotated datasets and

---

<sup>13</sup>Trained weights of the models, Disaster\_Tweet\_Classification



Table 7.15 The summary tweet classification studies.

Reference	Classification	Dataset	#Size	Algorithm	Features	Category	Best Accuracy
						In-disaster Out-disaster Cross-disaster	
[356]	Relatedness	Own data	7,490	SVM*, MaxEnt, NB	uni-grams, Word2Vec	✓	0.72
[54]	Relatedness	CrisisLexT26	28,000	NB, SVM,CNN*	CART, TF-IDF, Word2Vec	✓	0.83
[191]	Relatedness	CrisisLexT26	5,931	SVM	Semantic	✓	0.86
[401]	Relatedness	7 datasets; AIDR, CrisisLexT6, CrisisNLP, CrisisLexT26, CrisisNLP, CrisisMMD, Epic An- notations, collection by McMin et al.	123,166	feed-forward NN*,CNN	BERT, USE	✓	0.98
[275]	Relatedness	CrisisNLP3, CrisisLex, AIDR	21,021	SVM, LR, RF, CNN*	TF-IDF, Word2Vec and Cis embed- ings	✓	0.94
[370]	Relatedness	CrisisLexT26, Crowd- Flower10K	10,876	LR	TF-IDF, Word2Vec		
[239]	Informativeness	CrisisMMD	4,434	SVM,CNN, CNN and ANN*	n-grams	✓	0.76
[403]	Informativeness	CrisisLexT26, AIDR	6,780	RF, SVM, NB, LibLin- ear classifier*	n-grams		0.75
[290]	Informativeness	Own data	4,000	NB, SVM*	BOW	✓	0.87
[58]	Informativeness	CrisisLexT26 (flooding only)	5,577	ANN,SVM,CNN*	n-grams	✓	0.78
[5]	Informativeness	CrisisLexT26		Random Forest		✓	0.76
[277]	Informativeness	CrisisLexT26		CNN		✓	0.81
[274]	Topical	CrisisNLP		CNN, Bi-LSTM*		✓	0.62

\* The model having the best accuracy.

# Total number of tweets in the dataset.

bring together academia and industry to research automatically processing social media streams [373].

Word embedding is a key factor in improving the performance of a DL model for a text classification task. Multiple general-purpose word embeddings such as GloVe [294], fastText [50] and Word2Vec [255] and domain-specific word embeddings such as Crisis embedding [274] have been proposed. However, there is not much work done to examine the effectiveness of different deep learning architectures and different word embeddings in improving tweet classification models [30].

Many approaches for tweet classification focus on particular disaster types [5, 110, 58]. For example, Caragea et al. (2016), use flooding datasets extracted from CrisisLexT26 as for both training and testing datasets. Similarly, Gata et al. (2019) train SVM and NB models to detect tweets related to earthquake events. However, only a few studies comprehensively test classification across various disaster types [74, 118, 401]. Closer to our objective is the work by Graf et al. [118] and Wiegmann et al. [401]. Graf et al. (2018) introduce a cross-domain informativeness classifier based on SVM classifier. The study by Wiegmann et al. (2020) compares the effectiveness of three state-of-the-art machine learning models, namely CNN and two transformer models: BERT and Universal Sentence Encoder (USE) for the related tweet classification task. However, they explicitly consider only cross-disaster types. Also, these approaches have been mostly pursued in academic contexts and have not been made available to the public and responding organisations through easily accessible and integrable tools [54].

## 7.9 Methodology

We conduct experiments under three settings to evaluate twelve ML models and two DL models with three different word embeddings for the disaster-related tweet classification task.

### 7.9.1 Dataset

We extracted tweets from Disaster Data Corpus 2020 created by [401], that includes data from seven repositories, namely, CrisisLex T26 [283], CrisisLex T6 [284], CrisisNLP - RESOURCE # 1 [156], CrisisNLP - RESOURCE # 2 [161], CrisisNLP - RESOURCE # 5 [19], Epic Annotations [355], and the dataset collected by [249]. Furthermore, we downloaded additional non event-tagged Kaggle (“Real or Not? NLP with Disaster Tweets”) dataset<sup>14</sup> that was originally created by figure-eight<sup>15</sup>, and Appen Disaster Response Messages<sup>16</sup> and Kaggle (“Disasters on social media”) dataset<sup>17</sup>. Table 7.16 lists the 46 disasters contained in the datasets considered in this study and the number of Related and Not-Related tweets available for each of them before the preprocessing steps. We assigned each disaster to one of 8 disaster types, based on the work by Wiegmann et al. (2020). The combined dataset has 192, 948 labelled tweets in total.

During the noise reduction and preprocessing steps, we removed all non-English tweets, duplicate tweets and re-tweets (e.g., RT @username:) using string manipulations in Python Pandas

<sup>14</sup>Kaggle “Real or Not? NLP with Disaster Tweets” dataset, <https://www.kaggle.com/c/nlp-getting-started/overview>

<sup>15</sup>Appen Datasets Resource Center, <https://www.figure-eight.com/data-for-everyone/>

<sup>16</sup>Appen Disaster Response Messages, <https://appen.com/datasets/combined-disaster-response-data/>

<sup>17</sup>Kaggle “Disasters on social media” dataset, <https://www.kaggle.com/jannesklaas/disasters-on-social-media>

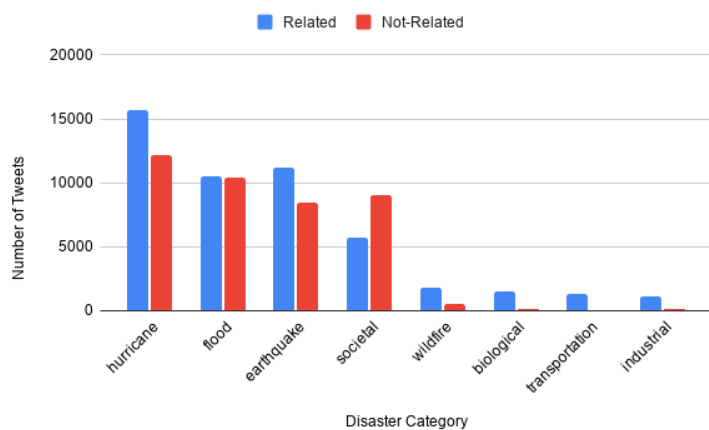
**Table 7.16** Related and Not-Related labelled tweets grouped by disaster category

Name	No of Tweets		Name	No of Tweets	
	Related	Not-Related		Related	Not-Related
<b>Flood (9)</b>			<b>Wildfires (3)</b>		
2012 Philipinnes	906	88	2012 Colorado	953	238
2013 Sardinia	926	68	2013 Australia	949	242
2013 Manila	921	47	2014 California	1,245	344
2013 Alberta	6,172	4,856	<b>Societal (2)</b>		
2013 Queensland	11,332	10,259	2013 Boston bombing	6,577	4,416
2013 Colorado	925	70	2013 LA airport shootings	912	87
2014 India	1,322	498	<b>Industrial (4)</b>		
2014 Pakistan	1,744	25	2012 Venezuela refinery explosion	339	58
2017 Sri Lanka	367	655	2013 West-Texas explosion	6,157	4,825
<b>Earthquake (12)</b>			2013 Brazil nightclub fire	952	40
2012 Costa Rica	909	399	2013 Savar building collapse	1,141	75
2012 Guatemala	940	108	<b>Transportation (4)</b>		
2012 Italy	940	50	2013 Glasgow helicopter crash	918	177
2013 Bohol	969	31	2013 New York train crash	999	0
2013 Pakistan	1,569	312	2013 Spain train crash	991	6
2013 California	1,595	106	2013 LA train crash	966	31
2013 Chile	1,590	342	<b>Hurricane (11)</b>		
2015 Nepal	10,583	5,801	2011 Joplin Tornado	1,756	976
2017 Mexico	1,030	350	2012 Hurricane Sandy	6,138	3,870
2017 Iraq and Iran	493	104	2012 Hurricane Pablo	907	68
2018 Nepal	3,410	2,820	2013 Typhoon Yolanda	940	71
<b>Biological (2)</b>			2013 Oklahoma Tornado	5,165	4,827
2014 Ebola	1,559	215	2014 Typhoon Hagupit	1,778	232
2014 Mers	1,331	27	2014 Hurricane Odile	1,219	43
<b>Other (3)</b>			2015 Cyclone Pam	1,508	496
2013 Russia meteor impact	1,133	271	2017 Hurricane Harvey	3,329	1,105
2013 Singapore haze	933	46	2017 Hurricane Maria	2,843	1,713
Kaggle and Appen datasets	24,800	14,725	2017 Hurricane Irma	3,548	956

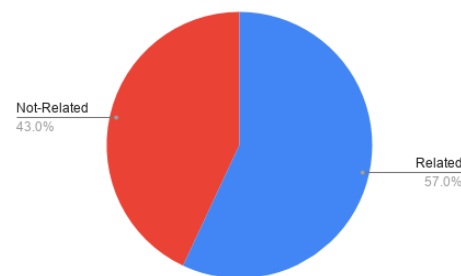
library<sup>18</sup>. Furthermore, all URLs, hashtags, special characters, emoticons, and emojis were removed. Also, we removed stop words, words having less than three characters and sentences having less than three words. After these steps, there were 117, 954 tweets, reducing around 40% of the tweets. We also applied lemmatization to convert words into their root forms as it improves the classification accuracy [5]. Figure 7.11 illustrates the number of Related and Not-Related tweets in each disaster category after the preprocessing stages.

We combined the Related and Informative and Related but not Informative into the *Related* class, and Not Applicable into the *Not-Related* class. For the datasets where there were topical classes, we combined them into *Related* class (e.g., “affected people”, “missing trapped or found people”). These two classes were then used for distinguishing crisis-related content from unrelated content for creating binary text classifiers [191]. Also, we combined the same disaster events across different datasets (e.g., Queensland Floods in CrisisLex and CrisisNLP). Figure 7.12 presents the

<sup>18</sup>Python pandas library: <https://pandas.pydata.org/>



**Figure 7.11** Related and Not-related tweets by category.



**Figure 7.12** Related and Not-related tweets in the dataset.

distribution of total Related and Not-related tweets in the dataset.

To avoid classification bias towards the majority class, we balanced the data from each category by matching the number of Related tweets with Not-Related ones. For example, after preprocessing, the number of related and not-related tweets of earthquake category were 6,946 and 4,650, respectively. We randomly selected not-related tweets from other categories except for earthquake category and made the dataset such as having 6,946 related and 6,946 not-related tweets.

## 7.9.2 Models

We selected five supervised ML algorithms that have been mostly explored for disaster tweet classification tasks namely Logistic Regression (LR), Decision Tree (DT), SVM, NB, ANN and RF [290, 186, 191]. In addition, six more ML models were selected that have rarely been studied for disaster tweet classification tasks in literature such as Gradient Boosting Classifier (GB), Ridge-Classifier, AdaBoost, k-Nearest Neighbors (KNN), xgboost, and catboost. All the algorithms were implemented in Python scikit-learn [292], using the default parameters. Furthermore, we used two DL algorithms, namely CNN and Bi-directional LSTM (Bi-LSTM).

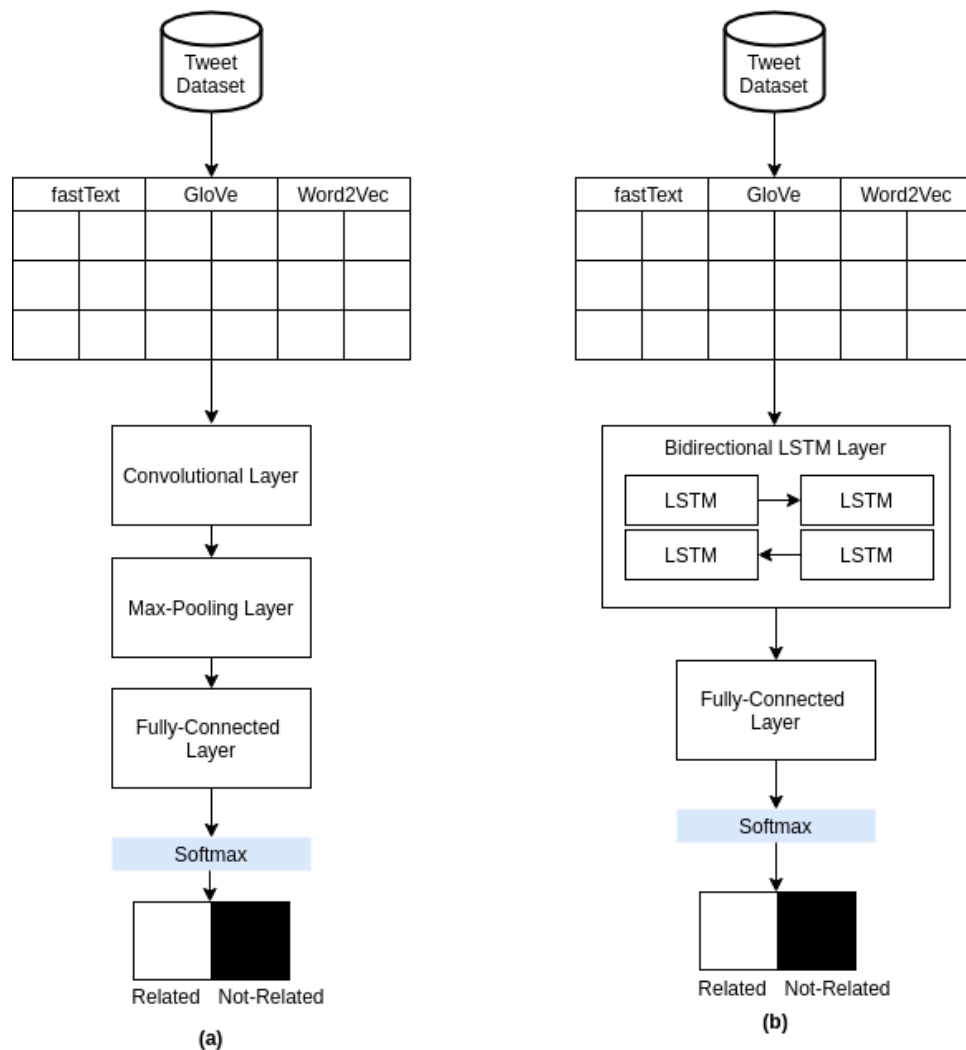
Text documents have to be converted into numerical vectors for the machine learning task. Single-dimensional bag-of-words (BOW) model with Term Frequency–Inverse Document Frequency (TF-IDF) representations have been widely adopted for traditional ML algorithms [275], whereas word embeddings such as Word2Vec and GloVe have been used for DL models [30, 54]. We extracted word-level unigrams from tweets as features for our ML models and converted to TF-IDF vectors by considering each tweet as a document.

Word embeddings generate a vectorized representation of words by mapping words to vectors instead of a one-dimensional space. Therefore, semantically close words should have a similar vector representation instead of a distinct representation. We used pre-trained Word2Vec model of Google News dataset about 100 billion words<sup>19</sup>, pre-trained fastText model of Wikipedia 2017, UMBC web base corpus having 999,995 word vectors<sup>20</sup> and pre-trained GloVe embeddings having 2 196,016 vectors<sup>21</sup> as features for our DL models. When embedding, each tweet is represented as a matrix of size  $n * k$ , where  $n$  is the maximum length of a tweet text (number of words) in the

<sup>19</sup>word2vec pre-trained word vectors, <https://code.google.com/archive/p/word2vec/>

<sup>20</sup>fastText pre-trained word vectors, <https://fasttext.cc/docs/en/english-vectors.html>

<sup>21</sup>GloVe pre-trained word vectors, <https://nlp.stanford.edu/projects/glove/>



**Figure 7.13** Illustration of (a) CNN and (b) Bidirectional LSTM of twitter text classification task.

training data and  $k$  is the embedding vector dimension. We used ( $k = 300$ ) for all three embedding models and applied zero-sequence padding for the tweet texts having the number of words less than  $n$ . Kim et al. (2014) described a CNN architecture for text classification tasks and has been mostly adopted in disaster tweet classification studies [193, 267, 54]. We adopted a similar model with a single convolution layer followed by a max-over-time pooling layer and a fully connected layer where the softmax function is applied to predict the document classes. Furthermore, our dropout rate was set to 0.5 for regularisation and ran the model for 100 epochs. However, early stopping was used to terminate the execution based on validation accuracy. The Long-Short Term Memory (LSTM) is a specialized version of Recurrent Neural Network (RNN) capable of learning long term dependencies. While LSTM can only see and learn from past input data, Bidirectional LSTM (Bi-LSTM) runs input in both forward and backward directions. This bidirectional feature of Bi-LSTM is critical for the various applications involved with understanding complex language [30]. ALRashdi et al. (2019) described a Bi-LSTM model for disaster tweet classification, and we adopted a similar architecture for our experiments. Figure 7.13 illustrates the architectures of the CNN and Bi-LSTM networks for the tweet classification task.

### 7.9.3 Experiments

We carry out experiments under the following three settings.

1. In-disaster balanced training dataset (disasters considered: Earthquake, Flood, Hurricane and Societal)
2. Out-disaster balanced training dataset (disasters considered: Earthquake, Flood, Hurricane and Societal)
3. Cross-disaster balanced training dataset (disasters considered: Earthquake, Flood, Hurricane, Societal, Wildfire, Industrial, Transportation and Biological)

**Table 7.17** In-disaster, out-disaster and cross-disaster experimental datasets

<b>In-Disaster</b>		<b>Cross-Disaster</b>	
Train Dataset	Test Dataset	Train Dataset	Test Dataset
Earthquake	Earthquake (2018 Nepal)	All data	Earthquake (2017 Iraq and Iran)
Flood	Flood (2017 Sri Lanka)	All data	Flood (2018 Nepal)
Hurricane	Hurricane (2017 Maria)	All data	Hurricane (2017 Maria)
Societal	Societal (2013 LA airport shootings)	All data	Societal (2013 LA airport shootings)
		All data	Biological (2014 MERS)
		All data	Transportation (2013 LA train crash)
		All data	Wildfire (2014 California)
		All data	Industrial (2013 Brazil nightclubfire)
<b>Out-Disaster</b>		<b>Out-Disaster</b>	
Train Dataset	Test Dataset	Train Dataset	Test Dataset
Earthquake	Flood (2017 Sri Lanka)	Hurricane	Earthquake (2018 Nepal)
Earthquake	Hurricane (2017 Maria)	Hurricane	Flood (2017 Sri Lanka)
Earthquake	Societal (2013 LA airport shootings)	Hurricane	Societal (2013 LA shootings)
Flood	Earthquake (2018 Nepal)	Societal	Earthquake (2018 Nepal)
Flood	Hurricane (2017 Maria)	Societal	Flood (2017 Sri Lanka)
Flood	Societal (2013 LA shootings)	Societal	Hurricane (2017 Maria)

We formulated our experiments for all three settings such that the tests are applied to the newest disaster dataset. For example, we selected the most recent disasters from each category and used that as the test dataset. In the case of multiple disasters in the same year, we chose the disaster with the fewest tweets as the test dataset. Therefore, our test datasets were; 2017 Sri Lanka floods, 2018 Nepal earthquake, 2014 MERS, 2014 California wildfires, 2013 LA airport shootings, 2013 Brazil nightclub fire, 2013 LA train crash and 2017 hurricane Maria. Hence, before training the algorithms, we removed those entire test datasets from each category to test

the models for unseen data. Table 7.17 lists the training and testing datasets considered for three experiments. We used 10 fold stratified sampling for cross-validation <sup>22</sup> while training the models. The performance of algorithms were measured using average F1-score. Altogether we carried out 36 model training and evaluations for in-disaster category. To reduce the number of training and evaluations, we selected the top three ML models and best DL model based on average F1-score for the out-disaster and cross-disaster experiments. All experiments were executed in the Google Collaboratory <sup>23</sup> environment.

## 7.10 Results and Discussion

**Table 7.18** Average F1-scores of the DL and ML models for the in-disaster experiments (The best scores are highlighted in grey, and the three best performing ML models and the best performing DL model are underlined).

Algorithm	Hurricane	Societal	Earthquake	Flood
<u>Linear SVM</u>	0.742	0.565	0.810	0.866
RidgeClassifier	0.718	0.571	0.813	0.800
<u>Logistic Regression</u>	0.743	0.576	0.799	0.819
Decision Tree	0.695	0.613	0.769	0.741
k-Nearest Neighbors	0.492	0.512	0.523	0.516
Gradient Boosting Classifier	0.687	0.491	0.676	0.727
<u>NB</u>	0.729	0.737	0.795	0.854
AdaBoost	0.715	0.605	0.731	0.792
Random Forest	0.524	0.687	0.740	0.789
Perceptron	0.628	0.632	0.751	0.753
xgboost	0.716	0.613	0.756	0.767
catboost	0.696	0.537	0.723	0.734
LSTM-fastText	0.770	0.690	0.794	0.914
CNN-fastText	0.791	0.779	0.769	0.905
LSTM-GloVe	0.776	0.787	0.753	0.911
CNN-GloVe	0.766	0.707	0.799	0.898
<u>LSTM-Word2Vec</u>	0.793	0.795	0.820	0.925
DL-Word2Vec	0.787	0.764	0.804	0.900

Figure 7.18 shows the F1-score for ML and DL algorithms for in-disaster experiments, where scores for the models range from 0.49 to 0.92. The Bi-LSTM model with Word2Vec features performs the best while the KNN algorithm produces the worst results. From the data in Figure 7.18, it is apparent that the DL algorithms outperform ML algorithms having F1-scores over 0.69. It can also be seen from the data that across all the experiments flood tweet dataset has achieved higher F1-values. A possible explanation for this can be the larger number of common words among the flood datasets considered.

Regarding the out-disaster experiments, any DL or ML model trained on hurricane dataset and tested on flood dataset has performed the best while the models trained on societal and applied on hurricane has performed the worst (see Figure 7.19). Overall, ML/DL model's performance applied for an out domain has obtained lower average F1-values, with scores ranging from 0.42 to 0.91. This

<sup>22</sup>Stratified ShuffleSplit cross-validator, [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.StratifiedShuffleSplit.html](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html)

<sup>23</sup>Google Collaboratory, <https://colab.research.google.com/>

**Table 7.19** Average F1-scores of the ML and DL models (LSTM-fastText (DL<sup>1</sup>), CNN-fastText (DL<sup>2</sup>), LSTM-GloVe (DL<sup>3</sup>), CNN-GloVe (DL<sup>4</sup>), LSTM-Word2Vec (DL<sup>5</sup>) and DL-Word2Vec (DL<sup>6</sup>)) for the Out-disaster experiments. The best scores are highlighted in grey.

Algorithm	Earthquake-Flood	Earthquake-Hurricane	Earthquake-Societal	Flood-Earthquake	Flood-Hurricane	Flood-Societal	Hurricane-Earthquake	Hurricane-Flood	Hurricane-Societal	Societal-Earthquake	Societal-Flood	Societal-Hurricane
SVM	0.632	0.531	0.492	0.597	0.530	0.504	0.653	0.827	0.510	0.522	0.519	0.507
LR	0.644	0.532	0.525	0.639	0.539	0.504	0.713	0.831	0.521	0.524	0.518	0.504
NB	0.759	0.522	0.551	0.844	0.641	0.581	0.822	0.841	0.630	0.729	0.789	0.616
DL <sup>1</sup>	0.844	0.593	0.600	0.833	0.612	0.658	0.732	0.896	0.610	0.677	0.762	0.427
DL <sup>2</sup>	0.782	0.495	0.602	0.784	0.680	0.622	0.793	0.903	0.566	0.743	0.801	0.469
DL <sup>3</sup>	0.839	0.610	0.624	0.682	0.624	0.665	0.808	0.911	0.621	0.783	0.730	0.567
DL <sup>4</sup>	0.837	0.606	0.678	0.742	0.649	0.634	0.783	0.886	0.605	0.735	0.810	0.525
DL <sup>5</sup>	0.864	0.653	0.657	0.856	0.739	0.677	0.826	0.907	0.644	0.795	0.819	0.634
DL <sup>6</sup>	0.794	0.583	0.626	0.733	0.740	0.616	0.776	0.886	0.618	0.660	0.730	0.608

finding implies that out-disaster experiments need to be carefully designed. Furthermore, choosing DL models over ML models yields better results. Among the DL models, the Bi-LSTM model with Word2Vec embeddings has performed the best.

Figure 7.20 illustrates the results of cross-disaster experiments, where average F1-scores ranging from 0.41-0.93. The Bi-LSTM model with Word2Vec features has achieved the highest F1-scores while KNN algorithm performing the worst. Overall, a model trained on a combined disaster dataset applied on flood data performs the best while industrial and societal categories perform poorly.

The current state-of-art for relatedness classification can be found in [356, 54, 191, 401, 275, 370]. The accuracy scores reported by them are as 0.72 for in-disaster experiments in [356], 0.83 for cross-disaster experiments in [54], 0.98 for cross-disaster experiments in [401] 0.86 for cross-disaster experiments in [191] and 0.94 for in-disaster experiments in [275]. It is important to note that

**Table 7.20** Average F1-scores of the ML and DL models (LSTM-fastText (DL<sup>1</sup>), CNN-fastText (DL<sup>2</sup>), LSTM-GloVe (DL<sup>3</sup>), CNN-GloVe (DL<sup>4</sup>), LSTM-Word2Vec (DL<sup>5</sup>) and DL-Word2Vec (DL<sup>6</sup>)) for the Out-disaster experiments. The best scores are highlighted in grey.

Algorithm	Wildfire	Hurricane	Industrial	Societal	Transport	Biological	Earthquake	Flood
SVM	0.625	0.727	0.535	0.590	0.744	0.797	0.712	0.891
LR	0.616	0.729	0.543	0.642	0.718	0.781	0.638	0.894
NB	0.606	0.612	0.607	0.606	0.674	0.836	0.668	0.898
DL <sup>1</sup>	0.793	0.741	0.412	0.513	0.752	0.628	0.729	0.919
DL <sup>2</sup>	0.781	0.712	0.563	0.518	0.771	0.720	0.779	0.911
DL <sup>3</sup>	0.795	0.743	0.553	0.499	0.771	0.729	0.814	0.902
DL <sup>4</sup>	0.763	0.689	0.593	0.559	0.711	0.583	0.708	0.916
DL <sup>5</sup>	0.798	0.766	0.608	0.644	0.775	0.804	0.816	0.928
DL <sup>6</sup>	0.770	0.716	0.588	0.622	0.770	0.794	0.805	0.896



these experimental settings are significantly different from ours. For example, in [275] in-disaster experiments contained tweets only from Cyclone PAM and in [401] cross-disaster experiments contained data from same disaster category.

In summary, from these results, it is interesting to note that DL models outperform the traditional ML algorithms. This finding supports previous research by [275] and [54] who showed that DL classifiers performed better than all non-DL classifiers. It seems possible that word embedding performs well than the BOW and TF-IDF representations. However, the training time for DL algorithms were higher than the classical ML models. Moreover, the difference in classification time for both approaches was negligible. Among the DL models, the Bi-LSTM model with Word2Vec features has performed the best across all three experimental settings. Another important finding is that with the default parameters, the KNN algorithm has performed the worst for all three experiments. We have considered data-rich disasters (having more than 25,000 tweets in the training dataset) for in-disaster and out-disaster categories while cross-disaster category having a combination. From the data in Figures 7.18 and 7.20 it is visible that there is no significant deviation among the results of in-disaster and cross-disaster results. An implication of this is the possibility of using a cross-disaster dataset if the training data unavailable. However, the F1-scores of the out-disaster category are generally lower except for DL models. Therefore, out-disaster experiments have to be carefully designed. The generalisability of these results is subject to certain limitations. For instance, we kept the default parameters for all our ML algorithms. As indicated in the literature, parameter tuning yields better results [83]. Therefore, future research has to be done to identify the best parameters for the ML models. Furthermore, our training datasets are of different sizes as we wanted to explore the maximum possible performance of the individual classifier.

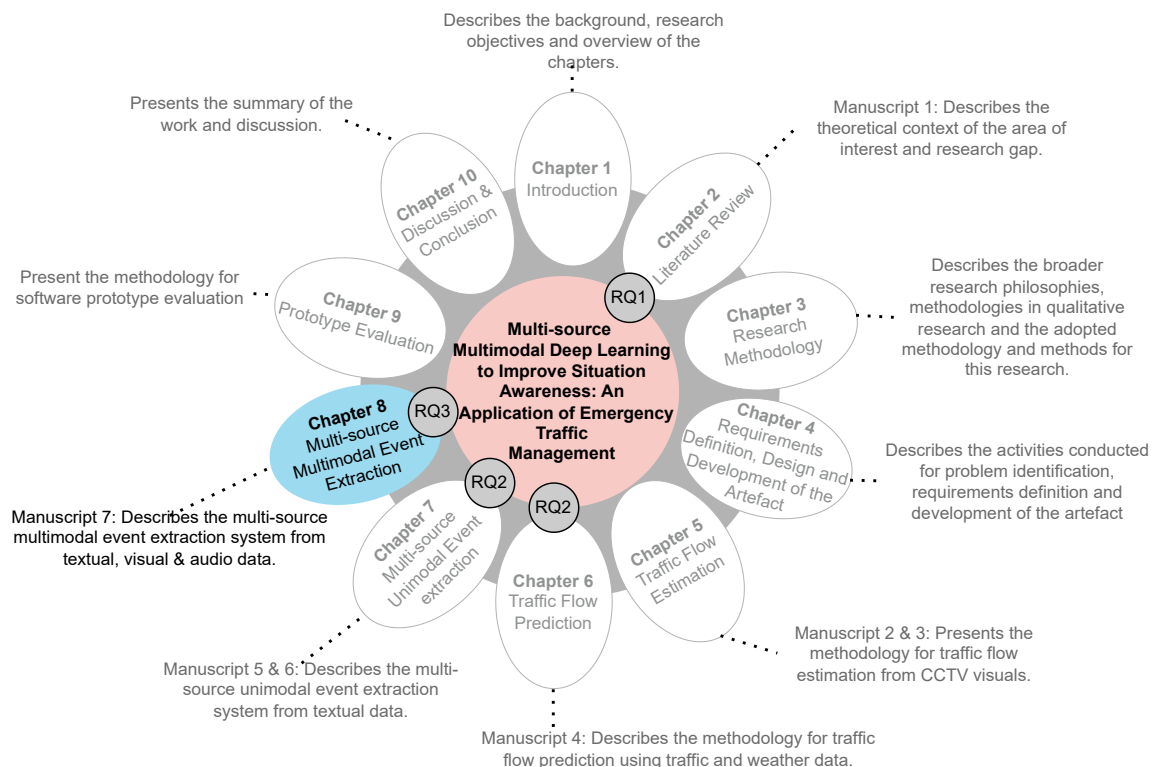
## 7.11 Conclusion

This study investigated the identification of the best-performing ML or DL models, to classify disaster-related tweets in three settings: in-disaster, out-disaster and cross-disaster. The research findings suggest that generally, DL models outperform traditional ML models for the tweet classification task. The use of different embedding plays a significant role in text classification. It was shown that the Bi-LSTM model with Word2Vec features performing the best for all three experimental settings considered, namely, in-disaster, out-disaster and cross-disaster.

This is the largest study so far, evaluating  $\approx 0.2$  million labelled tweet dataset. The evidence from this study suggests that classifiers can be trained to identify disaster-related tweets in all three categories. However, these findings are limited by using default parameters for the ML algorithms and considering only English tweets. Therefore, in our future work, we plan to identify the best parameters for the experimented models. Furthermore, we will study the capacity of studied models for reproducing or replicating for future researchers.

## Chapter 8

# Real-time Information Extraction from Multi-source Multimodal Data: An Application of Emergency Traffic Management



This chapter contains the seventh manuscript on the design, development, and evaluation of the fourth and final component of the software artefact. The third research question, “How can the integration of multi-source multimodal data effectively support disaster response by cross-validating social media data?” is addressed in this chapter. The article explains how multimodal data from multiple sources can be utilized successfully to validate SM content and support the SA of disaster responders through real-time event templates.

## Abstract

Traditionally, actionable information for disaster response has been gathered via analyzing reports made by officials on disaster sites. However, these approaches are labor-intensive, time-consuming, and susceptible to human errors. Today, people are increasingly using social media platforms to report situational information such as urgent needs, dead or injured people, and damage to properties during disasters. Despite their usefulness, the majority of this pertinent data is not available to humanitarian organizations during emergencies, mainly due to several data processing and data quality issues. Therefore, we developed a Multi-Source Multimodal Event Extraction System for Disaster Response (M<sub>2</sub>E<sub>2</sub>S for DR), this is a real-time system that focuses on integrating text and images to extract answers to the *What (semantic)*, *Where (spatial)* and *When (temporal)* (3W) questions, as well as impact information. The data for the system is extracted from online news and social media channels, and thus, the integrity of the user-generated content is cross-validated. M<sub>2</sub>E<sub>2</sub>S was evaluated using news and tweets in the traffic and transportation domain, and it achieved an overall generalized precision of 0.87 for 3W+Impact questions. We have made publicly available a demonstration version of the M<sub>2</sub>E<sub>2</sub>S including all main functionalities for future researchers at <https://rangikanilani.github.io/events.html>.

## 8.1 Introduction

Whether large or small, responding to a crisis requires a massive amount of data to assess the situation, analyze the impact, and make clear decisions to support the community [162, 409]. Traditionally, disaster data acquisition was performed by sending officials to disaster locations and reviewing the reports they sent. As a result, the reporter was often exposed to a significant life threat. Moreover, these methods are labor intensive, time-consuming, prone to human errors, and can be biased [145]. Today, the majority of people can produce and share information as it happens, thanks to Social Media (SM) and other associated applications available on mobile devices. Research studies have shown that the general public use SM platforms during disasters to report critical information such as early warnings, cautions, and damage to infrastructure such as roads, bridges, and buildings [159, 237, 85, 242]. Furthermore, SM-based communication has challenged traditional media systems, information pathways, and hierarchies, with the most recent real-world events being first reported in SM [301, 181].

Despite the benefits of high volumes of SM data, it brings multiple challenges for real-time processing of the information to be used for emergency response. Most importantly, responding organizations have concerns about the trustworthiness of information available in SM channels [165]. For example, user-generated SM content can not be directly incorporated for understanding the situation as it may contain fake news, rumors, and misinformation [196]. As a result, a significant amount of timely, valuable information on SM platforms gets wasted without being properly utilized. However, in addition to SM platforms, multiple sensors and other sources such as satellite monitoring, online news, and crowd-sourcing generate a massive amount of data during a disaster event [411, 13]. Generally, the data from various sources come in different modalities such as text, images, audio, and video. Integrating multiple other data sources and modalities provides mechanisms to triangulate and validate the information while also adding more contextual information [281, 166]. Hence, the information acquired through SM channels can be further strengthened.

Moreover, cross-validation provides ways to reduce errors and improve overall accuracy.

A considerable amount of literature is available on analyzing the ways of using data from different sources and modalities. However, a vast majority of them have considered a single source or a single modality. For example, Aipe et al. [10] explored how text data from SM can be used for disaster-related information classification. Kumar et al. [200] used Twitter text data for location identification during emergencies. An image dataset extracted from remote sensing was used by Chen et al. [67] for damage assessment. Similarly, image datasets extracted from SM posts have been used for disaster-related information filtering [14]. One of the most significant drawbacks of using a single source of data or modality is that it leads to information gaps in understanding the comprehensive view of a disaster situation. Comparatively, data from different sources and formats bring complementary information regarding an event that leads to more robust inferences [166].

Figure 8.1 shows several online news headlines, associated images, and a tweet extracted during a vehicle crash incident that occurred in June 2021 in Auckland, New Zealand. Although the crash is reported in the headlines, the nature and extent of the damage cannot be inferred from the text alone. However, if we consider the images attached, it is easy to understand the destruction caused by the accident. Moreover, the tweet provides eyewitness information and instructions for the public to avoid closer roads.



**Figure 8.1** Online news, images and tweet text with complementary information

Traditionally, disaster-related information extraction from SM and verification have relied heavily on manual methods, which need high levels of human intervention [107]. However, recently, researchers have explored Machine Learning (ML) models to automate tasks such as filtering useful data, and classification [290, 191, 273]. More recent attention has focused on Deep Learning (DL) techniques and has provided improved results over classical machine learning approaches for multiple disaster data processing tasks [268, 203, 242]. This performance is due to the capability of DL models to perform automatic feature engineering compared to the handcrafted feature engineering required for the classical machine learning approaches [211]. In this paper, we describe the software application developed using DL-based architectures to fuse the text and visual data extracted from online news and SM named as Multi-Source Multimodal Event Extraction System for Disaster Response ( $M_2E_2S$  for DR). The outputs (real-time event templates) generated by  $M_2E_2S$  can be used by emergency responders for their decision-making tasks. The architecture of  $M_2E_2S$  addresses the following research questions.

1. How can multimodal data from multiple sources be extracted in real-time?
  - This question explores the ways of extracting live data in many formats such as text and visuals from multiple sources.
2. How can data be pre-processed and grouped?
  - In this question, the ways to reduce noise from raw data and preprocessing steps are discussed.
3. How can semantic, spatial, and temporal descriptive features can be identified?
  - This question identifies methods to extract answers for semantic- (*what*), spatial- (*where*) and temporal- (*when*) questions from text data.
4. How can impact information be identified?
  - In this question, the ways of identifying impact information such as road deaths, injuries, and damages to infrastructure are explored.
5. How can event templates with geolocation be provided for responding organizations?
  - This question identifies ways of presenting the event templates with geolocation for the emergency responders.

The rest of the paper is organized as follows. In Section 8.2, we provide a comprehensive review of the literature. Then, in Section 8.3, the details of the methodology for developing the proposed system are discussed. Next, in section 7.4 experimental scenarios and results are discussed. Finally, Section 8.5 presents the concluding remarks.

## 8.2 Related Work

Recently social media applications like Twitter and Facebook have become more popular as an important source of data during disasters. Therefore, the extraction and analysis of disaster-related data from SM content have received much attention in the literature. For example, the text data from SM posts have been extensively analyzed for disaster-related information classification [10, 16, 56, 180], damage assessment [244, 311], and event detection tasks [276]. Image data extracted from SM posts have also been used for disaster-related information classification [199, 327], damage assessment [219, 218, 272], and event detection [34]. However, the user-generated content available in SM channels contains fake information, rumors, and misinformation [62]. Therefore, responding organizations have reported concerns regarding the trustworthiness of SM content [165].

In addition to SM channels, there are other sources that provide disaster-related data, such as remote sensing, online news, CCTV monitoring, and crowdsourcing [411]. Remote sensing, mainly satellite monitoring, and drones have been major sources of visual data for disaster response activities. These data have been analysed for damage assessment [37, 67, 94, 112, 144, 177, 222, 257, 270, 341, 357, 381, 421], disaster mapping [8, 265, 316], rescue and resource allocation [46], and event detection [32, 189, 228, 229]. Data from CCTV cameras have been used to detect disasters such as early fires, and flooding [232, 263]. Crowdsourced data has been largely used for location

reference identification during disaster events [184]. However, the vast majority of this research has only considered a single source or a single data modality.

Combining multiple sources and modalities leads to more information than learning from a single data modality alone [166, 25]. The semantic association of multi-source multimodal data enables users to reap the benefits of a more complete set of knowledge [368]. Furthermore, data integration can help to reduce errors and improve the integrity of the information provided [25]. As a result, the trustworthiness of the content provided through SM channels can be further improved.

More recently, researchers have analyzed multimodal datasets and demonstrated their usefulness in many disaster response tasks. For example, Mouzannar et al. [261] proposed a deep learning multimodal classification of disaster-related SM posts. Convolutional Neural Networks (CNNs) were used to process raw images and text before classifying social media posts into one of the six classes using softmax layers. Researchers have recently proposed multimodal systems utilizing both tweet text and images to find relevant information from SM. Rizk et al. [326] developed a multimodal disaster-related classifier to classify Twitter data into the built-in infrastructure damage and nature damage classes. They concatenated semantic features from tweet text and visual features from the image and achieved an accuracy of 92.43%, whereas a model that uses only visual features achieved an accuracy of 91.10%. Mouzannar et al. [261] developed a multimodal system based on the deep learning framework to classify user posts into Fire, Floods, Natural landscape damage, Infrastructural damage, Injuries and dead people, and Non-damage classes. They used a CNN-based Inception model for images and a CNN model for text and combined textual and visual features to classify users' posts and achieved an accuracy of 92.62%. Apart from that, multimodal deep learning-based approaches have been used in many other disaster response tasks such as disaster damage assessment [7, 132, 261], disaster event detection [20], and disaster-related information filtering and classification [1, 55, 109, 152, 153, 202, 217, 241, 238, 259, 306, 307, 326, 281]. A summary of these multimodal deep learning applications in disaster management is presented in Table 8.1.

The accuracy of the DL approaches shown in Table 8.1 is higher than the accuracy of DL algorithms employed to learn from a single modality alone. However, these methods have been mostly explored using static datasets available offline [1, 202, 7]. Considering the time-critical nature of disaster environments, one of the main concerns is that the responding organizations need real-time information for their decision-making tasks. Work that has addressed real-time multimodal data extraction and fusion includes the Advanced System for Emergency Management (ASyEM), which was proposed by Foresti et al. [107] to integrate data from physical sensors, social media applications, and Unmanned Aerial Vehicles (UAV)s on demand. The goal was to enhance SA during disasters by generating alarms when an event is detected. The paper describes the system's initial architecture but does not detail how data is integrated from all sensors. Moreover, ASyEM largely depends on intelligent sensors that are capable of analyzing data obtained, learning normal patterns, and making local decisions about anomalous events. However, these sensors are not practical to implement in every location. Quelloffene Integrierte Multimedia Analyse (QuOIMA) is another research project that aims to combine traditional and social media contents for disaster situation awareness [329]. Their focus is on the gathering and analysis of (incoming) information rather than the management of active (outgoing) communication [38, 329]. The project's early work describes the conceptual architecture of the proposed system, but the practical application is not described. Therefore, both these systems have three main limitations: (1) They do not discuss

how different data sources and modalities are fused. (2) They do not evaluate how successful the fusion algorithms are (3) They do not provide examples of use of the proposed architectures for real-time applications.

**Table 8.1** Applications using multimodal data for disaster response

Article	DR Task	Data Source	DL Algorithm	DL Architecture
[1]	Disaster Related Information Classification	CrisisMMD <sup>1</sup>	CNN	DenseNet
[7]	Disaster Damage Assessment	CrisisMMD	CNN, Recurrent CNN	Inception-v3 pretrained on Imagenet
[20]	Disaster Event Detection	Media-Eval 2020 <sup>2</sup>	CNN, BERT	BERT, VggNet16 pre-trained on the Places dataset and VGG-16 pretrained on Imagenet, ResNet152
[55]	Disaster Related Information Filtering	CrisisLexT26	CNN	
[109]	Disaster Related Information Filtering	CrisisMMD	CNN	VGG-19, BERT
[132]	Disaster Damage Assessment	Twitter API, Flickr API	CNN	ResNet18 CNN pre-trained on the Places365 dataset
[152]	Disaster Related Information Filtering	Web mining	CNN	
[153]	Disaster Related Information Filtering	Web mining	CNN	Word2Vec, Inception-V3
[202]	Disaster Related Information Classification	CrisisMMD	CNN	VGG-16
[217]	Disaster Related Information Filtering	CrisisMMD	CNN, RNN	AlexNet
[241]	Disaster Related Information Filtering	CrisisMMD	CNN	VGG-16
[238]	Disaster Related Information Classification	CrisisMMD	CNN, Bi-LSTM	VGG-16, BERT, crisis word embeddings, DenseNet
[259]	Disaster Related Information Filtering	Twitter API	CNN	VGGNet, ResNet, Inception-v3
[261]	Disaster Damage Assessment	CrisisLexT26, CrisisNLP, web mining	CNN	Inception-v3, Inception-v4, VGG16, InceptionResnet-v2, w2v, Glove
[306]	Disaster Related Information Classification	web mining	CNN, LSTM	Inception-v3, SoundNet
[307]	Disaster Related Information Classification	CrisisMMD	CNN, BERT, RoBERTa	AlexNet VGG19 ResNet-50, ALBERT-base BERT-base RoBERTa-base

<sup>1</sup>CrisisMMD, <https://crisisnlp.qcri.org/crisismmd>

<sup>2</sup>Media-Eval 2020, <https://multimediaeval.github.io/editions/2020/>

[326]	Disaster Related Information Classification	SUN database <sup>3</sup>	CNN	Inceptionv3, W2V
[281]	Disaster Related Information Classification	CrisisMMD	CNN	VGG-16

Structured event extraction from text data has been mostly explored using news data corpora. This includes the extraction of answers for 5W1H questions (who, what, when, where, why, and how), which can be used to describe the main event. Hamborg et al. [126] and Norambuena et al. [278] describe syntactic rules to extract answers for 5W1H questions using news datasets published in English. For instance, they used named entities, noun phrases (NP), verb phrases (VP), NP-VP patterns for the extraction of words and phrases. Finally, candidate scoring is used to find the best answers. A more recent survey by Gupta et al. [121] provides a comprehensive summary of disaster event detection research that utilized text data from SM. So far, however, there has been little discussion about extracting structured events by combining multiple sources and data formats. Considering these limitations in the literature, we developed M<sub>2</sub>E<sub>2</sub>S to fuse multi-source multimodal data in real-time for DR tasks, and the details of data acquisition, preprocessing, algorithm development, and fusion are discussed in depth.

### 8.3 Methodology

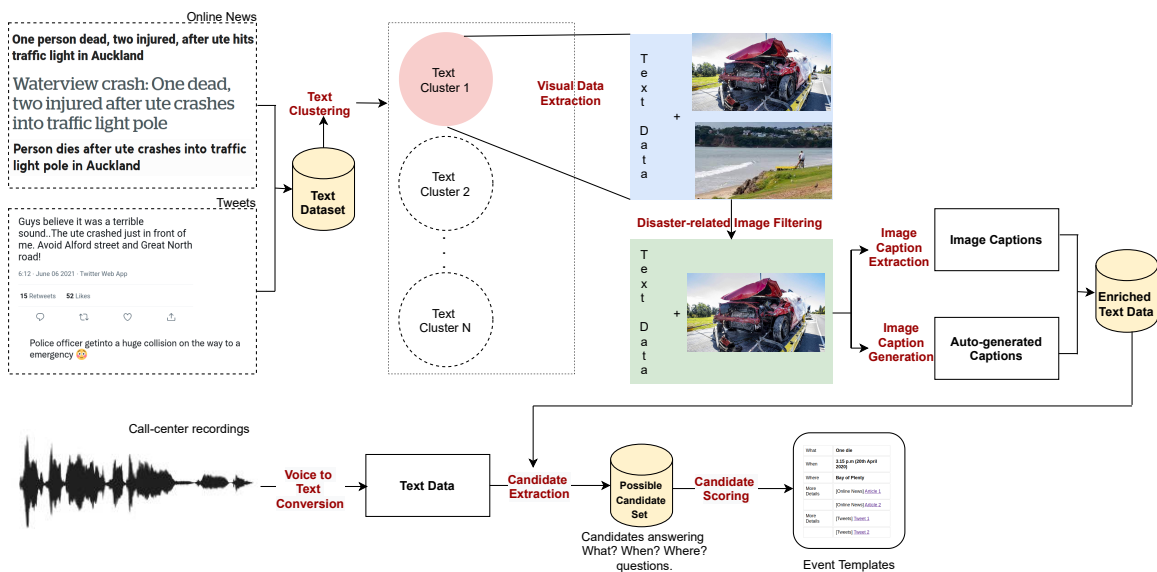


Figure 8.2 The proposed architecture

Fig. 8.2 shows the general architecture of M<sub>2</sub>E<sub>2</sub>S. The system has five core modules (MoDs) and sixteen sub-modules (SMoDs) to collect, preprocess, analyze and extract event templates in real-time as follows.

<sup>3</sup>SUN database, <https://vision.princeton.edu/projects/2010/SUN/>



### Multi-source multimodal data collection and processing

- **MoD<sub>1</sub>: Text data handling**

- *SMoD<sub>1.1</sub> Text data extraction: collect online news and tweets as a scheduled task and filter using key-word based heuristics.*
- *SMoD<sub>1.2</sub> Noise filtering: remove noise from text such as symbols and web links.*
- *SMoD<sub>1.3</sub> Relevant tweet identification: identify disaster-relevant tweets using a supervised algorithm.*
- *SMoD<sub>1.4</sub> Clustering: group news headlines with tweets.*

- **MoD<sub>2</sub>: Visual data handling**

- *SMoD<sub>2.1</sub> Visual data extraction: download images from online news and Twitter posts.*
- *SMoD<sub>2.2</sub> Visual data classification: identify disaster-related images*
- *SMoD<sub>2.3</sub> Visual data captioning: generate natural language sentences describing the scenes in images*

### Candidate selection

- **MoD<sub>4</sub>: Candidate extraction**

- *SMoD<sub>4.1</sub> Candidate extraction from news, tweets, and image captions: jointly extract words/phrases answering “What”, “When” and “Where” (3W) questions from news headlines, body, tweets and image captions.*

### Candidate scoring, event template creation and geolocation

- **MoD<sub>5</sub>: Candidate scoring and Event template creation**

- *SMoD<sub>5.1</sub> Candidate scoring for news, tweets, and image captions: select the best candidates that could answer 3W questions from the possible candidate set in SMoD<sub>4.1</sub>.*
- *SMoD<sub>5.2</sub> Impact information scoring.*
- *SMoD<sub>5.3</sub> Event template creation from news, tweets and image captions: Create event templates answering 3W questions and geolocate.*

The next sections describe the modules and submodules in M<sub>2</sub>E<sub>2</sub>S in detail.

#### 8.3.1 Text data handling - MoD<sub>1</sub>

##### 8.3.1.1 Text data extraction - SMoD<sub>1.1</sub>

Text data for M<sub>2</sub>E<sub>2</sub>S comes from online news and tweets. A Linux scheduled task (cronjob) is used to download online news and tweets every 30 minutes. Online news is collected from three main

online news providers in New Zealand, namely, rnz news<sup>4</sup>, nzherald news<sup>5</sup> and stuff news<sup>6</sup> using their rss feeds.

Tweets are extracted in real-time from the Twitter streaming Application Programming Interface (API) using the python tweepy library<sup>7</sup>. We set the geographical boundary to collect tweets generated by the users within New Zealand. Additionally, to avoid exceeding API limits, the maximum number of tweets downloaded per iteration of the extraction system is limited to 1000 tweets. A rule-based matcher is used to retrieve tweets and news headlines relating to our keywords. Currently, the algorithm captures event types, such as traffic and transport (e.g., crash), weather (e.g., rain, cyclone, flood, storm, snow, wind, hurricane, tornado), societal (e.g., armed conflict, terrorism), fire, earth slip, earthquake, and pandemic. The keyword-based rule matcher used in the system can be modified to limit the extraction for a single event or to add additional event types.

### 8.3.1.2 Noise filtering - *SMoD*<sub>1.2</sub>

The noise filtering module has three components to remove noise from news headlines, tweets, and news body text. This module aims to clean raw text data for further processing. Therefore, we remove any special characters except for letters and numbers from news headlines. To clean up the news body text, special characters, videos, images, embedded tweets, advertisements, and other links are removed.

Due to the use of abbreviated terms, being multilingual, and having irrelevant and redundant data, user-generated Twitter contents compose a significant level of noise. Hence, we remove duplicate tweets, non-English tweets, and Re-Tweets (e.g., RT@ user:). Tweet text is cleaned by removing hashtags, links, words having a length of less than three characters, stop words, special characters, and short sentences (fewer than three words). We also expand abbreviated words (e.g., “ur” → “your”, “2morow” → “tomorrow”). Then we apply standard preprocessing techniques such as sentence splitting, lemmatization, part-of-speech (POS) tagging, dependency parsing, and Named Entity Recognition (NER) using the python spaCy library<sup>8</sup> before applying the clustering algorithm discussed in section 8.3.1.4.

### 8.3.1.3 Relevant tweet identification - *SMoD*<sub>1.3</sub>

Not all tweets are useful for analysis, and it is, therefore, important to distinguish relevant from irrelevant tweets. Thus, this module filters relevant tweets for the analysis tasks. A considerable amount of literature has been published on using ML [59, 336, 216] and DL approaches [267, 17, 401] for identifying useful twitter posts for disaster response tasks. We adopted a Bi-directional LSTM model for relevant tweet classification as described by Algiriyage et al. [24].

### 8.3.1.4 Clustering - *SMoD*<sub>1.4</sub>

The aim of this module is to group similar news headlines and tweets based on the content. We converted the text data into weighted vectors before clustering [346]. The highest accuracy for clustering was achieved while using the Word2Vec model for converting sentences to vectors, known as word embeddings. In the Word2Vec model, vectors are learned so that words with similar meanings will be located near each other in the vector space. Thus, the semantic relationship between words is preserved [254]. We use the Density-based spatial clustering of applications with

<sup>4</sup>rnz news, <https://www.rnz.co.nz/>

<sup>5</sup>nzherald news, <https://www.nzherald.co.nz/>

<sup>6</sup>stuff news, <https://www.stuff.co.nz/>

<sup>7</sup>Python tweepy library version 3.9.0, <https://pypi.org/project/tweepy/>

<sup>8</sup>Python Spacy version 2.3.2, <https://spacy.io/>

noise (DBSCAN) [100] algorithm implemented in the python sklearn library<sup>9</sup> to group similar news headlines and tweet vectors together. This choice was made as the DBSCAN algorithm does not require the number of clusters to be provided [77]. Instead, the algorithm decides the number of clusters it can generate given the data. The DBSCAN algorithm groups together data points that are close to each other based on a distance measurement and a minimum number of points. Therefore, the algorithm requires two parameters to be specified by the user: *eps* - how close points should be to each other to be considered a part of the same cluster and *minPoints* - the minimum number of points needed to form a dense region. At the end of the clustering phase, a collection of clusters with news headlines and tweets or clusters with only tweets is produced.

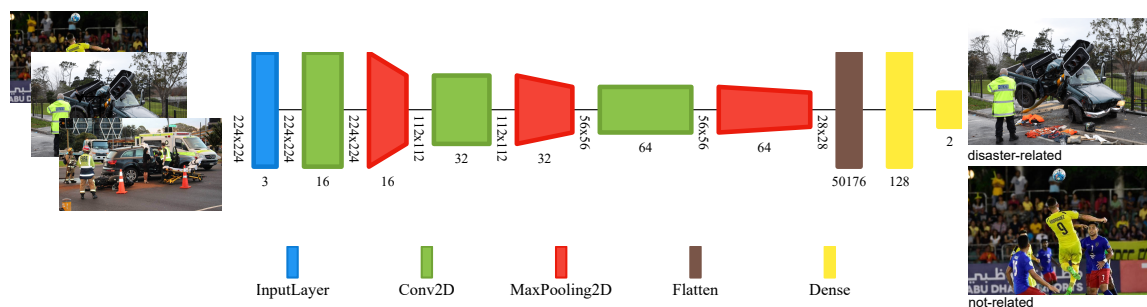
### 8.3.2 Visual data handling - $MoD_2$

#### 8.3.2.1 Visual data extraction - $SMoD_{2,1}$

In this module, images for each of the text clusters identified in the  $SMoD_{1,4}$  are extracted. These images are downloaded from news web pages using the python BeautifulSoup library<sup>10</sup>, and images attached to tweets are extracted using the wget package<sup>11</sup>. As a result, all the images available on news web pages and images posted in tweets of each cluster are downloaded.

#### 8.3.2.2 Visual data classification - $SMoD_{2,2}$

Images from tweets and online news must be filtered to identify disaster-related images. Therefore, this module aims to identify useful images for the analysis task. We used a supervised approach to filter disaster-related images. Our approach uses the Inception-v3 algorithm trained on the CrisisMMD dataset. Additionally, we used 1000 vehicle crash and non-crash images from news web pages in New Zealand to train the algorithm to categorize related and unrelated disaster images [361, 18]. The architecture of the adopted image filtering algorithm is illustrated in Fig 8.3. We trained the model for 500 iterations in the New Zealand eScience Infrastructure (NeSI) high-performance computing cluster. The overall accuracy of the classification model was 92.3%.



**Figure 8.3** Inception-v3-based disaster-related image classification model

#### 8.3.2.3 Visual data captioning - $SMoD_{2,3}$

The aim of this module is to generate natural language descriptions for the filtered images in  $SMoD_{2,2}$ . Image captioning is a well-established field in computer vision that deals with providing a textual description of the scene [382]. Multiple datasets have been developed for the captioning

<sup>9</sup>DBSCAN algorithm, <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>

<sup>10</sup>Beautiful Soup v4.10 <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

<sup>11</sup>wget, <https://www.gnu.org/software/wget/>

tasks such as MSCOCO [224], Flickr 8k [317] and Flickr 30k [410]. However, disaster scenes are significantly different from normal images available in these datasets. The automatic captions generated by pre-trained deep learning algorithms on these datasets do not provide impact information for the images, which is essential for developing situational awareness during disaster response. Therefore, we selected the CrisisMMD dataset for generating image captions [18]. Additionally, we used 587 crash-related images downloaded from New Zealand online news from the previously created image database in section 8.3.2.2. We employed undergraduate student annotators to manually generate the captions with a complete disaster-scene description, including impact information. Arriaga et al. [35] developed instructions for captioning images of dangerous situations such as fires, injured persons, and car accidents. We adopted the guidelines presented by Arriaga et al. [35] as follows for annotating the dataset.

- Write a single English sentence for each image.
- The sentences have to be in the present or present continuous tense.
  - e.g., “a man is lying on the ground, while emergency service personnel are assisting him”.
- Write primarily about the accident/incident/event in the image.
  - When possible, use the impact information (e.g., serious, moderate, minor).
- When possible, be explicit with the vehicle type (e.g., car, van) and the number of vehicles in the image.
  - The sentence should not contain any digits (i.e. 1, 2).
  - Use written numbers instead of digits (i.e. ‘one’, ‘two’).
  - If there are more than two vehicles, always use the term “multiple”.
- If it is visible in the image that people who are attending, be specific and say “police are ...,” or “emergency services are...”.
- There is no limitation in the length of the sentence. However, it is advised to use between 7 to 18 words per sentence [35].

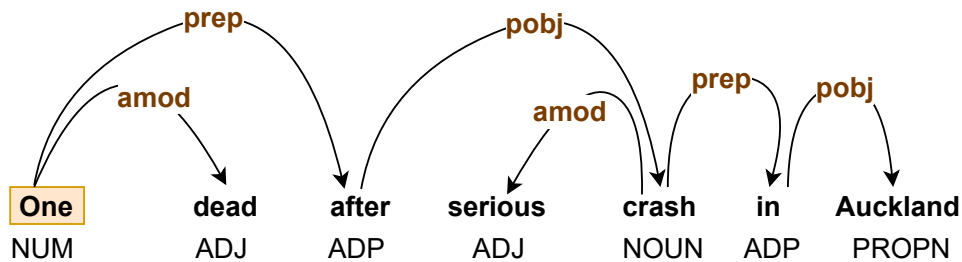
Four annotators were employed and the average cross pairs annotator agreement for a sample dataset resulted in a kappa value of 0.72 [248]. To extract the image features for the captioning model, we used the VGG-16 model. VGG-16 is a deep CNN architecture designed to classify ImageNet datasets into 1000 classes. It consists of 13 convolutional layers, followed by three fully connected layers. It takes an image size of  $(224 \times 224 \times 3)$  as an input and performs convolution operation using a  $(3 \times 3)$  filter. A detailed description of the layers and parameters of the VGG-16 network can be found in [349].

The uniform architecture of VGG-16 has led to it becoming the preferred choice for extracting features from images. It has been effective for several image captioning tasks due to its strong performance [139, 378]. The overall architecture of the VGG-16 model is presented in Fig. 8.4. We take the Global Vectors for Word Representation (GloVe) embeddings of the text captions [295]. The output of the embedding is fed into an LSTM with 256 states. The output of the LSTM (256



python newspaper3k<sup>12</sup> package. The candidate extraction module extracts the answers for *what*, *when*, and *where* questions and words/phrases indicating impact information from news headlines, tweets, image captions, and news body text separately for each cluster.

Dependency parsing and POS tagging are used to extract the part of the text to answer the *what* question. Dependency parsing analyzes the grammatical structure of a sentence, establishing relations among words, whereas POS tagging identifies the grammatical categories of words such as tense [71]. Fig 8.6 and Table 8.2 show the dependencies and POS tagging of the news headline “One dead after serious crash in Auckland”.



**Figure 8.6** Dependency parsing of news headline “One dead after serious crash in Auckland”. The *ROOT* node is highlighted in orange.

**Table 8.2** Dependencies and POS of the news headline “One dead after serious crash in Auckland”

word	lemma	dependency	POS
One	one	ROOT	NUM
dead	dead	amod	ADJ
after	after	prep	ADP
serious	serious	amod	ADJ
crash	crash	pobj	NOUN
in	in	prep	ADP
Auckland	Auckland	pobj	PROPN

We analyzed the news headline reporting pattern of over 700 New Zealand news articles published in 2020 and 2021 on vehicle crash events and created rules based on the syntactic structure of the news headlines. Our algorithm starts from the *ROOT* node, selects multiple sub-trees, and builds a new text from that. By analyzing the grammatical structures of the headlines, we were able to identify several relationship patterns of the *ROOT* node and other nodes. Fig 8.7 illustrates some of these patterns. According to the patterns, we created rules as illustrated in algorithm 2. Let’s consider the example sentence “One dead after serious crash in Auckland” in Fig 8.6. In this case, the *ROOT* node is **One**, and its *POS* is **NUM**. According to algorithm 2, the “subtree2” function is called, and dependency of *amod* is passed to check. The “subtree2” function checks subtrees for children of **One** having a dependency of *amod*. As shown in Fig 8.6, there is only a single child **dead**. So the word **dead** is returned by the function “subtree1” and finally, “extractWhat” function returns the joined phrase “**One dead**” as the result. If there are multiple sentences (e.g., tweets and news bodies), our algorithm first splits the text into sentences before applying dependency parsing. Therefore, several candidates are extracted for each sentence.

To extract *where* candidates, we use the NER implementation in spaCy<sup>13</sup>. Our algorithm

<sup>12</sup>Python newspaper3k, <https://newspaper.readthedocs.io/en/latest/>

<sup>13</sup><https://spacy.io/api/annotation#named-entities>

<b>dependency</b>	nummod	nsubj	aux	ROOT	prep	nummod	punct	compound	pobj	prep	pobj
<b>word</b>	One	person	has	died	after	two	-	vehicle	crash	in	Katikati
<b>POS</b>	NUM	NOUN	AUX	VERB	ADP	NUM	PUNCT	NOUN	NOUN	ADP	PROPN

<b>dependency</b>	compound	nsubj	ROOT	compound	dobj	nummod	dobj	prep	pobj
<b>word</b>	Truck	crash	closes	State	Highway	39	south	of	Hamilton
<b>POS</b>	NOUN	NOUN	VERB	PROPN	PROPN	NUM	ADV	ADP	PROPN

<b>dependency</b>	nsubj	ROOT	prep	pobj	prep	compound	pobj
<b>word</b>	Three	taken	to	hospital	following	Timaru	crash
<b>POS</b>	NUM	VERB	PART	VERB	VERB	PROPN	NOUN

<b>dependency</b>	nsubj	ROOT	prep	pobj	prep	poss	case	compound	pobj
<b>word</b>	Person	dead	after	incident	on	Auckland	's	Southern	Motorway
<b>POS</b>	NOUN	ADJ	ADP	NOUN	ADP	PROPN	PART	PROPN	PROPN

**Figure 8.7** Relationship patterns of ROOT node and other nodes

extracts items tagged as Geo-Political Entities (GPEs) (e.g., countries, cities, states), Non-GPE locations (LOC) (e.g., mountain ranges, bodies of water), and FAC (e.g., buildings, airports, highways, bridges) by the NER tool.

We extract *when* candidates, including both relative and absolute time references from the text (see Table 8.3) using both NER and regular expressions. For relative time expressions, our system includes the news reported date and time or tweet generated date and time to extract the precise temporal information.

**Table 8.3** Absolute and relative time expressions.

Type	Expression	Example
Absolute time	Day Month Year Hour:Minute	21 May 2019 21:27
Relative time	today, afternoon, morning	3:50 this afternoon, 10 AM in the morning

Rules constructed using spaCy rule-based matcher<sup>14</sup> are used to extract *impact* information. The following are the ten rules to match impact data.

<sup>14</sup>spaCy rule-based matching, <https://spacy.io/usage/rule-based-matching>



**Algorithm 2** Extract answer for “what” question

```

1: function SUBTREE1(word, pos)
2:   if children for POS=“pos” then
3:     return (join children)
4:   else
5:     return nothing
6:   end if
7: end function

8: function SUBTREE2(word, dep)
9:   if children for DEP=“dep” then
10:    return (join children)
11:  else
12:    return nothing
13:  end if
14: end function

15: function EXTRACTWHAT(sentence)
16:   for each word in sentence do
17:     if word.dependancy = “ROOT” and
        word.pos=“VERB” then
18:       join [subtree2(word,“nsubj”),
19:            word,
20:            subtree1(word,“PART”),
21:            subtree2(word,“dobj”)]
22:       else if word.dependancy = “ROOT” and
        word.pos=“NUM” then
23:         join [word,
24:              subtree2(word,“amod”)]
25:       else if word.dependancy = “ROOT” and
        word.pos=“NOUN” then
26:         join [subtree(word,“NOUN”),
27:              word,
28:              subtree(word,“VERB”)]
29:       else
30:         subtree2(word,“nsubj”),
31:         word
32:       end if
33:     end for
34: end function

```

```

impact_words = ['hurt', 'injury', 'serious', 'die', 'dead', 'moderate', 'minor']
p1 = ['POS': 'NOUN', 'OP': '*', 'POS': 'PROPN', 'OP': '*', 'LOWER': 'closed']
p2 = ["LOWER": "closes", 'LOWER': 'state', 'LOWER': 'highway', 'POS': 'NUM']
p3 = ['POS': 'NUM', 'LOWER': 'people']
p4 = ["TEXT": "IN": impact_words]
p5 = ['POS': 'NUM']
p6 = ['POS': 'PROPN', 'OP': '+', 'POS': 'PROPN', 'OP': '+', 'POS': 'NUM']
p7 = ['POS': 'VERB', 'POS': 'ADP', 'POS': 'ADJ', 'OP': '?', 'POS': 'NOUN']
p8 = ['LOWER': 'closes', 'POS': 'PROPN', 'OP': '+', 'POS': 'PROPN', 'OP': '+', 'POS':
'NUM', 'OP': '?']
p9 = ['LOWER': 'closes', 'POS': 'NOUN']
p10 = ['LOWER': 'closes', 'POS': 'PROPN']

```

To further explain our approach to candidate extraction, we consider a cluster generated from the extraction system of news and tweets relating to a vehicle crash event in the Bay of Plenty, New Zealand. Tables 8.4, 8.5, 8.6, and 8.7 show candidate extraction results for the news headlines, tweets, news body text and image captions respectively.

**Table 8.4** Candidates extracted from news headlines

News Headline	What?	When? Where?	Impact
‘One dead after two motorbikes crash in Bay of Plenty’	One dead	- Bay of Plenty	one dead
‘Motorcycles crash leaves one dead near Whakatāne’	Motorcycles - crash leaves one dead	Whakatāne	one dead



**Table 8.5** Candidates extracted from tweets

<b>Tweet</b>	<b>What?</b>	<b>When?</b>	<b>Where?</b>	<b>Impact</b>
‘I found a colleague was in the serious crash in Bay of Plenty today. Here’s hoping it’s a speedy recovery‘	I found, hoping	today	Bay of Plenty	serious crash
‘Farmer dies in bike crash in Bay of Plenty‘	Farmer dies	-	Bay of Plenty	dies

**Table 8.6** Candidates extracted from news body text

<b>News Body Text</b>	<b>What?</b>	<b>When?</b>	<b>Where?</b>	<b>Impact</b>
‘One person has died after two motorbikes crashed in the Bay of Plenty. The crash occurred at 3.15pm on Bell Rd in Nukuhou, south of Whakatāne, police said. Another person suffered minor injuries in the crash. WorkSafe had been advised and the Serious Crash Unit was in attendance, police said‘	[One person died, police said, Another person suffered minor injuries]	3.15pm	Bay of Plenty, Bell Rd, Nukuhou, Whakatāne	one died, minor injuries, serious crash
‘One person has died and another has minor injuries following a serious crash in Nukuhou, near Whakatāne in Eastern Bay of Plenty. The crash involved two motorcycles and was reported around 3.15pm., The Serious Crash Unit and WorkSafe are attending the scene‘	[One person died, The crash involved two motorcycles, The Serious Crash Unit and WorkSafe attending the scene]	3.15pm	Nukuhou, Bay of Plenty, Whakatāne	one died, minor injuries, serious crash

### 8.3.4 Candidate scoring and event template creation - $MoD_5$

#### 8.3.4.1 Candidate scoring for news, tweets, and image captions - $SMoD_{5,1}$

Candidate scoring is mainly associated with determining the best candidates to answer the 3W questions after the candidate extraction process. Therefore, our algorithm jointly extracts the best answers for 3W candidates from news, tweets, and image captions. We assume that online news is reported using the inverted pyramid structure; a system of news writing that arranges facts in descending order of importance [278]. Candidates are scored based on their position, frequency, and appearance across news, tweets, and images (cross-media reference). Furthermore, a novel location relatedness score is introduced in identifying the best candidate for the *where* question. Table 8.8 illustrates the different scores used for candidate scoring, and each of the scores is defined below.

*Position score* ( $S_{pos}(C)$ ): A higher score is assigned if candidates are found early in the text

**Table 8.7** Candidates extracted from image captions

Image caption	What?	When?	Where?	Impact
‘Serious crash investigators are looking into the cause of the crash’	One dead	-	-	serious crash
‘police responding to crash’	Police responding	re-	-	responding to crash

Where	When	What
Position Score	Position Score	Position Score
Frequency Score	Frequency Score	Frequency Score
Cross-media Reference Score	Cross-media Reference Score	Cross-media Reference Score
Location Relatedness Score		

**Table 8.8** Scores used to determine the best candidates among the candidate set extracted from news headlines, body, tweets, and image captions

considering the inverted pyramid structure of news reporting. For occurrences in the first sentence of the body text or in the headline, a score of 1 is assigned. For occurrences in subsequent sentences, the score follows an exponential decay, decreasing with an increase in position,  $p$ ,  $S_{pos}(C) = e^{-dp}$ , with  $e$  being the exponential constant and decay coefficient,  $d > 0$  [278]. If we select logarithmic decay (i.e.  $d = \log(2)$ ), we divide the score by half whenever we move farther away from the headline or first sentence in the lead paragraph. For example, consider the location candidates in the headlines and body in Tables 8.4 and 8.6. Let  $X = [x_1, x_2, \dots, x_n]$  be the vector of news headlines and  $Y = [y_1, y_2, \dots, y_n]$  be the vector of the news body text in the cluster. We calculate the position score for candidates presented in a headline  $S_{pos}(C_h) = \sum_{i=1}^n S_{pos}(C_{x_i})$ , and the position score for candidates in body text  $S_{pos}(C_b) = \sum_{i=1}^n S_{pos}(C_{y_i})$ . Finally, the full position score is obtained using Eq. 8.1.

$$S_{pos}(C) = S_{pos}(C_h) + S_{pos}(C_b) \quad (8.1)$$

Table 8.9 shows an example of full position score calculation for *where* candidates of the example news headlines and body text in Tables 8.4 and 8.6.

**Table 8.9** Position score calculation for *where* candidates of the news in Tables 8.4 and 8.6

Locations in headline	$S_{pos}(C_h)$	Locations in body	$S_{pos}(C_b)$	$S_{pos}(C)$
Bay of Plenty	1.0	Bay of Plenty	2.0	3.0
-	-	Bell Rd	0.5	0.5
-	-	Nukuhou	1.5	1.5
Whakatāne	1.0	Whakatāne	1.5	2.5

*Frequency score* ( $S_{freq}(C)$ ): We score candidates by their frequency of occurrence in the news tweets and image captions. The frequency values are transformed by scaling them between 0 and 1. Table 8.10 illustrates frequency scores for *where* candidates of the examples in Tables 8.4, 8.5, 8.6, and 8.7.

*Location relatedness score* ( $S_{rel}(C)$ ): One of the main problems in location identification from text is the location ambiguity where several distinct locations have the same name (e.g., Kingston

**Table 8.10** Frequency score calculation for *where* candidates of the examples in Tables 8.4, 8.5, 8.6 and 8.7

Location	Frequency	$(S_{frq}(C))$
Bay of Plenty	5	1.0
Bell Rd	1	0.0
Nukuhou	2	0.25
Whakatāne	1	0.0

in New York and Kingston in Jamaica or London) [297]. Our algorithm resolves ambiguity issues based on spatial proximity clues. We introduce a new location relatedness score to identify the most accurate spatial candidate. First, the location names are transformed into coordinates using Nominatim<sup>15</sup>, which uses free data from OpenStreetMap<sup>16</sup>. Then, the geodesic distance among all location candidate pairs is calculated using Python geopy library<sup>17</sup>. Finally, we assign a higher score for the closest location candidates and lower scores for the distant location candidates following the logarithmic decay function. Table 8.11 illustrates the location relatedness scores obtained for *where* candidates.

**Table 8.11** Location relatedness score calculation for *where* candidates of the news and tweet examples in Tables 8.4, 8.5 and 8.6.

Location pairs	Pair wise distance	$S_{rel}(C)$
Bay of Plenty, Bell Rd	459.9687	0.03125
Bay of Plenty, Nukuhou	14.0604	0.50000
Bay of Plenty, Whakatāne	3.0771	1.00000
Bell Rd, Nukuhou	457.6903	0.03125
Bell Rd, Whakatāne	461.4393	0.03125
Nukuhou, Whakatāne	11.8114	0.50000

*Cross-media reference score* ( $S_{cross}(C)$ ): The system analyses the occurrences of candidates across tweets, online news, and image captions and assigns a score of 1 if the candidates appeared in all media. If the candidate appears in two media items, then a score of 0.5 is allocated. Otherwise, the score is 0. For example, the location Bay of Plenty is in both tweets and online news and thus gets a score of 0.5.

Finally, we obtain the full candidate score by summing the individual score values (see Eq. 8.2) and select the highest-scoring candidate as the correct answer to the question. As can be seen from Table 8.12, the final score is designed to be between 0 and 1. Therefore, the system chooses Bay of Plenty as the most suitable candidate for the *where* question in our example. Similarly, the scores mentioned in Table 8.8 are calculated to select the candidates for the *what* and *when* questions.

$$S_{full}(C) = S_{pos}(C) + S_{frq}(C) + S_{rel}(C) + S_{cross}(C) \quad (8.2)$$

<sup>15</sup>Nominatim version 3.5.1, <https://github.com/osm-search/Nominatim>

<sup>16</sup>OpenStreetMap, <https://www.openstreetmap.org/#map=2/-41.2/-6.6>

<sup>17</sup>geopy 2.2.0 <https://pypi.org/project/geopy/>

**Table 8.12** Final *where* candidate selection of the examples in Tables 8.4, 8.5, 8.6 and 8.7

Location	Final Score
Bay of Plenty	1.0
Bell Rd	0.0
Nukuhou	0.5
Whakatāne	0.5

### 8.3.4.2 Impact information scoring - $SMoD_{5.2}$

A disaster can result in a variety of impacts on people and the environment. For example, we observe that during many crash incidents, there exist impacts such as road deaths, injuries, and road closures (see Table 8.13). Moreover, the level of impact is highlighted using certain keywords (e.g., serious, major, and minor).

**Table 8.13** Traffic accident-related impacts during 2020 and 2021 in New Zealand

Impact	Number of Articles	Example
Road death	141	“Person <i>dead</i> after incident on Auckland’s Southern Motorway”
Injury	115	“One person has <i>died</i> after two-vehicle crash in Katikati” “Serious <i>injuries</i> after truck and car crash near Omokoroa in Western Bay of Plenty” “Critical <i>injuries</i> from single-vehicle crash near Tauranga”
Road closure	53	“Canterbury highway <i>closed</i> following three-vehicle crash” “Serious crash <i>closes</i> SH1 in Northland”

Therefore, we remove duplicates of the extracted words/phrases indicating impact information. Then we assign a score considering both impact and impact level as listed in Table 8.14.

**Table 8.14** Scoring of impact information

Impact keyword	score
“serious” OR “severe” OR “major” OR “critical” OR “fatal” OR “dead” OR “die” OR “kill”	1.0
“moderate” OR “injury”	0.5
“minor” OR “close” OR “closure”	0.25

### 8.3.4.3 Event template creation from news, tweets, and image captions - $SMoD_{5.3}$

Event templates are created showing answers to 3W questions jointly from online news, tweets, and image captions to improve SA of disaster responders. The main screen of M<sub>2</sub>E<sub>2</sub>S shows the summarized information regarding the event. The location of the incident is pinpointed on a map and more information is provided using the drill-down approach to avoid information overload for the emergency responders. For example, when the user clicks on the pinpoint of the map, the first screen shows the 3W questions and answers, some images, and the impact information. Then, when they click the ‘view more’ link, they can see the related online news, tweets, and images (see Figure 8.8).

A live demonstration of M<sub>2</sub>E<sub>2</sub>S is available through <https://rangikanilani.github.io/events.html>. Furthermore, the symbols shown in Fig. 8.9 are used to display the impact information.

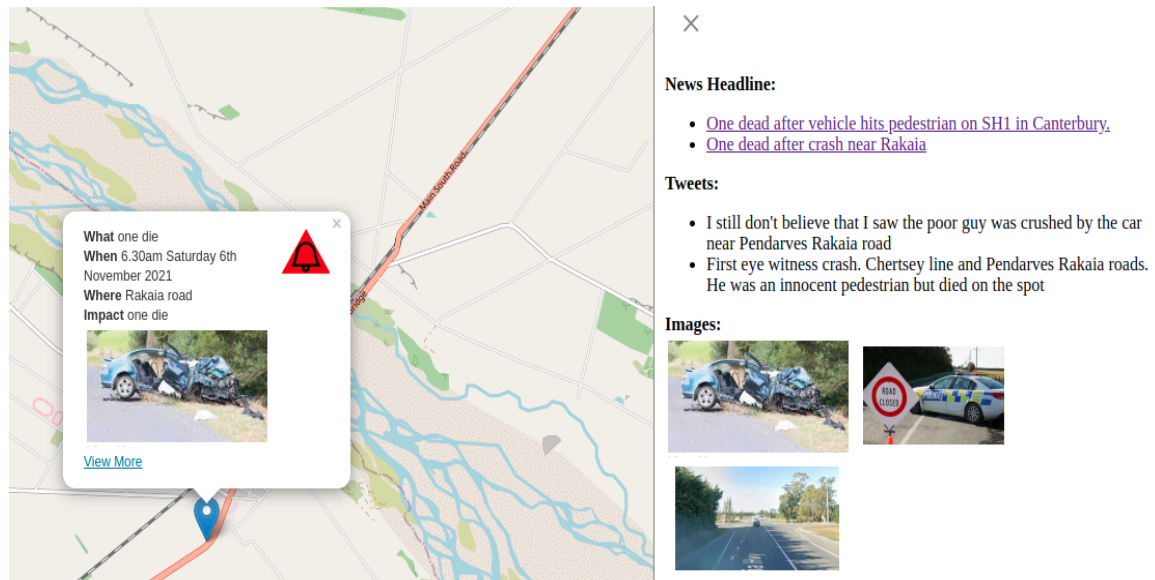


Figure 8.8 Impact information visualization in event templates



Figure 8.9 Symbols used to show impact information

## 8.4 Evaluation

The developed system was evaluated using generalised precision ( $gP$ ), a score suitable for retrieval performance evaluations as described by Kekäläinen et al. [188] and used by Hamberg et al. [125, 127].  $gP$  is calculated using the following formula

$$gP = \sum_{d \in R} r(d)/n \quad (8.3)$$

where  $R$  is the set  $n$  event templates from a database  $D = d_1, d_2, \dots, d_N$ . Let the event template  $d_i$  in the database have relevance scores of  $r(d_i)$  being real numbers ranging from 0.0 to 1.0.

To conduct the evaluation experiments we executed  $M_2E_2S$  for three months from 1st of January 2022 to 31 March 2022 to generate real-time events. The dataset contained 56 news articles and 63 tweets relating to traffic incidents. We generated event templates with  $M_2E_2S$  using only tweets text data, tweets + news text data, and finally tweets + news text and visual data. There were 56 event templates from news text data and 17 event templates from both news and text data. We recruited two graduate students as assessors. They were given the grouped tweets and news that the system used to identify event templates. After reading each tweet, article, or image caption in a group we provided the assessors the 3W and impact phrases that had been extracted by the system and asked them to judge the relevance of each answer on a 3-point scale as follows;

$$r(d_i) = \begin{cases} 0, & \text{if an answer contained no relevant information} \\ 0.5, & \text{if only part of the answer was relevant or if the information was missing} \\ 1, & \text{if the answer was completely relevant without missing information} \end{cases}$$

Table 8.15 shows the Average generalized precision ( $gP$ ) scores

**Table 8.15** Generalised precision scores of 3W+Impact

**Table 8.16** Tweets text data

	$gP$ Score
What	0.68
Where	0.74
When	0.85
Impact	0.71
Avg (3W)	0.76
Avg (3W)+Impact	0.75

**Table 8.17** News + tweets text data **Table 8.18** News + tweets text and visual data

	$gP$ Score		$gP$ Score
What	0.78	What	0.80
Where	0.92	Where	0.94
When	0.86	When	0.86
Impact	0.86	Impact	0.89
Avg (3W)	0.85	Avg (3W)	0.87
Avg (3W)+Impact	0.86	Avg (3W)+Impact	0.87

Our system achieved a  $gP$  score of 0.68 for the *what* answer extraction using only tweets text data. Furthermore, the  $gP$  score has increased after using both tweet and news text data, as can be seen from Table 8.17. Moreover, after integrating visual data with tweets and news text data, the  $gP$  score increased by 0.02. Similarly, *where* answer extraction has also improved after integrating tweet and news data. For instance, location extraction from only tweet data has achieved a  $gP$  of 0.74. However, the extraction has considerably improved after using news data, having a  $gP$  score of 0.92. The combination of textual and visual data has strengthened the *where* answer extraction by increasing the  $gP$  score by 0.02. Using both tweet and news data has led to an improvement in the  $gP$  score of *when* answer extraction. However, there is no significant improvement after incorporating visual data. A possible explanation for this might be that the temporal information is not inferred from visual data through captioning. The extraction of impact data follows a similar pattern to that of *what* and *where* answers. For example,  $gP$  score has increased with the incorporation of news data and visuals.

Although there is no similar system available for direct comparison, we compare our results with Hamborg et al. [125]. Their event extraction system, “Giveme5W1H” has achieved an average precision of 0.79 for *what* answer extraction and our system has achieved 0.80. Moreover, M<sub>2</sub>E<sub>2</sub>S has got a  $gP$  of 0.94 for *where* extraction which is 0.16 higher than the average precision of Giveme5W1H [125]. Furthermore, we achieved a  $gP$  of 0.86 for *when* answer extraction, whereas Hamborg et al. achieved 0.78.

As shown in Table 8.18, the results indicate that the inclusion of news data has increased the of 3W+Impact answer extraction accuracy than using only tweets. Moreover, the fusion of both text and visual data has significantly improved the extraction results. The most striking observation to emerge from the results is that the fusion of multi-source multimodal data significantly improves the event extraction accuracy.

## 8.5 Discussion and Conclusions


A high volume of data is generated from smart mobile devices through social media applications during disasters. However, current disaster response activities barely use most of these data due to data processing and integrity issues. To address this problem, our study set out to cross-validate social media data with other disaster data sources and data modalities in real-time. As a result, a novel software system, namely M<sub>2</sub>E<sub>2</sub>S, for extracting real-time event templates from multi-source multimodal data, is introduced. The present study integrated tweets with online news text data and visual data. Furthermore, this study found that combining text and visual data from tweets and online news results in better information extraction performance (overall  $gP = 0.87$ ) than utilizing them alone (overall  $gP = 0.75$ ).

The findings in this paper are subject to at least three limitations. Firstly, our system extracts answers for the 3W questions. A comprehensive event extraction system could provide more information by extracting answers for the more specific '5W1H' questions. However, finding answers for 5W1H from natural language processing is challenging, as differentiating and identifying answers is incredibly complex, even for humans. Therefore, future research should focus on the need for extracting more information from natural language to potentially inform such answers. Secondly, we have evaluated the proposed system in the traffic and transportation domain. However, it would be interesting to see if the developed methods are applicable to other disaster events such as floods, earthquakes, and landslides. Therefore, further experiments should be carried out to evaluate the M<sub>2</sub>E<sub>2</sub>S for other disaster scenarios. Finally, we have integrated text and image data for the event extraction. However, in addition to those data formats, video data contain a lot of information covering a wider range of aspects. For example, people upload video data more frequently when they tweet. Furthermore, news reporters also provide video recordings of disaster incidents. Additionally, CCTV monitoring provides real-time surveillance footage. Due to the availability of video data, further work needs to modify the architecture of M<sub>2</sub>E<sub>2</sub>S to accommodate video data integration for event extraction.

A demonstration of M<sub>2</sub>E<sub>2</sub>S is available at:<https://rangikanilani.github.io/events.html>

## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

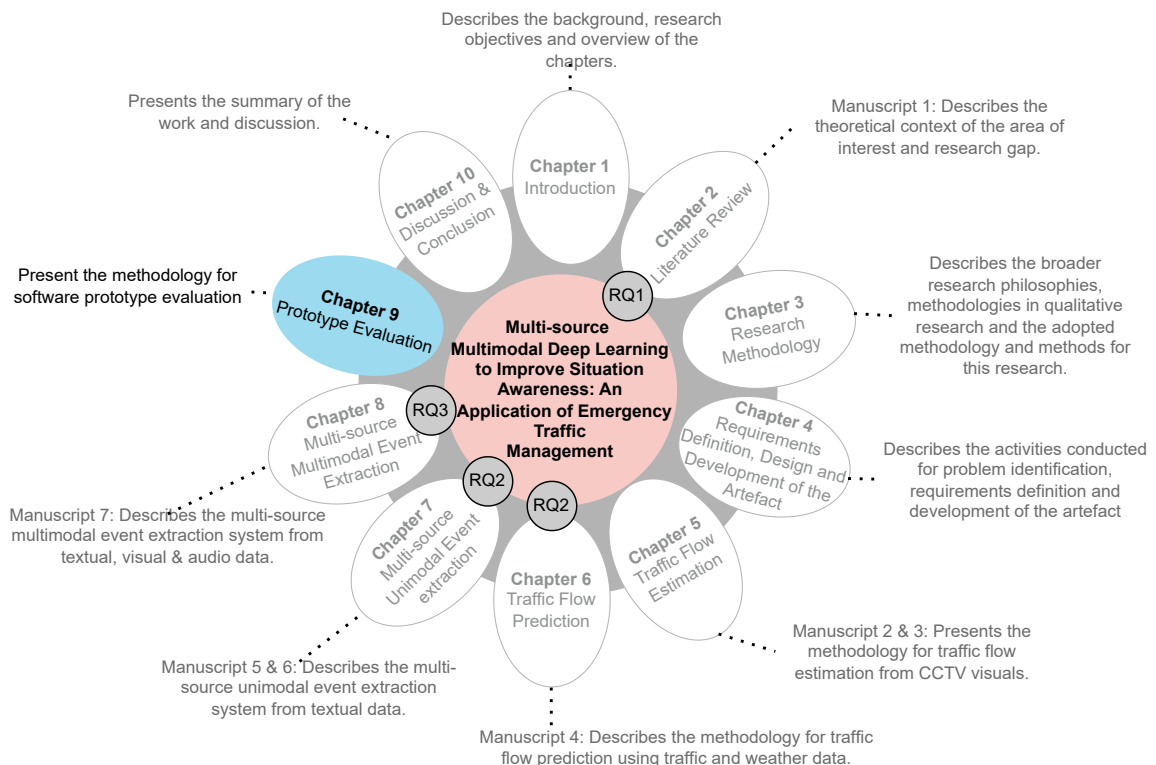
Name of candidate:	Rangika Nilani
Name/title of Primary Supervisor:	Dr Raj Prasanna
In which chapter is the manuscript /published work:	Eight
<p>Please select one of the following three options:</p> <p><input type="radio"/> The manuscript/published work is published or in press</p> <ul style="list-style-type: none"> <li>• Please provide the full reference of the Research Output:</li> </ul> <p><input type="radio"/> The manuscript is currently under review for publication – please indicate:</p> <ul style="list-style-type: none"> <li>• The name of the journal:</li> <li>• The percentage of the manuscript/published work that was contributed by the candidate: <span style="float: right;">85.00</span></li> <li>• Describe the contribution that the candidate has made to the manuscript/published work: The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.</li> </ul> <p><input checked="" type="radio"/> It is intended that the manuscript will be published, but it has not yet been submitted to a journal</p>	
Candidate's Signature:	
Date:	11-Dec-2022
Primary Supervisor's Signature:	Raj Prasanna <small>Digitally signed by Raj Prasanna Date: 2022.12.12 04:53:50 +13'00'</small>
Date:	12-Dec-2022

This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/publication or collected as an appendix at the end of the thesis.



# Chapter 9

## Prototype Evaluation



This chapter describes the process used to evaluate the artefact and the findings of the evaluations. Artefact evaluation is an essential step in DSR, which aims to determine if the process developed in a study incorporates user needs. Therefore, the three activities of the method framework provided by Johannesson et al. [175], were used as the roadmap for designing and evaluating the software prototype. These activities include “Analyse Evaluation Context”, “Select Evaluation Goals and Strategy” and “Design and Carry Out Evaluation”. Sections 9.2, 9.3 and 9.4 discuss how these activities were adapted in this research to evaluate the software prototype.

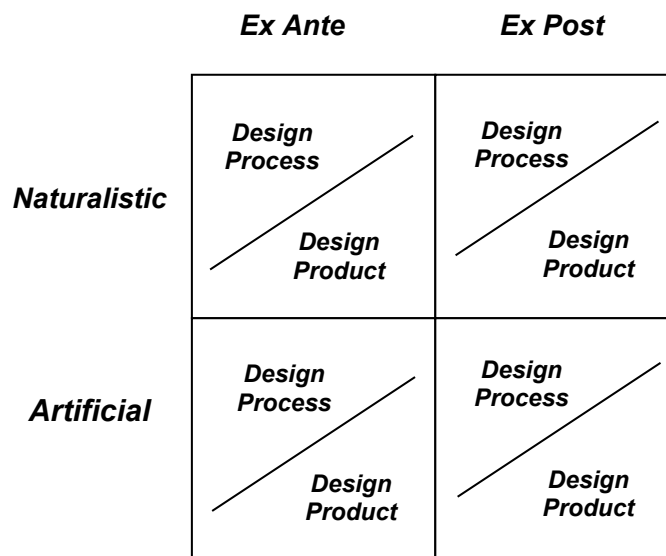
### 9.1 Evaluation in DSR research

In DSR, evaluation is regarded as a crucial activity to prove rigor and label the research as “science” [385, 245, 178, 141]. The artefact evaluation addresses the question, “How well does the artefact solve the explicated problem and fulfil the defined requirements?” [175]. Furthermore, the evaluation process helps to validate if the artefact is relevant to a problem domain and helps to

increase the overall quality of the resulting artefact. Moreover, it contributes to demonstrating the artefact’s feasibility and efficacy by confirming the underlying requirements used in its development [141, 253].

Venable et al. [376] suggest two types of evaluations in DSR, namely “formative” and “summative”. Formative evaluation involves evaluating an artefact while it is still being designed to learn how to make it better during subsequent design processes. As a result, formative evaluation is a part of an iterative design process, in which the artefact is designed and evaluated during multiple iterations. In comparison, summative evaluation aims to evaluate an artefact after its design and development have been finalised. The results of a summative evaluation are utilised to acquire a final assessment of the utility of the artefact rather than feeding back into the design process.

Hevner et al. [385] identified evaluation as “crucial” and noted the requirement for researchers to rigorously evaluate design artefacts. They identified five evaluation methods: observational, analytical, experimental, testing, and descriptive. However, they do not provide much guidance in choosing among the various evaluation methods. Therefore, addressing this gap, Pries-Heje et al. [310] proposed a strategic framework for choosing among evaluation strategies and methods after analysing a large portion of the literature. A strategic framework could serve (at least) two purposes. First, it could help Design Science researchers build strategies for evaluating their research outcomes and achieving improved rigor in DSR. Second, it could be used descriptively to better understand unstated evaluation strategies in existing DSR reports [310]. The proposed strategic framework for evaluating the DSR artefacts involves two dimensions: time, and evaluation method (see Figure 9.1).



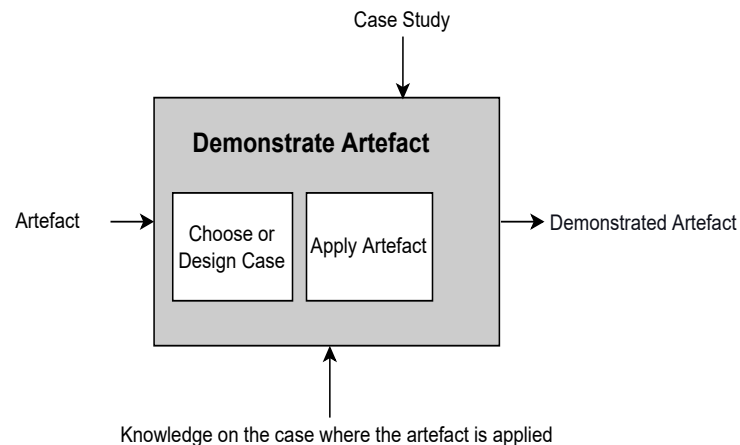
**Figure 9.1** Evaluation framework proposed by Pries-Heje et al. [310]

Pries-Heje’s [310] framework seeks to answer four important questions.

- What is actually being evaluated?
- How is it being evaluated?
- When was it evaluated?
- Who is evaluating?

“What” is being evaluated involves choosing between the design product and the design method. “How” to evaluate decides between naturalistic or artificial forms of evaluation. In a naturalistic evaluation, the artefact is evaluated in its intended context or the real world. Therefore, a naturalistic evaluation involves real users using real systems to solve real problems. In comparison, an artificial evaluation evaluates an artefact in a fabricated and artificial environment, such as a laboratory [360]. “When” to evaluate is selected from ex ante, ex post, or both. Ex ante evaluations occur before the system is constructed, and ex post evaluations occur after the system is constructed. “Who” includes elements such as the evaluation context, which includes real users, organisations, and problems.

This research followed the DSR method framework developed by Johannesson et al. [175], as mentioned in Chapter 3 section 3.2.4. The last two activities of Johannesson’s framework are Demonstrate Artefact and Evaluate Artefact. Demonstrate Artefact, activity shows how to apply the artefact in a specific case, proving its feasibility. Johannesson suggests a demonstration as a lightweight evaluation [175]. Therefore, this activity addresses the question, How can the developed artefact be used to address the explicated problem in one case?. There are two sub-activities of Johannesson’s framework: choose or design case and apply artefact [175]. The first activity is related to developing a well-documented case from the literature, a real-life case, or a combination of these. The second sub-activity is to apply the artefact to the chosen case, which includes documenting the outcome of the application. This activity is summarised in Figure 9.2.

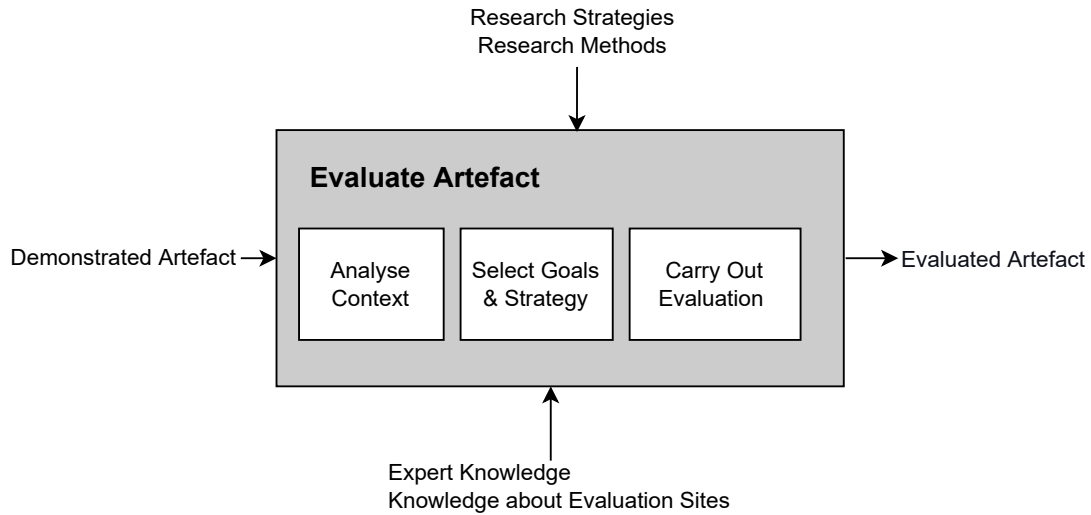


**Figure 9.2** Demonstrate Artefact 8.1

The final activity of Johannesson’s method framework is Evaluate Artefact, in which researchers decide how well the artefact solves the stated problem and to what extent it meets the requirements [175]. Therefore, the activity answers the question, “How well does the artefact solve the explicated problem and fulfil the defined requirements?”. There are three activities to be carried out during the artefact evaluation as follows:

1. Analyse Evaluation Context—Researchers analyse and describe the evaluation context
2. Select Evaluation Goals and Strategy—Researchers decide on evaluation goals, strategies, and methods to apply.
3. Design and Carry Out Evaluation—Researchers design the evaluation study in detail and then execute it.

Evaluate Artefact activity is illustrated in Figure 9.3.



**Figure 9.3** Evaluate Artefact activity

During this research, the fourth and fifth activities of Johannesson’s framework were combined together.

Sections 9.2, 9.3 and 9.4 discuss the evaluation process.

## 9.2 Analyse Evaluation Context

During this activity, the project was clearly analysed to understand the context, including resource constraints, such as time and access to users or organisations. Therefore, as the first step, the answers to the main questions raised by Pries-Heje’s [310] were identified as follows:

- What is actually being evaluated?

The software prototype developed during this research was described in Chapter 4. The final artefact comprised four components developed individually. These components are discussed in Chapter 5, Chapter 6, Chapter 7 and Chapter 8. The final software prototype comprising all four components was assessed during the evaluation process.

- How is it being evaluated?

The choice between an artificial and a naturalistic strategy depends on the qualities of the artefact to be evaluated. Naturalistic evaluations are carried out in real settings, while artificial evaluations are done in artificial settings such as lab environments. The naturalistic evaluation was chosen considering the real-time nature of the artefact and the significance of stakeholder participation in the evaluation.

- When was it evaluated?

The software prototype was evaluated ex ante. The rationale for ex ante is that the artefact was an early prototype that had not been fully developed.

- Who is evaluating?

Previous studies have shown that the feedback provided by experts is extremely useful for artefact evaluation [80, 84, 293]. Furthermore, multiple experts in a study can also help to identify different perspectives on an artefact and can help to produce better results by combining their views [141]. Therefore, the software prototype was evaluated by a group of experts representing transport management and city councils in New Zealand. The following criteria were considered while selecting candidates for the evaluation.

- Candidate participated in the initial requirement collection interviews.  
and/or
- Candidate works in the transport management domain.  
and/or
- Candidate has experience in developing/working with transport-related Information Systems (ISs)

Seven experts were selected for the evaluation, and the majority were involved in the design process of the artefact of this project from beginning to end. The background information of the selected participants for the prototype evaluation is summarized in Table 9.1.

**Table 9.1** Summary of the selected experts for the prototype evaluation

<b>Position</b>	<b>Organization</b>	<b>Years of Experience</b>
Program Director	Auckland Transport	15+ years
Computer Vision Specialist	Auckland Transport	10+ years
Project Manager	City Council	15+ years
Optimisation Delivery Manager	Transport Operations Centre	15+ years
Journey Manager	Auckland Transport	15+ years
Manager	Transport Operations Centre	15+ years
Senior ITS Engineer	Transport Operations Centre	15+ years

The experts chosen had considerable expertise in their domains, and the rationale for selecting them from various transportation management perspectives was to evaluate the artefact based on their everyday decision-making needs from several angles.

### 9.3 Select Evaluation Goals and Strategy

Johannesson et al. [175] suggest that a researcher must choose the goals and evaluation strategy, as well as research methodologies and methods, based on an understanding of the evaluation context. The goals of the evaluation have to be decided by answering the questions such as “Is the evaluation formative or summative?”, “Which are the most important attributes to evaluate?”. Therefore, the goals for the evaluation are selected as follows:

- Is the evaluation formative or summative?

Based on the discussion in Section 9.1, a summative evaluation aims to assess an artefact after its design and development have been fully completed. On the other hand, formative

evaluation provides a way for subsequent design activities to iteratively improve an artefact still under design. Therefore, formative evaluation was chosen as the prototype was in the early stage of design.

- Which are the most important attributes to evaluate?

This question seeks to identify which attributes of the artefact should be evaluated. There are a number of quality models in the software engineering literature, and they define several quality characteristics (also known as attributes or factors) of software artefacts.

Among them, one of the earliest models introduced by McCall et al. [247] addresses different quality attributes in three categories: a) product revision that consists of maintainability, flexibility, and testability; b) product operations that consist of correctness, reliability, efficiency, integrity, and usability; and c) product transition that consists of portability, reusability, and interoperability attributes. Boehm et al. [49] introduced a software quality model in 1978 to automatically and quantitatively evaluate the quality of software. The model is hierarchical and consists of a) high-level or primary characteristics, b) intermediate-level, and c) lower-level or primitive characteristics. The high-level attributes include utility, maintainability, and portability. There are seven intermediate-level factors: portability, reliability, efficiency, usability, testability, understandability, and flexibility. The primitive characteristics have been proposed to provide the foundation for defining quality metrics (e.g., accuracy, completeness, and accountability).

Dromey et al. [93] proposed a product-based quality model that has attributes in four categories: a) correctness category that includes functionality, and reliability; b) internal category consisting of maintainability, efficiency, and reliability; c) contextual category consisting of maintainability, reusability, portability, and reliability; d) descriptive category that includes maintainability, efficiency, reliability, and usability attributes. The ISO 9126 model [170] was first introduced in 1991 called Software Product Evaluation - Quality Characteristics and Guidelines for Their Use (ISO 9126). From 2001 to 2004, the ISO published an expanded version containing both the ISO quality models and inventories of proposed measures for these models. This standard categorizes software quality into a) internal and external quality category that consists of functionality, reliability, usability, efficiency, maintainability, and portability; and b) quality in use category that consists of effectiveness, productivity, safety, and satisfaction attributes (ISO9126, 2001). Grady et al. [344] introduced the FURPS model in 1992, with five characteristics: functionality, usability, reliability, performance, and supportability. Each of these attributes has further sub-categories or sub-components as follows.

- Functionality: Feature set, Capabilities, Generality, Security
- Usability: Human factors, Aesthetics, Consistency in the user interface, Online and context-sensitive help, User documentation, Training materials
- Reliability: Frequency and severity of failure, Recoverability, Predictability, Accuracy, and Mean time between failures
- Performance: Speed, Efficiency, Availability, Accuracy, Throughput, Response time, Recovery time, and Resource usage
- Supportability: Testability, Extensibility, Adaptability, Maintainability, Compatibility, Configurability, Serviceability, Installability, and Localizability

The different quality attributes discussed in the McCall [247], Boehm [49], Dromey [93], ISO 9126 [170] and FURPS [344] models are summarized in Table 9.2.

**Table 9.2** Quality attributes of multiple quality models

Attribute	McCall	Boehm	Dromey	FURPS	ISO 9126
Maintainability	✓		✓		✓
Flexibility	✓				
Testability	✓	✓			
Correctness	✓				
Efficiency	✓	✓	✓		✓
Reliability	✓	✓	✓	✓	✓
Integrity	✓				
Usability	✓		✓	✓	✓
Portability	✓	✓	✓		✓
Reusability	✓		✓		
Interoperability	✓				
Understandability	✓	✓			
Functionality	✓		✓	✓	✓
Performance	✓			✓	
Supportability	✓			✓	

A number of researchers have reviewed these quality models (for example, [315, 39, 252, 358]). These reviews suggest that the selection of different quality attributes depends on the artefact and its functionality. Therefore, the developers can select quality models and attributes based on their specific requirements, capabilities, environment, and the functional aspects of their artefact [333, 286].

The artefact developed during this study was a software prototype, and thus, “functionality” and “usability” attributes, as discussed by McCall, Dromey, FURPS and ISO 9126 models, were chosen to evaluate during the first iteration because the artefact was in its very early stage of design. However, “reliability” and “performance” attributes were chosen to be evaluated during the second evaluation round after considering the feedback received during the first round of evaluation. Other attributes such as dependability and maintainability were not relevant to evaluate as the artefact was not fully developed. The aspects considered under the FURPS model were useful to comprehensively test the various features, scenarios, capabilities and results generated by the artefact. For example, the FURPS model provides sub-categories of “functionality”, and “usability”, “reliability” and “performance” attributes as follows:

- Functionality aspects:
  - \* main features for the intended usage
  - \* generalizability for other scenarios
  - \* support to the decision-makers for making better decisions
- Usability aspects:
  - \* the simplicity of use
  - \* clear and easy to understand the interface screens

- \* comfort and acceptability of usage
- Reliability aspects:
  - \* accuracy of the event templates generated;
  - \* feedback when the user makes an error;
  - \* possibility of recovering from a user error;
- Performance aspects:
  - \* response time of user interfaces;
  - \* the speed of event generation;

The second sub-activity of Select Evaluation Goals is to identify appropriate research strategies and methods for evaluation. Venable et al. [376] provide guidelines for selecting evaluation strategies and methods (see Table 9.3). Having chosen *ex ante* and naturalistic evaluation, Venable et al. suggest action research, focus groups and interviews as suitable research strategies and methods. Among these strategies and methods, interviews are a powerful instrument for acquiring stakeholder opinions and perceptions regarding the usage and value of an artefact [175]. Interviews also allow researchers to delve further into the perspectives of stakeholders by asking follow-up questions as needed. Moreover, interviews allow the researcher to collect complex, deep, unique, and sensitive data while establishing contacts for future discussions. Considering these advantages, interviewing was selected as the best method for evaluating the software prototype.

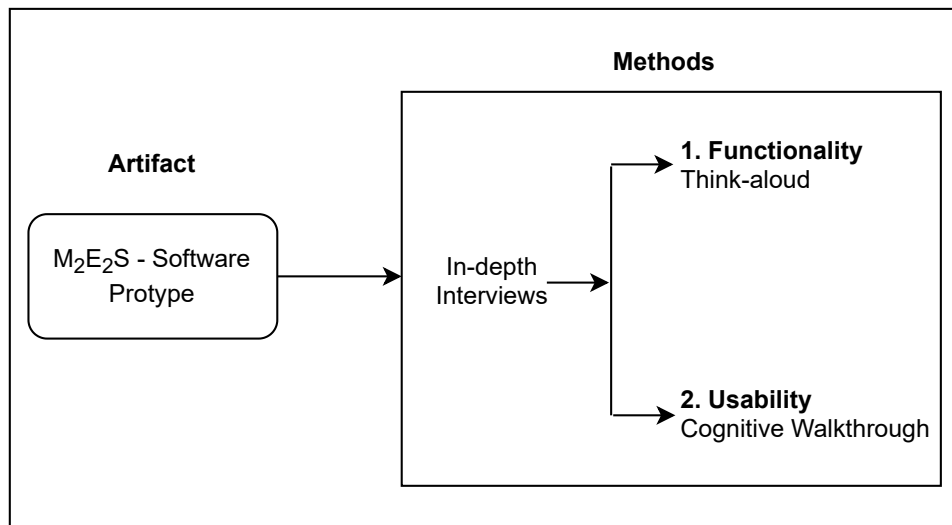
**Table 9.3** Research strategies and methods for different evaluation strategies adapted from Venable et al. [376]

	Ex ante	Ex post
Naturalistic	Action research	Action research
	Focus group	Case study
	Interview	Ethnography
		Phenomenology
		Survey
		Focus group
		Participant observation
Artificial	Mathematical or logical proof	Mathematical or logical proof
	Computer simulation	Computer simulation
	Lab experiment	Role-playing simulation
	Informed argument	Lab experiment
		Field experiment

Apart from guidelines provided by Venable et al. [376], there are other methods available in the literature that support interviews for evaluating software artefacts. For example, Wright et al. [404] suggest that artefact evaluation using the think-aloud method is an effective testing method that provides significant gains for the evaluators. Moreover, Rieman et al. [325] describe cognitive walkthrough as a method that helps designers take on a potential



user's perspective and identify some of the problems that might arise in interactions with the system. Having considered the capabilities of these evaluation methods, the researcher used in-depth interviews combined with think-aloud and cognitive walkthrough methods as illustrated in Figure 9.4 to obtain valuable information from the responders.



**Figure 9.4** Evaluation process used in this research study

## 9.4 Design and Carry Out Evaluation

In this activity, researchers design the evaluation in detail by designing interview questions and booking responders for interviews [175]. According to the formative evaluation, two interview rounds were designed. An early prototype was presented for evaluation in round 1, and an improved version that had been modified in response to feedback from round 1 was presented for evaluation in round 2. Therefore, questionnaires were developed to capture the evaluation details from the experts. The questions developed for Round 1 interviews are available in Appendix D, and Round 2 questions are available in Appendix E. The evaluation interviews had to be done using online methods due to COVID-19 travel restrictions. Therefore, zoom, a popular video conferencing platform, was used to conduct the interviews. Interviewees were asked to select a convenient date and time. Section 9.4.1 details round 1 interviews while Section 9.4.3 details round 2 interviews.

### 9.4.1 Round 1 Evaluations

Figure 9.5 illustrates interfaces of the software artefact evaluated during the round 1 interviews. The interviews started with preliminary questions to understand the participants' background and expertise in using Information Systems. Then, a combination of think-aloud and cognitive walkthrough techniques was used for the evaluations and are described in the next section. These two techniques were chosen as they provide robust and flexible forms to carry out a complete spectrum evaluation. Finally, a few questions were asked to understand how satisfied they are with capturing their requirements in the prototype system.

- Think-aloud evaluation

Before conducting the interviews, we produced a short user guide and requested interviewees to come for the interview after reading it; thus, the training was limited to reading the short



Figure 9.5 Software artefact evaluated during Round 1 interviews

user guide. This user guide is available in Appendix D. We adopted the think-aloud method used by Wright et al. [404] in their study which was called “co-operative evaluation”. The candidates were told to think of themselves as co-evaluators of the system. They were occasionally asked questions such as “*What will the system do if...?, Why did you want to do that?*”. When a user asked questions such as what to do next, the interviewer asked more questions to determine the user’s understanding of the available operations, their interpretation of the screen, and so on. The candidates were asked to note any problems they identified and their suggestions for improvement.

- Cognitive walkthrough

Rieman et al. [325] suggest that the usability evaluation should start early in the design process, optimally in the early prototyping stages. It is a practical evaluation approach based on Lewis and Poison’s CE+ theory of exploratory learning. The CE+ theory is a human cognition information-processing model that describes human-computer interaction in four steps [215, 304, 400].

1. The user sets a goal to be accomplished with the system (e.g., “check the spelling of this document”).
2. The user searches the interface for available actions such as menu items, buttons, and command-line inputs.
3. The user selects the action that seems likely to progress toward the goal.
4. The user performs the selected action and evaluates the system’s feedback to prove progress toward the current goal.

The four steps were adapted while conducting the walkthrough with the interview candidates. The participants were given a set of goals to begin the walkthrough. However, since the interviews were conducted online, the researcher demonstrated each action and asked participants to evaluate the system’s feedback. The candidates were requested to note the problems and suggestions for improvement. The interviews were transcribed and analyzed individually. Thematic analysis was used to analyze the transcripts of interviews. Section 3.4.2 of Chapter 2 describes the process of the thematic analysis process. According to thematic analysis, the transcripts were classified as a problem or a suggestion and then into sub-categories. The following sections discuss the results of the analysis.

### 9.4.2 Findings of the Round 1 interviews

Table 9.4 summarises the results of the preliminary interview analysis.

**Table 9.4** Results of the preliminary interview analysis

Participant	Question	Description
P1	Daily duties	Project Management, Innovation
	Current ISS	(Systems Applications and Products (SAP)), Office365
	Real-time systems	Smart View (an in-house system), pedestrian counting from CCTV, Earthquake monitoring system
	Social media usage	Yes (communication)

	Previous experience	Project Management
	Formal Training	Yes
P2	Daily duties	Project Management, Technology
	Current ISs	Traffic counting system from sensor data
	Real-time systems	CCTV monitoring system
	Social media usage	Yes (pushing information for the public)
	Previous experience	Emergency call centre
	Formal Training	No (On the job training)
P3	Daily duties	Handle transport operations, Technology
	Current ISs	Edge device monitoring system, Automatic traffic management system
	Real-time systems	CCTV monitoring system, Incident detection camera system, Traffic signal control system
	Social media usage	No
	Previous experience	Intelligent Transportation Systems (ITS)
	Formal Training	Yes (Geographic Information System (GIS))
P4	Daily duties	Optimization and design of intersections
	Current ISs	Office365, Traffic modeling systems, Communication systems
	Real-time systems	CCTV monitoring, Traffic signal control system
	Social media usage	Yes (communication)
	Previous experience	Telecommunication
	Formal Training	No (On the job training)
P5	Daily duties	Development, Training Neural Networks, Technology
	Current ISs	Databases, Analytic systems
	Real-time systems	CCTV monitoring system
	Social media usage	No
	Previous experience	Development
	Formal Training	No
P6	Daily duties	Project Management
	Current ISs	Data Management systems, SAP
	Real-time systems	CCTV monitoring system
	Social media usage	Yes (communication)
	Previous experience	Program Management
	Formal Training	Yes (SAP)
P7	Daily duties	Operations planning, Coordination of activities, Incident management
	Current ISs	Traffic counting system from sensor data
	Real-time systems	CCTV monitoring system, detour allocation system, Traffic Watch system
	Social media usage	Yes (pushing notifications, analyzing the context after incidents)
	Previous experience	Incident monitoring
	Formal Training	No (On the job training)

According to Table 9.4, three of the participants handle daily activities related to project management, two of them work on handling and planning transport operations, and one does software development. SAP, Office365, and a traffic counting system based on sensor data are the most popular offline tools, as they are used by at least two of the participants. Among the other offline software, they use multiple internal systems such as edge device monitoring, automatic traffic management, traffic modeling, communications, databases/ data management, and analytic tools.

All participants use real-time systems, and CCTV monitoring systems are the most common among them, with all participants using them. Two of the participants use traffic signal control systems. Other real-time tools such as earthquake monitoring, smart view, traffic watch, incident detection from cameras, and detour allocation systems are also used. Four participants use social media to perform their daily job tasks mainly for communication purposes, such as push notifications for the public. However, three participants are not using any social media applications on a daily basis.

The participants came from seven backgrounds: project management, emergency call centers, ITS, telecommunication, development, program management, and incident monitoring. Four of the participants received no formal training in information systems and depended only on on-the-job training. However, two participants received formal SAP training and one on GIS.

The thematic analysis process of Round 1 interviews resulted in a total of 17 problems and 20 suggestions. On average each individual participant identified 6.3 of the problems and suggestions. Tables 9.5 and 9.6 show the problems identified by the experts and the suggestions given for the improvement of the artefact.

**Table 9.5** Problems of the artefact identified from the cognitive walkthrough

<b>Module/Goal</b>	<b>Problem</b>	<b>Comments made by the participant</b>
Real-time Traffic Counting Footage vs Traffic Flow Prediction	Flow from CCTV Not clear what the numbers are	“it’s not quite clear to me what exactly I’m looking at [...] what exactly those numbers are?”
	Can not see the time for traffic counts	“It is not clear if you are identifying for all of the time or particular time windows?”
	Not clear what is meant by bike	“bike part is a bit misleading for me in the sense that I thought it was cyclists and not motorbikes”
Real-time Traffic Counting Footage vs Traffic Flow Prediction	Information presented in not clear	“I think it just needs a little bit more tidying up”
	Confusing traffic flow visualization	“If the incident happened in an intersection, I don’t understand if that’s particularly talking about traffic from which road”

		Not clear how external influences to traffic is handled	“things like school holidays, road-works, university holidays and then[...]if there’s a big sporting event or a concert, then your model doesn’t know about that. So it could get lots of false-positive notifications”
		Do not show smooth predictions	“they look not very smooth, often real data is not smooth, but I would have expected predictions to be more smooth”
		A predictive model can not be accurate	“it’s going to be very difficult to get a reasonably accurate predictive model for traffic conditions as they are dynamic”
Multi-Source modal Event Extraction for Disaster Response	Multi-	Small incidents are not picked	“I don’t know how useful it would be in real-time[...]so small incidents may not be picked up by news or tweets, but may still have a significant impact on the network[...]if people aren’t tweeting, or there’s no news on it may not actually get picked up” “I don’t see it, having its greatest benefit, during short duration events”
		Not clear how you differentiate news tweets	“A lot of the time, news articles seem to be tweeted out as well. So I wonder how other tweets just going to be a repeat of the link being shared of the news article, or will it be people actually providing further context or insight”
		Difficult to integrate call-centre data	“call centre operators don’t always know exactly what’s happening on the network, and therefore they look at our internal systems[...]but not every crash event is actually logged[...]lot of the time, we don’t get the first response phone calls the police or the emergency services do. so I think it would be quite complex to try and integrate information there”
		Twitter is not the primary social media channel in New Zealand	“in the New Zealand context, it’s a little bit different to what you might see internationally. Twitter isn’t always used as the primary social media tool to share information”

	Not reliable	“specially in the South Island, where it’s quite remote, unfortunately, sometimes there isn’t cell coverage, so no one will be able to be updating social media” “I guess is that it would be very difficult to make this. Reliable”
Geo-location	The map is hard to understand & use Not clear what is meant by blue dot in the map	“it doesn’t tell me anything” “But actually, what’s that blue dot? because if you’ve got multiple accidents on the same street”
Entire system	Not clear how the system can be integrated to existing systems  User Interface (UI) is not friendly from the operational perspective	“can this system be integrated into existing systems?, they’re not interested in having any more systems[...]or at least if a system gets upgraded to something else gets removed, there’s so many systems” “there’s probably some UI stuff that would need to be done for this to be more user-friendly for an operational environment”

**Table 9.6** Suggestions received for the improvement of artefact through cognitive walkthrough

Module/Goal	Suggestion	Comments made by the user
Real-time Traffic Flow Counting from CCTV Footage	Include other modes of transportation	“I would say it will be more useful if we get the other forms of transport as well”  “ I guess if your solution could also detect pedestrians and cyclists, and those using micro-mobility devices” “given the high number of cyclists, it would be good to see if that could be incorporated in the future as well”
	Count vehicles on street parking	“counting the number of vehicles on the street park would help to know at what times the parking is full and vice versa”
	Identify the condition of road surface	“So actually understanding that the surface of the road conditions would be beneficial [...] it requires on the operators paying attention to where the weather patterns are”
	Include pedestrian counting	“it might be good to know pedestrian counts as well”  “it would be really interesting to understand as well the pedestrian element[...]specially in like high school pedestrian zones ”
	Allow behavioural inspection	“also incorporate behavioural observations like wrong-way driving, illegal parking and stuff like that”
Real-time Traffic Flow vs Traffic Flow Prediction	Manual interventions are not considered	“interventions such as changing traffic signals or putting up signage where we have signs that could affect the prediction, currently not considered“
	Include error bars	“ I’d like to see the prediction with sort of error bars so that you could really see when actual data deviates outside of that”
	Include real-time weather data	“It would be really good to represent live met service data as well”
Multi-Source modal Event Extraction for Disaster Response	Include Facebook data	“How about Facebook data?”



Geo-location	Improve the map	“geolocate the incident on a map, once I click that, it actually provides information [...] a pop up that would say, what you have in component three”
	Allow the operator to manually update the location	“sometimes, the geolocation auto-populates might only be able to narrow down to a city[...]I think having the operator be able to manually update the location would be beneficial”
	Identify secondary locations	“So if a detour is needed because the road has been closed, then actually be able to understand where the closure points are on the map[...]sometimes it could be 10 kms away”
	Identify locations of interest in the impacted area	“there can be schools, other services that may have restricted access or may be impacted by an event. Also, it could be things like petrol stations, where they can actually help to get the information out about an event”
Entire system	Have a single system instead of four components	“Information can be presented more cohesively. I want to see it, in one go. It could be that in that particular map, you could click it and then say, what particular road or area, If you right-click it, there is an option to real-time and more information”
	Integrate other sources of data (e.g., earthquake monitoring sensor data)	“in the initial version start with a map, a dashboard view will be ideal” “in Christchurch, we have earthquake monitoring sensors [...] so I’m assuming that if you can get data from those sensors, they can overlay. It’s the same with flooding sensor data.”
	A reporting tool	“you can look at historical information and then correlate them..this allows predictive planning as well as real-time planning [...] mitigating risk when the events happen.”

	“A lot of the time, after an incident then we’ll get a debrief or lessons learned. And I think that’s where this tool will help to collate information around a debrief[...]will help people preparing for those debrief documents and information”
Consider the holistic view of the road network	“rural and urban context, understanding how interventions would actually affect the predicted traffic”
Able to manually add data	“ I want to manually be able to add any additional news articles or anything I found”
A way to tackle if we have answered the questions in social media	“Have we actually covered off what some of the commentary and the questions the people asking?”
Include the confidence of the information	“give a confidence value is probably quite important in helping operators make the right decisions”
Visualize high-priority events	“it will be good to identify high-priority events”

Three sub-categories of the problems/suggestions were identified during the thematic analysis process such as operational, functional, and visual/cognitive. The mapping of these problems and suggestions to sub-categories of FURPS model and the sub-categories of the thematic analysis process is listed in the Table.

**Table 9.7** Mapping of the problems and suggestions to sub-categories of FURPS model and the sub-categories of the thematic analysis process

<b>Problem/Suggestion</b>	<b>Sub-category of FURPS model</b>	<b>Sub-category of thematic analysis</b>
Not clear what the numbers are	clear and easy to understand the interface screens	visual
Can not see the time of traffic counts	clear and easy to understand the interface screens	visual
Not clear what is meant by bike	clear and easy to understand the interface screens	visual
Information presented is not clear	clear and easy to understand the interface screens	visual
Confusing traffic flow visualization	clear and easy to understand the interface screens	visual

Not clear how external influences to traffic is handled	generalizability for other scenarios	operational
Do not show smooth predictions	main features for the intended usage	visual
A predictive model can not be accurate	main features for the intended usage	functional
Small incidents are not picked	main features for the intended usage	functional
Not clear how you differentiate news tweets	main features for the intended usage	functional
Difficult to integrate call-centre data	comfort and acceptability of usage	operational
Twitter is not the primary social media channel in New Zealand	comfort and acceptability of usage	functional
Not reliable	comfort and acceptability of usage	functional
The map is hard to understand & use	the simplicity of use	visual
Not clear what is meant by blue dot in the map	clear and easy to understand the interface screens	visual
Not clear how the system can be integrated to existing systems	support to the decision-makers for making better decisions	operational
User Interface (UI) is not friendly from the operational perspective	comfort and acceptability of usage	operational
Include other modes of transportation	generalizability for other scenarios	functional
Count vehicles on street parking	generalizability for other scenarios	functional
Identify the condition of road surface	generalizability for other scenarios	functional
Include pedestrian counting	generalizability for other scenarios	functional
Allow behavioural inspection	generalizability for other scenarios	functional
Manual interventions are not considered	support to the decision-makers for making better decisions	operational
Include error bars	support to the decision-makers for making better decisions	visual
Include real-time weather data	support to the decision-makers for making better decisions	functional

Include Facebook data	support to the decision-makers for making better decisions	functional
Improve the map	the simplicity of use	visual
Allow the operator to manually update the location	support to the decision-makers for making better decisions	functional
Identify secondary locations	generalizability for other scenarios	operational
Identify locations of interest in the impacted area	generalizability for other scenarios	operational
Have a single system instead of four components	the simplicity of use	visual
Integrate other sources of data (e.g., earthquake monitoring sensor data)	support to the decision-makers for making better decisions	functional
A reporting tool	support to the decision-makers for making better decisions	operational
Consider the holistic view of the road network	support to the decision-makers for making better decisions	operational
Able to manually add data	support to the decision-makers for making better decisions	functional
A way to tackle if we have answered the questions in social media	support to the decision-makers for making better decisions	functional
Include the confidence of the information	support to the decision-makers for making better decisions	functional
Visualize high-priority events	support to the decision-makers for making better decisions	visual

The next sections describe the problems and suggestions identified by the participants and sub-categories such as “Operational”, “Functional” and “Visual/cognitive”.

- Operational

All the participants recruited for the interviews are actively working in the traffic, transport or urban planning domain. Therefore, they could identify problems from an operational perspective. For example, they had concerns regarding tackling external influences such as holidays, roadworks, and major events by the system. Moreover, they identified multiple issues in integrating call-centre data. Firstly, police or emergency services first record call centre data and it is very rare that NZTA/TOCs get them. Secondly, call-centre data are not always logged into any of their internal systems. One of the other main concerns was how the system could be integrated into existing ISs. NZTA has multiple systems/tools to handle its day-to-day operations, such as traffic count, CCTV monitoring, and traffic signal control systems. Therefore, as the operators can not use multiple systems, any new tool is expected to be integrated into the existing systems to have a smooth workflow. Moreover, they wanted

to have a consistent UI for the operators, similar to the UIs they handle daily. Apart from that, they highlighted that manual interventions such as changing traffic signals and putting up signage boards are not considered by the system.

From an operational perspective, they gave some suggestions to improve the system. One is identifying secondary locations where an event can impact the decision-making process (e.g., identifying road closure points). Moreover, identifying and plotting points of interest such as schools and petrol stations nearby would help identify restricted access locations and establish communication channels during an event. Also, they highlighted the need for a reporting tool for correlation, debriefing, or lessons learnt purposes.

- Functional

The participants identified several functional issues during the demonstration. Because traffic networks are dynamic, one of the key problems was the accuracy of a prediction model. Furthermore, they were unsure if the system could identify small-scale events that are not often tweeted or reported in the news. Moreover, since the news is mostly tweeted, they had issues with how we differentiate them from the tweets generated by the general public. Also, they pointed out that Twitter is not the primary social media channel in New Zealand, and people might be putting more content on Facebook than on Twitter. Finally, due to many issues in social media content, such as fake news, rumours and misinformation, they identified that making the proposed system reliable is very difficult.

Interviewees identified multiple suggestions for improvement of the functionality of the system. For example, they pointed out that including other modes of transportation such as scooters, cyclists, and counting pedestrians would be beneficial. Moreover, counting the number of vehicles in the street parking, identifying the road surface condition, and human behavioural inspection (e.g., wrong-way driving and illegal parking) from CCTV footage would provide an operator with additional details. Furthermore, incorporating live weather data would further detail the responder for his decision-making task. They also suggested including other data sources such as earthquake and flood sensor data, live bus tracking data and Facebook data. The ability of an operator to manually add data to the system was also suggested. Furthermore, it was highlighted that the system could allow the responder to handle questions and concerns on social media channels. Finally, the importance of incorporating confidence levels in the information provided by the system was emphasised.

- Visual/cognitive

Spivey et al.[351] identify that designing information formats so that the human mind can process the content more effectively is a key principle of system design. Language, memory, perception, learning, and attention are some examples of cognitive skills.

This study found that most users had issues regarding the presentation of information, such as not being clear about the numbers on screens, not seeing the time for traffic counts, and being unclear about what is meant by bike in traffic count component. Also, they did not like presenting traffic flow in a simple bar graph as it was difficult to understand the deviation. Furthermore, they were concerned about the geolocation in the map, as it currently just shows the location as a pinpoint and does not show any information such as the event details and related visuals.

The participants gave multiple suggestions to improve the user interface. Among them, the most important one is to have a single system instead of four components. In addition, they suggested showing the map view and then allowing drill-down options for the responder to select and see more information as needed. Moreover, they wanted to see error bars in the traffic prediction chart and visualize high-priority events.

As described, a few questions were asked to understand how satisfied the participants were with capturing their requirements in the prototype system. Following are some of the responses received.

“...I am really interested that you have managed to get that to work really well. So that is something that we have been trying for years to get to work, and many large consultancy firms like Baker Jacobs GHD have been trying for years to get to work...”

“...This is excellent, and it would be interesting to test it in a real-world environment, like in New Plymouth or a small city like that, to see how well it actually does...”

“...That’s quite impressive, you’ve done well to get it to where it is now, in comparison to some of the other products I have seen. And indeed, the flow prediction piece as well. I know that that is a nightmare in itself. It’s looking really good...”

Based on the in-depth interview analysis and the comments, it is possible to infer that the system performed in accordance with the users’ expectations. The following is a summary of the main findings of the interview analysis process:

- Bringing together information from multiple sources in real-time during disasters is very helpful for responders. The software prototype demonstrated all the functionality of such a system which has been needed for a long time.
- Overall, the user interface was easy to understand. However, some screens need help guides for required inputs from the user.
- A few improvements are needed to integrate different components in the current prototype into a single system.
- Special attention must be paid to ensure that the final system can be integrated with the existing in-house systems.

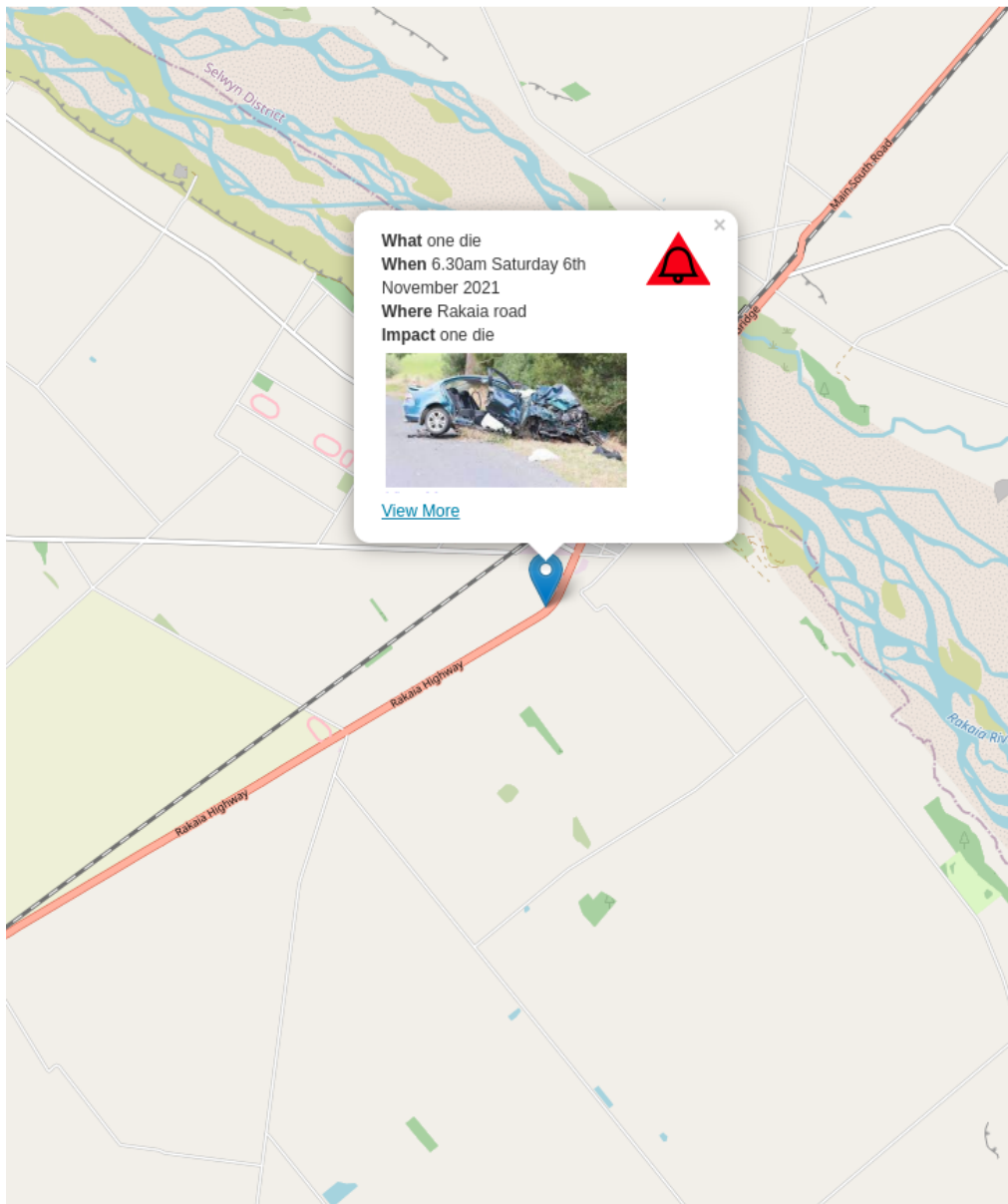
### 9.4.3 Round 2 Evaluations

In Section 9.4.2, 37 problems and suggestions for improvement were identified during Round 1 interviews. Formative assessment requires that the next round of evaluations be conducted after addressing the issues identified in earlier rounds [175]. Therefore, all 37 problems and suggestions were carefully considered and prioritized to be incorporated into the software prototype’s next version. Problems/ Suggestions such as “include other modes of transportation”, “count vehicles on street parking”, “identify conditions of road surface”, “pedestrian counting”, and “behavioral inspections of road users” were not listed in the initial requirements and required an extensive

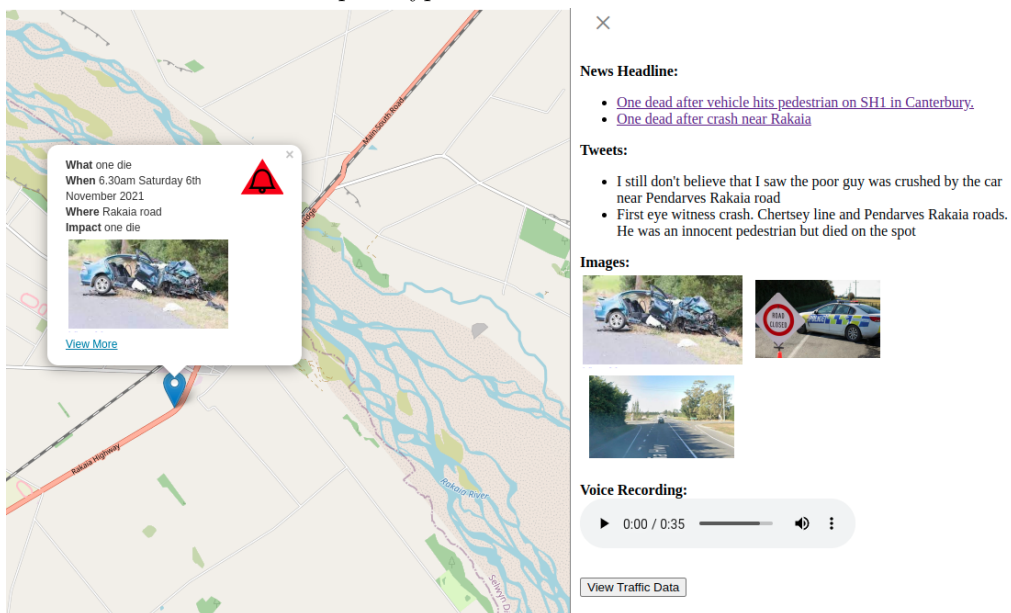
development time. Therefore, they were considered out of the scope of within the time frame of the doctoral research. Moreover, due to technical difficulties like the lack of a real-time weather API and Facebook data being private, problems/suggestions like “add real-time weather data” and “include Facebook data” were determined to be impossible to implement. Finally, problems/suggestions such as “integrating with existing systems” and a “reporting tool” should be considered in the development of the final system. Therefore, the following problems/suggestions were identified and addressed in the software prototype for the second round of evaluations:

- The map is hard to understand & use
- Not clear what is meant by the blue dot on the map
- User Interface (UI) is not friendly from the operational perspective
- Improve the map
- Have a single system instead of four components
- Able to manually add data
- Visualize high-priority events

Some interfaces of the final software prototype are illustrated in Figures 9.6 and 9.7. Furthermore, a live demonstration of this is available at <https://rangikanilani.github.io/events.html>. During the evaluation interviews, the Software Artefact was demonstrated, highlighting functionality, usability, reliability, and performance attributes in the FURPS model (Round 2 questions are available in Appendix E.). The Experts provided feedback using a five-point scale of Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree. The results were then calculated using a 5-point standard Likert scale, where Strongly Disagree = 1, Disagree = 2, Neutral = 3, Agree = 4, and Strongly Agree = 5.



(a) Initial screen of the software prototype



(b) Second screen of the software prototype

Figure 9.6 Initial screen and view more screen



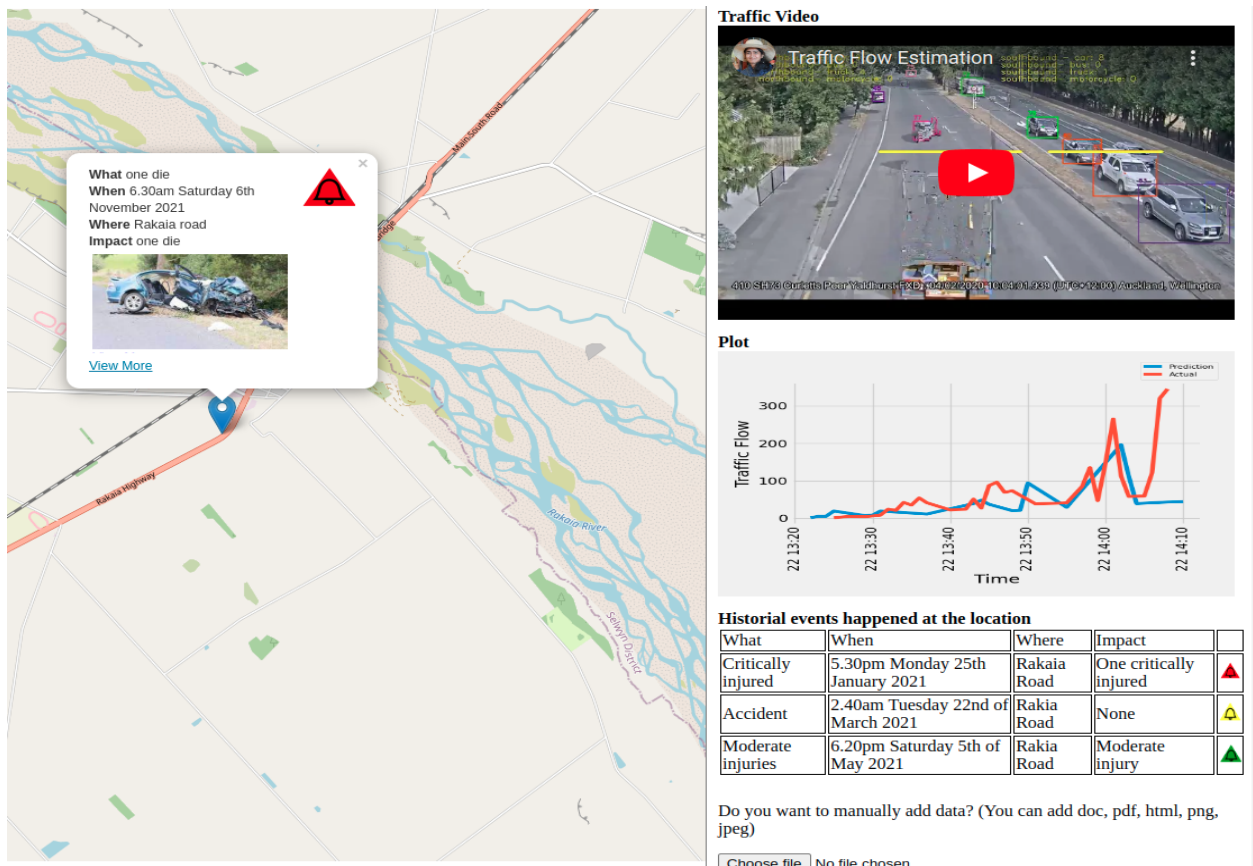


Figure 9.7 Third screen of the software prototype

#### 9.4.4 Findings of the Round 2 interviews

During Round 2 interviews, functionality, usability, reliability, and performance attributes of the FURPS model were evaluated, and the findings are discussed in the following sections.

##### 1—Evaluation of the Functionality attribute

The FURPS model states that functional requirements determine the features and capabilities of a software system [344]. Therefore, the goal of analyzing the functional features of the artefact during this research was to see how well it supports the responders in utilizing SM during daily job tasks and making decisions. The aspects evaluated under the functionality attribute are as follows:

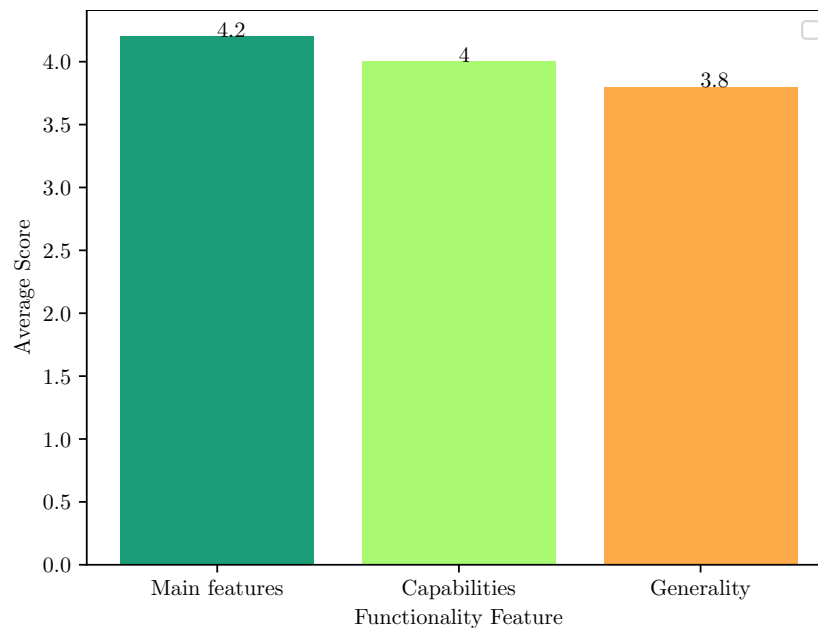
- Main features for the intended usage;
- Functional capabilities;
- Generalizability for other scenarios;
- Support to the responders for making better decisions;

Average scores of the responses for the questions obtained from the evaluators are illustrated in Figure 9.8. Main features and capabilities have received a higher score from the evaluators indicating that the software prototype could demonstrate the features required by responders for their decision-making. The response from one of the evaluators during the interview was:

“...It is great to see that you have included most of the functionalities in the system that we discussed earlier...”

However, the generality aspect has received an average score of 3.8. This is because the evaluators could not see a real-world scenario of generalizing the prototype for other events except for traffic emergencies and thus marked most of the responses as “Neutral”. An evaluator discussed the generality feature by saying:

“...I am going with neutral for this...You know, the fact is that I don’t see how we can actually get data during other scenarios in this demonstration...”

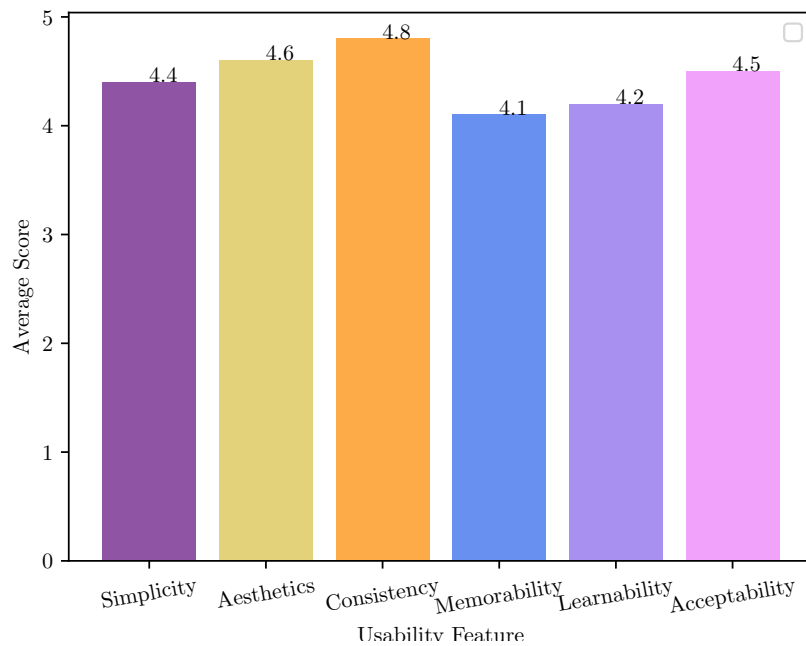


**Figure 9.8** Results of Functionality Evaluation

## 2—Evaluation of the Usability attribute

Rawashdeh et al. [318] define usability as the “capability of a software product to be understood, learned, used and attractive to the user when used under specified conditions”. Usability attribute in the FURPS model deals with aspects such as human factors, aesthetics, consistency in the user interface, and context-sensitive help [344]. The following usability aspects were evaluated during this research, and the results are shown in Figure 9.9.

- The simplicity of use;
- Clear and easy-to-understand screens;
- Consistency of information presented on screens;
- Comfort and acceptability of usage;
- Learnability for new users;
- Avoidance of information overload;



**Figure 9.9** Results of Usability Evaluation

As illustrated in Figure 9.9 all usability factors have achieved a score higher than 4.0. This indicates that evaluators are generally satisfied with the usability aspects provided by the software artefact. Evaluators commented on the usability features as follows:

“...I don’t have any comments regarding usability aspects... Everything is clear, well presented, easy to understand, easy to follow...”

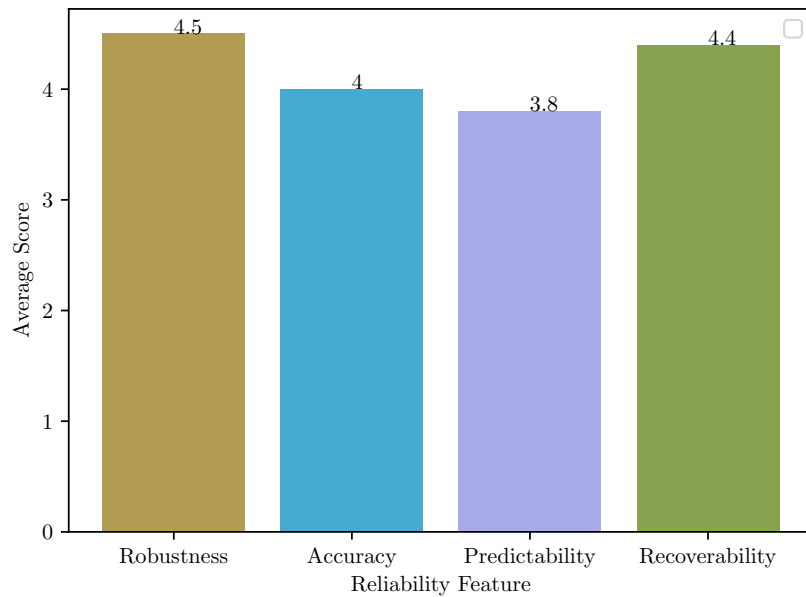
“...I don’t think I have any comments regarding the usability... It’s all good...”

### 3—Evaluation of the Reliability attribute

Reliability in the FURPS model includes factors such as frequency and severity of failure, recoverability, predictability, and accuracy [344]. Therefore, the reliability attribute determines the ability of a software artefact to maintain a specified level of performance under different conditions [315]. During this research, the following reliability factors of the software artefact were evaluated, and the results are illustrated in Figure 9.10.

- Minimized number of errors through the user interface;
- Accuracy of the event templates generated;
- Feedback when the user makes an error;
- Possibility of recovering from a user error;

As shown in Figure 9.10, the evaluators were overall satisfied with Robustness, Accuracy and Recoverability. The evaluators commented on errors that users can make while using the system.



**Figure 9.10** Results of Reliability Evaluation

“...I see that you don’t take many inputs from the users. That’s good. People make many mistakes if they are to enter too much stuff, especially during an emergency...”

However, the predictability factor achieved a lower score as many of the evaluators allocated “Neutral” as their feedback. One evaluator commented on predictability as follows:

“...It is very hard for me to decide this as the system is not fully developed...”

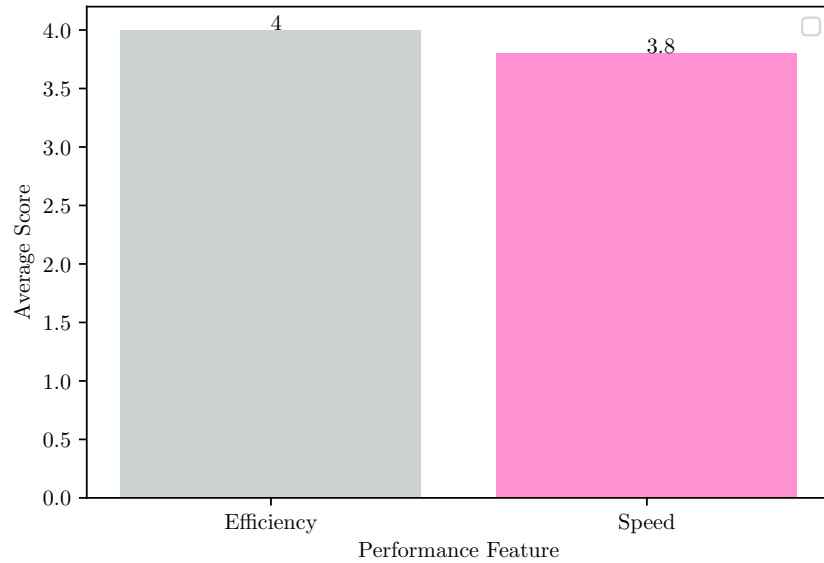
#### 4—Evaluation of the Performance attribute

Performance attribute in the FURPS model includes factors such as speed, efficiency, response time, recovery time, and resource usage [344, 315]. Among those, the following factors were considered during the evaluation of the software artefact during this research. The results of the evaluation are shown in Figure 9.11.

- Response time of user interfaces;
- The speed of event generation;

The evaluators were happy with the efficiency of the software prototype. However, it was difficult to witness the speed as the system was demonstrated not for a real-time event. Therefore, most of the evaluators marked their response as “Neutral”. The following are some of the comments made by evaluators.

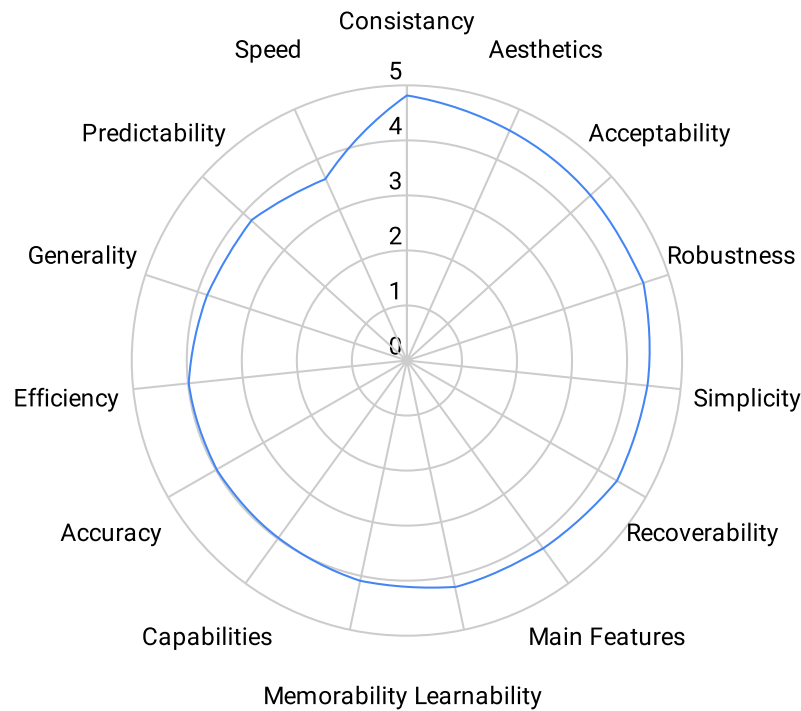
“...No, I don’t have major comments. One thing you might actually do is to put some graphics on the system to say when the next cycle of events will be available in the system...”



**Figure 9.11** Results of Performance Evaluation

“...I think the speed is quite good based on your values. But I’d rather like to see that in an actual system...”

The experts selected for the evaluation interviews represented various fields of transportation, including traffic management, project management and software development. Section 9.4.4 examined their feedback for the second time on different areas of the attributes in the FURPS model. The average scores of all these attributes were satisfactory and above the minimum acceptance level, that is, 3. Figure 9.12 summarises the fourteen factors evaluated and scores obtained. These results are the average scores of the responses of all seven evaluators, and as illustrated in Figure 9.12, there is a slight variation in the average score. This indicates that the design and functionality of the software artefact met the expectations of the experts who provided the initial requirements.



**Figure 9.12** Results of Performance Evaluation

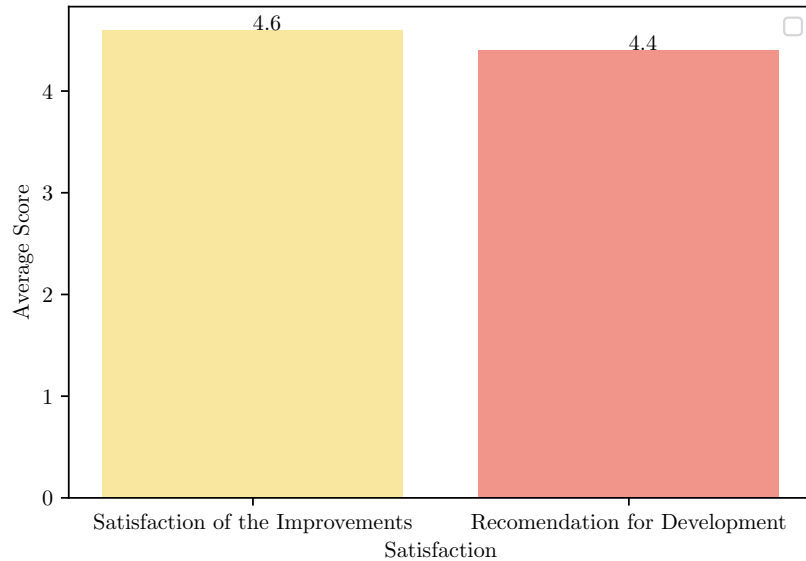
In addition to evaluating the quality attributes, there were questions to determine if the evaluators were happy with the progress and the next stage of development. The following are the results achieved. Figure 9.13 shows the scores obtained for satisfaction evaluation questions. Overall, the high scores achieved indicate that the responders were highly satisfied with the improvements. The following are some of the responses:

“...In its current state as a proof of concept, I would surely recommend exploring it further ...”

“...Very good tool, simple to use, quite intuitive and makes good use of technology. I am very impressed by the amount of work that has been done to bring it to life...”

The software artefact evaluation process led to significant findings that can be summarised as follows.

- Overall, the information presented in the artefact is clear, simple and easy to understand. A future extension of this work can be a mobile application where the responders get an alert on the go.
- There is a 30-minute time lag between each run of the system. In future, this can be improved with more research.
- Currently, the software artefact displays results in the dashboards, and it is not clear how to integrate the system with the other in-house tools. Therefore, the actual system has to provide an API where the software results can be easily integrated into other existing systems.



**Figure 9.13** Results of Satisfaction Questions

According to the findings, two iterations of interviews provided sufficient information to improve the artefact. However, the evaluation process was limited in several ways. First, due to COVID-19 travel restrictions, the evaluation interviews had to be conducted via online techniques. Therefore, the evaluators' feedback was mainly based on the demonstration of the artefact. However, they could have provided more feedback if they had been given the opportunity to work with the artefact. Second, the evaluation was limited to two iterations due to the time constraints of the PhD research. However, more iterations of evaluations would have provided more details to improve the artefact. Finally, the evaluators were selected from a few locations of TOCs in New Zealand. However, their feedback may not reflect the needs of all TOCs in the country, limiting the generalizability of the results.

## 9.5 Chapter Summary

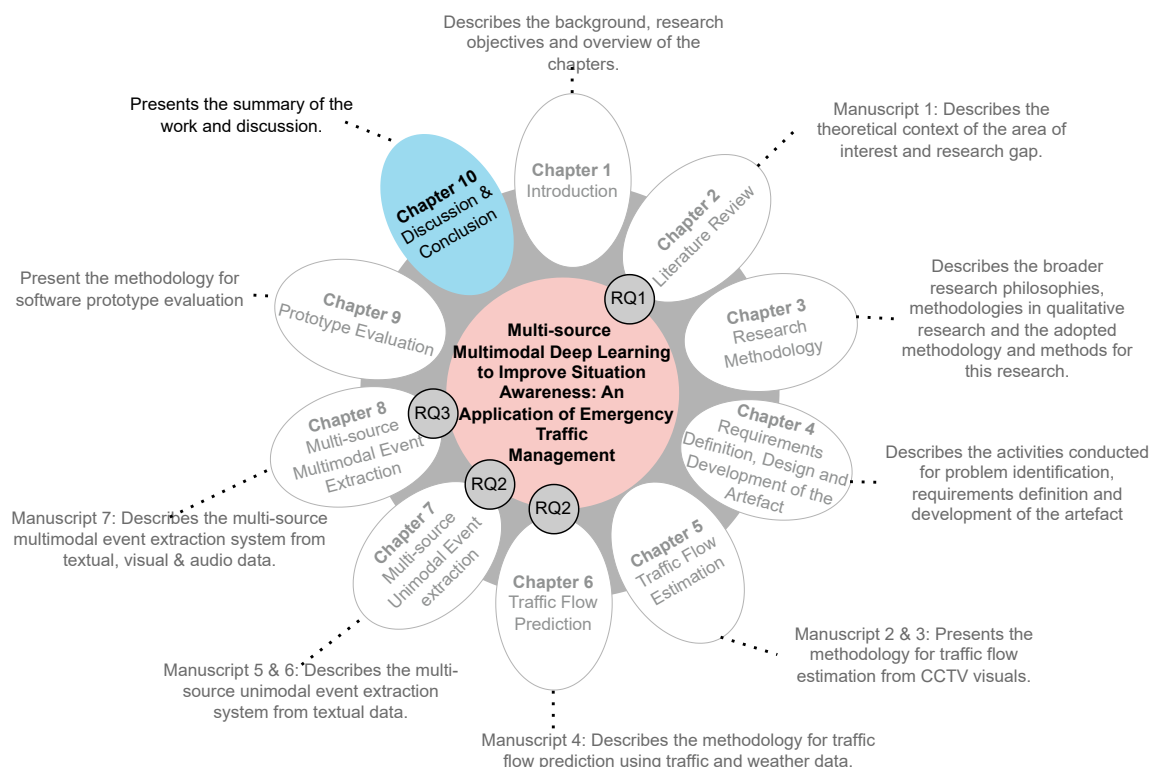
This chapter presented the evaluation of the software artefact according to the DSR. Therefore, evaluations were conducted following the method framework provided by Johannesson et al. [175], which includes three activities: "Analyse Evaluation Context", "Select Evaluation Goals and Strategy" and "Design and Carry Out Evaluation". During the "Analyse Evaluation Context" activity, key questions in evaluations in Pries-Heje's [310] framework, such as *What is being evaluated?*, *How is it being evaluated?*, *When was it evaluated?* and *Who is evaluating?* were answered. As a result, ex ante and naturalistic evaluation strategies were chosen. Seven experts from the traffic and transportation domain in New Zealand were identified as evaluators. The second activity of Johannesson's framework was to "Select Evaluation Goals and Strategy". A formative evaluation was decided, where the artefact is evaluated during the design stage to learn how to improve it during subsequent design processes. Four software quality attributes in the FURPS model were evaluated, including functionality, usability, reliability and performance [344]. Interviews were chosen as the appropriate evaluation method to collect data. The final activity in Johannesson's

method framework for artefact evaluation is “Design and Carry Out Evaluation”. During this activity, the artefact was evaluated in two consecutive rounds. Results of the first round identified 37 problems and suggestions for improvement, including hard-to-understand user interface, inability to add data manually, and visualising high-priority events. After a careful analysis, seven problems and suggestions were selected to be implemented in the design of the second version of the artefact. This artefact was presented again for the same set of evaluators for the second round of evaluations. Interviews were analyzed using a standard Likert scale of 1 to 5. Overall, the evaluation results indicate that the evaluators were highly satisfied with the improvements and recommended the development of the actual system. Some future works were also identified, which will be included in the next version of the software.



# Chapter 10

## Discussion and Conclusion



This chapter brings together the components presented in the thesis, including discussions on the motivation for the study, addressing research questions, and key outcomes, followed by a synthesis of contributions to research and practice. The chapter concludes by summarising limitations, highlighting potential extensions, and shedding light on future research directions.

### 10.1 Motivation for study

SM has become the most prominent method for rapidly disseminating information across a large community since it is free and offers simple access through smart mobile devices [10, 200]. According to research, the general public uses social media platforms to report essential information about disasters, including early warnings, missing or dead people, and damage to infrastructure such as roads, bridges, and buildings [159, 237, 85, 242]. Most importantly, multiple recent real-world events have been first reported in SM, challenging traditional media systems [301, 181]. Therefore, SM platforms provide a powerful tool for disaster responders to get quick information

from the people on disaster sites. However, the majority of first responders to an emergency or disaster are reluctant to use SM for SA due to reliability concerns, such as user-generated social media content, including rumors, false information, and misinformation [165, 196]. For instance, during the problem identification phase, interviews were conducted with emergency responders from multiple bodies, such as city councils, regional emergency management groups, and the ministry of civil defense and emergency management in New Zealand (see Chapter 4). They highlighted the following challenges related to the reliability issue:

“...Ground truth information flows in here from many ways, our public information management team monitors social media, all social media forms, for example, they see that on Facebook, someone has posted it’s got the trees down in this street. What is the reality? Is that true? How am I going to know?...”

“...Yeah, we do use social media. But it is quite unreliable. We have to see if that is a fake account or a brand new account or spam, or something like that...”

The study was motivated by the current difficulties faced by multiple emergency responding agencies in validating and utilising a huge volume of streaming SM data within seconds of a disaster. Previous studies have reported that combining multiple sources and modalities yields more information than learning from a single data modality alone [166, 25]. Furthermore, the semantic association of multi-source multimodal data allows users to gain access to a comprehensive set of knowledge [368]. In addition, data integration can help in minimizing errors and enhance the accuracy of the information provided [25]. As a result, the trustworthiness of the content provided through SM channels can be further improved [25]. Therefore, the overarching aim of this study was to validate SM content by fusing different modalities of data extracted from multiple sources to enhance the SA of disaster responders in real-time. Machine learning and deep learning approaches have been widely used in research to automate tasks in the disaster domain that were previously unattainable using solely human brain capability [290, 191, 273]. Thus, this doctoral research set out to answer three questions; (1) How have different deep learning algorithms been applied to data from various sources to support disaster response tasks? (2) How can data from multiple sources be fused to support disaster response? (3) How can the integration of multi-source multimodal data effectively support disaster response by cross-validating social media data?

## 10.2 Addressing the research questions

This section summarises how research questions were answered through the accomplishment of the research objectives that were described in Chapter 1.

### 10.2.1 Research Question 1: How have different deep learning algorithms been applied to data from various sources to support disaster response tasks?

The answer to the first research question was accomplished through the systematic literature review presented in Chapter 3. The systematic review included 83 articles, and the analysis was centered around the components of learning, a set of aspects that govern the application of machine learning for a given problem domain. Five research questions and eight sub-questions were derived

based on the components of the learning model to guide the literature analysis process.

The systematic review identified different modalities of data that have been analyzed for DR tasks. For example, image data have been mostly used for disaster damage assessment tasks, text data for disaster-related information filtering and disaster-related information classification tasks, and video data for disaster-related information filtering, classification, and disaster event detection tasks. Furthermore, the following are some of the findings relating to the most common DL models used by researchers for DR tasks:

- The CNN algorithm has been largely adopted for disaster-related image analysis
- RNNs, LSTMs, and Bi-LSTMs have been used to analyze varying length sequence data such as sentences (e.g., tweet text).
- The vast majority of research utilized pre-trained networks on larger data sets such as Places365 and ImageNet.
- Researchers have also adopted multiple DL architectures in a single research project, especially if the research involves multiple modalities of data. For example, CNN's are often used to extract image features, while RNNs, LSTMs, or BiLSTMs are used for text feature extraction.

### 10.2.2 Research Question 2: How can data from multiple sources be fused to support disaster response?

The second question set out to explore methods for fusing data from multiple sources to support disaster responders. Two key research activities were conducted to explore this.

The first project explored the fusion of weather and traffic data for traffic flow prediction. A literature review revealed that existing work has mostly focused on predicting traffic using traffic-only data [391, 422, 405, 407]. However, as traffic systems are dynamic and get affected by many conditions such as weather, planned/unplanned maintenance, school days/holidays, vehicle crashes, and disasters, a novel DL algorithm for traffic flow prediction using traffic-only data and using both traffic and weather data was developed. The experimental results indicated a higher accuracy for short-term traffic prediction combining traffic and weather data, demonstrating that integrating multi-source data contributes to increasing the prediction model's accuracy. This research is detailed in Chapter 5.

The objective of the second project was to combine multiple web sources for disaster event extraction. First, a literature review was conducted and identified that multiple studies had explored the extraction and analysis of disaster events from news web texts [161, 168, 129]. Moreover, there were systems described in the literature that extract crisis-relevant information from SM such as Tweedr [36], CrisisTracker [330], Twitcident [2] and AIDR [158], all of which processed, classified and clustered tweets for disaster response. Then, methods to effectively use multiple web text sources to assist disaster management based on their various features were explored based on the literature analysis. As a result, an algorithm to extract events, answering semantic- (*what*), spatial- (*where*), and temporal- (*when*) questions was developed having the following six following modules.

1. **News and tweet extraction:** collect online news and tweets as a scheduled task and filter using key-word based heuristics.

2. **Relevant tweet identification:** identify relevant tweets using a supervised algorithm.
3. **Noise filtering:** remove noise from text such as symbols and web links.
4. **Clustering:** group news headlines with tweets.
5. **Candidate extraction:** jointly extract words/phrases answering *what*, *when* and *where* questions from news headlines, tweets and news bodies.
6. **Candidate scoring:** select the best candidates from the possible candidate set in step 5.

Experimental results showed that the accuracy of event extraction improves significantly when additional sources are integrated. This study is discussed in Chapter 6.

### 10.2.3 How can the integration of multi-source multimodal data effectively support disaster response by cross-validating social media data?

The final question investigated how SM content can be cross-validated utilizing multi-source multimodal data. In addition, the ways of assisting disaster responders with their SA were investigated, and the details of this project are discussed in Chapter 7.

During the literature review, it was discovered that integrating multiple other data sources and modalities provides mechanisms to triangulate and validate the information while also adding more contextual information [281, 166]. Hence, the information acquired through SM channels can be further strengthened. Therefore, the methodology developed in answering research question 2 in Section 10.2.2 was extended to include multiple modalities such as images and audio. The architecture had five core modules and sixteen sub-modules to collect, preprocess, analyze, and extract event templates in real-time. Moreover, the system extracts answers to the *What (semantic)*, *Where (spatial)*, and *When (temporal)* (3W) questions, as well as impact information. Experimental results illustrated that 3W answer extraction significantly improves after integrating multi-source and multimodal data.

## 10.3 Contributions of the study

This research has made a wide range of contributions, including theoretical and practical knowledge. From the results, this thesis adds value to the research and practice communities, as it is relevant for design science and supporting SA during disaster response. These contributions are discussed in the following sections.

### 10.3.1 Contribution to research

This study makes several contributions to the existing knowledge related to using multi-source multimodal data for SA during disaster response. Making the datasets, annotations, and pertinent metadata publicly available is one of the main contributions shared by almost all the sections below. This is crucial for DL research, as many researchers have difficulty obtaining datasets. Furthermore, unlike most other disciplines, DL-based research continues to improve models benchmarking previous research. Therefore, to assess how well the recently created approach compares to earlier work, the researchers need access to previous datasets and a solid understanding of hyperparameters. The contributions to research are addressed in Sections 10.3.1.1 through 10.3.1.7.

### 10.3.1.1 Systematic literature review:

The amount of data produced during a disaster has significantly increased due to technological innovation, and first responders find it difficult to prioritize and accurately identify crucial data for their decision-making. For example, during the requirement capturing interviews, the responders said:

“...we get inputs into these phones, text messages, e-mails, both from outside and also within, we have the council teams, we have social media, we use radar. All this information comes in, and it should get filtered. Who is going to sit and do this? We don't have a sophisticated tool for this...”

“...So, in the end, to make a decision, whether to do something based on incomplete, possibly inaccurate, non-helpful information or do nothing until better information comes in. And then, how long is that going to take? How long to wait and do nothing while someone is hurt or dies?..”

DL techniques have been extensively used to learn high-level representations through deep features in many fields, including speech recognition, event detection, and multimedia retrieval [150, 388, 379]. A literature analysis suggested that multiple review articles discussed the use of DL for disaster response tasks [235, 279, 392, 335, 22, 198]. However, these reviews were especially focused on addressing a single source of data and how it can be used for disaster response. Therefore, a systematic literature review presented in Chapter 2 was conducted to explore how different data sources and modalities have been used to support disaster response activities. A roadmap, namely learning components proposed by Abu Moftha [366], was used to structure the literature analysis for the first time in the disaster domain. In contrast to the most recent work by Sun et al. [359], a more comprehensive discussion of datasets, preprocessing, DL architectures, hyperparameter tuning, challenges, and solutions in processing data for DL tasks was provided, and future research directions were clarified.

The literature review findings are published as a journal article in “SN Computer Science” [26]. Furthermore, the public can access the complete details of the analysis process, as well as the resources, through an online appendix [23].

### 10.3.1.2 Traffic flow estimation from CCTV images:

Estimating traffic flow is the first step in identifying road traffic patterns, contributing to traffic modeling, urban planning, and design processes of all aspects of a road network [104]. Traditionally, traffic flow has been estimated using inductive loops, pneumatic road tubes, and manual counts [42]. However, these methods are labor-intensive and time-consuming. For example, during interviews, one of the responders highlighted the problems using traditional methods:

“...We use inductive loops to count traffic...but the problem is we need to dig the road, and it damages the road surface, reduces the life of road. Then again, this is not continuous... contractors lay loops for a certain period...”

Recently, CCTV systems have been increasingly mounted in many public places, and thus research has found that they can be used to obtain traffic flow counts [104, 296]. However, the

difficulties in moving, storing, and developing efficient, intelligent algorithms for processing and analyzing CCTV big data have been identified as significant challenges [104]. During the interviews, one of the responders highlighted the current status of utilizing CCTV for traffic counts in New Zealand:

“...We have smart cameras at certain locations. We use them to detect debris in tunnels, stopped vehicles, crashes, and inverse vehicles. We had some agreements with companies for projects to get traffic counts from CCTV, but none of them worked perfectly well so far...”

There is limited research exploring traffic counting from CCTV data. For instance, the study by Fedorov et al. [104] used a DL architecture, Faster R-CNN, for a video data set to identify traffic flow. Moreover, a study evaluated the accuracy of two DL algorithms, MobileNet, and faster R-CNN trained on COCO data set [296] and demonstrated the accuracy of faster R-CNN is more for their CCTV image data set. Overall, these studies have identified the lack of public accessibility for CCTV data, difficulties in moving and storing CCTV data, and developing efficient, intelligent algorithms for processing and analyzing as significant challenges [104]. Furthermore, these studies have used a very limited dataset for training DL models, and the direction of vehicle movement was not considered during the flow estimation process.

New Zealand Transport Agency (NZTA) provides an Application Programming Interface (API) to access CCTV images in real-time across the country <sup>1</sup>. Taking advantage of the open access CCTV image series and considering the research gap, a novel method for estimating traffic flow was studied. As a result, an algorithm to obtain traffic flow from CCTV images was developed, considering the challenges and gaps. This research is presented in Chapter 4.

This work is published in the proceedings of the 17<sup>th</sup> International Conference on Information Systems for Crisis Response and Management (ISCRAM 2020) [28]. Furthermore, the developed algorithm and the project’s details are publicly available for future researchers for re-implementation <sup>2</sup>.

### 10.3.1.3 Traffic flow estimation from CCTV footage:

As discussed, understanding road traffic behavior is essential for developing an emergency traffic response plan [104]. Visual datasets obtained from surveillance cameras and aerial vehicles have recently been explored for many traffic monitoring applications [155, 417, 187, 6]. However, real-time traffic flow counting from CCTV footage has been less explored due to challenges in accessing, storing, and the requirement of special graphical processing capabilities [104]. It was highlighted in Section 10.3.1.2 that in New Zealand, traffic counts are currently obtained mostly using inductive loops installed by contractors in key locations in the road network for a certain period. As a result, traffic operation centers are currently unable to collect a continuous flow, and they only get counts at specific locations. Furthermore, interview participants highlighted that smart CCTV cameras are placed only in certain locations in the New Zealand road network, and other cameras produce low-quality footage. Therefore a study was set out to answer the research question: “Can traffic flow be estimated from low-quality CCTV video footage in real-time?”.

<sup>1</sup>NZTA’s traffic cameras API, <https://www.nzta.govt.nz/traffic-and-travel-information/infoconnect-section-page/about-the-apis/traffic-cameras/>

<sup>2</sup>Traffic flow estimation from CCTV images, <https://github.com/rangikanilani/Traffic-Flow-Estimation>

An analysis of the literature revealed that Fedorov et al. [104] had used the DL approach, Faster-RCNN and SORT tracker to obtain traffic flow from surveillance footage. Therefore, this study extended Fedorov’s work by downloading and annotating a New Zealand-based vehicle dataset for training, considering movement direction and vehicle class in traffic counting. Moreover, a novel DL algorithm, YOLOv4 was used and obtained a higher accuracy than Fedorov et al. [104]. This result implies the ability to obtain a highly accurate real-time traffic flow using CCTV video data that is discussed in Chapter 4.

This study is published in the Proceedings of the 18<sup>th</sup> International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021) [27]. In addition, the annotated dataset and pseudo-code for the proposed traffic estimation algorithms are made available for future researchers <sup>3</sup>.

#### 10.3.1.4 Short-term traffic flow prediction:

Short-term traffic flow prediction helps road traffic management in a variety of ways, including anticipating potential congestion and informing people about alternative routes, enabling traffic redistribution to avoid potential congestion in city environments [415, 183]. During the interviews, one of the responders highlighted the current status of traffic flow prediction in New Zealand:

“...So, to give you an example, the Wellington urban motorway has a very similar system that is now turned off because no one could get it to work. It was supposed to predict traffic flow. But it simply couldn’t work...”

The short-term traffic flow prediction problem has recently become more complex as more traffic data has become available, requiring models with greater data modeling capabilities. DL algorithms such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and recurrent neural networks (RNNs) <sup>4</sup> have been used for multiple traffic prediction tasks [391, 422, 405, 407]. They have received improved results over traditional parametric approaches such as autoregressive models and the Kalman filter [214, 375, 213, 182, 402, 96, 256, 92, 250].

Traffic systems are dynamic and susceptible to various external factors, including weather conditions, road maintenance activities, vehicle crash incidents, and other disaster events. As a result, traffic flow prediction models must also consider such parameters to make a more accurate forecast. However, in the literature, few studies have explored weather conditions as part of the traffic flow prediction problem with DL models. For example, a study by Koesdwiady et al. [194] used weather parameters such as temperature, humidity, visibility, wind speed, wind gust, dew point, and cloud layer height for a Deep Belief Network (DBN)-based traffic prediction model. They used decision-level fusion to combine traffic and weather data. Zhang et al. [415] proposed a Recurrent Neural Network (RNN)-based model using both weather (average wind speed, precipitation, maximal temperature, minimal temperature) and traffic data. In a recent study, Hou et al. [147] used a Stacked Autoencoder (SAE) and Radial Basis Function (RBF) neural network to predict traffic flow and showed improved results over using traffic-only data for the traffic prediction problem when combined with weather data. However, these studies have two main limitations; (1) they considered a very small dataset (e.g., three months). (2) The correlation between traffic flow and

<sup>3</sup>Traffic flow estimation from CCTV footage, <https://github.com/nilani-rangika/Traffic-Flow-from-Footage>

<sup>4</sup>A short introduction to DL techniques is provided in Appendix A



weather conditions plays an important role while developing DL models for traffic prediction [147], but none of these studies have considered the correlation patterns [415].

Therefore, the study presented in Chapter 5 went beyond Hou et al. [147] by proposing a novel DL model to predict traffic flow given weather parameters based on stacked Bi-LSTM networks. Furthermore, a correlation analysis was conducted using the Pearson coefficient value and received expert meteorologists' and transport experts' feedback for the first time for choosing weather parameters for the fusion model. Findings of Chapter 5 showed that the proposed DL architecture with Bi-LSTM networks achieved better results than Hou's [147] model.

### 10.3.1.5 Event extraction from multi-source unimodal data:

Emergency responders need actionable information associated with disaster situations in order to understand the development of a disaster (SA) and to facilitate decision-making, policy formulation, response, and resource allocation [98, 394]. As discussed in Section 10.3.1.1, information flows from multiple sources such as emails, telephone calls, social media, police reports, and first responder reports during a disaster. However, so far, there is no proper system to support responders in their SA by utilizing data from these sources.

“...So the challenge for our intelligence team is to create the common operating picture or COP...Currently, we do not have any sophisticated tool for this...”

People at the disaster sites provide quick updates using social media platforms; thus, recently, there has been an increasing interest in using social media data for disaster response [124, 130, 164, 412, 339, 185]. However, timely processing and analyzing an overwhelming amount of social media data brings several challenges. For example, data from social media platforms come in different modalities, such as text, image, audio, and video, and are inherently noisy. Moreover, the content is informal, mainly in colloquial language, and very brief with casual acronyms. Most importantly, SM contains rumors, fake information, and misinformation [163]. Therefore, as discussed previously, responders are reluctant to use SM data during emergency response.

Considering the time-critical nature of disaster environments, one of the main concerns is that the responding organizations need real-time information for their SA. Combining multiple sources leads to more information than learning from a single data source alone [166, 25]. As a result, the trustworthiness of the content provided through SM channels can be further improved. Online news provides mostly validated content, and multiple studies have explored the extraction and analysis of disaster events from news web texts [161, 168, 129]. However, very little work has explored jointly extracting information from online news and SM. For example, Verma et al. [380] analyzed a tweet and news corpora collected during the Nepal 2015 earthquake. They paired news with tweets as a supervised classification task using a Support Vector Machine (SVM) algorithm with a precision of 0.47. Petroni et al. [297] presented an online event extraction system using both news articles and tweets, trained to recognize breaking news events by co-referencing both media.

The existing research has multiple limitations, as follows:

- The work by Petroni et al. [297] does not provide adequate details of event extraction systems that allow for re-implementation by other researchers.
- The evaluation data sets used in these two research are not publicly available, which means extensions of these approaches are difficult.



- Both these studies do not provide examples for utilizing their approaches in real-time scenarios.

Furthermore, the structure of sentences, the information concerns, and the reporting perspectives vary among different web text sources [129]. Therefore, the characteristics of disaster information from varied web texts need to be further explored. As a result, a novel system named DEES-Disaster Event Extraction System was proposed to extract events and associated Spatio-temporal information from online news and social media in real-time using Natural Language Processing (NLP) techniques, specifically identifying the answers to *what*, *when*, and *where* questions. Therefore, the study presented in Chapter 6 goes beyond the work by Verma et al. [380] and Petroni et al. [297] in the following ways.

- introducing a new location-relatedness score in identifying the geolocation of the event.
- introducing a novel cross-media reference score for candidate scoring.
- introducing an extensible system for any disaster (i.e. additional event types can be added with minor modifications).
- training a DL algorithm for related tweet filtering while the mentioned research use rule-based and topic models to filter relevant tweets, and
- introducing dependency parsing for impact factor identification while Verma et al. [380] use an SVM classifier.

A demonstration of DEES is available at: <https://mu-clab.github.io/>

Relevant tweet filtering was necessary to complete the event extraction project. Therefore, a sub-project was conducted as described in Section 10.3.1.6.

#### 10.3.1.6 Disaster-related tweet classification:

Tweets are inherently noisy, and classifying disaster-related tweets is important for them to be utilized for disaster response [30]. A vast majority of the existing literature has focused on classifying tweets of the same event type and mostly used a single dataset for training classifiers [5, 110, 54, 191]. However, communication patterns of people change over the years, and, therefore, classification accuracy using classifiers trained on older datasets may not be high for future events [118]. Furthermore, supervised learning algorithms work well with more and complete training data covering the full spectrum of inputs the model should handle during the classification task.

Many approaches for tweet classification focus on particular disaster types such as flooding or earthquakes [5, 110, 58]. Only a few studies comprehensively test classification across various disaster types, such as the work by Graf et al. [118] and Wiegmann et al. [401]. Graf et al. [118] introduce a cross-domain informativeness classifier based on the SVM classifier. Wiegmann et al. [401] compare the effectiveness of three state-of-the-art machine learning models, namely CNN and two transformer models, BERT and Universal Sentence Encoder (USE), for the related tweet classification task. However, they explicitly consider only cross-disaster types. Also, these approaches have been mostly pursued in academic contexts and have not been made available to the public and responding organizations through easily accessible and integrable tools [54].

Therefore, large-scale machine learning and DL models were designed, and conducted evaluation experiments to identify disaster-related tweets combining multiple datasets. As discussed in

Chapter 6 this research extends Wiegmann’s [401] study by evaluating a large-scale ML, DL model for disaster-related tweet classification in three settings as follows:

- In-disaster: training and test data belong to the same disaster type.
- Out-disaster: training and test data belong to different disaster types.
- Cross-disaster: training set consists of tweets of various disaster types.

Moreover, all the learning weights are publicly available so that the response agencies can quickly adopt the trained models for an ongoing disaster.

This work is published in the Proceedings of the 18<sup>th</sup> International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021) [24]. The datasets, data pre-processing steps, models, and trained weights are publicly made available for future researchers<sup>5</sup>.

### 10.3.1.7 Real-time information extraction from multi-source multimodal data for disaster response

Section 10.3.1.5 emphasized the advantages of social media content for disaster response and the problems with reliability that prevent using SM. The content available in SM is multimodal because people include images, videos, and audio clips in addition to text posts. Researchers have recently analyzed multimodal datasets and demonstrated their usefulness in many disaster response tasks. Multimodal DL approaches have achieved higher accuracy than the DL algorithms employed to learn from a single modality alone. However, the majority of these approaches have only been explored with offline static datasets [1, 202, 7]. As discussed earlier, one of the primary problems is that the responding organizations require real-time information for their decision-making because disaster situations are time-critical. Work that has addressed real-time multimodal data extraction and fusion includes the Advanced System for Emergency Management (ASyEM), which was proposed by Foresti et al. [107] and Quelloffene Integrierte Multimedia Analyse (QuOIMA) introduced by Rogova et al. [329]. However, both these systems have three main limitations; (1) They do not discuss how different data sources and modalities are fused. (2) They do not evaluate how successful the fusion algorithms are (3) They do not provide examples for using the proposed architectures for real-time applications.

Therefore, previous work mentioned in Section 10.3.1.5 was extended to include text, visual and audio data for event extraction. In contrast to Foresti et al. [107] and Rogova et al. [329], the study presented in chapter 7 describes each component of the architecture in detail for future development. Moreover, the architecture was evaluated using the generalised precision score, and the findings demonstrate that integrating multiple sources and modalities yields higher accuracy than using a single source or modality for event extraction. The system that demonstrates functionality for real-world scenarios is available at <https://rangikanilani.github.io/events.html>.

Apart from the research or theoretical contributions, this study has some practical contributions for both disaster-related software developers and end-users. End users of the system are considered to be operators at TOCs (disaster responders) and the following sections discuss how the system can support them during daily operations.

---

<sup>5</sup>Disaster tweet classification, <https://github.com/nilani-rangika/Disaster-Tweet-Classification>

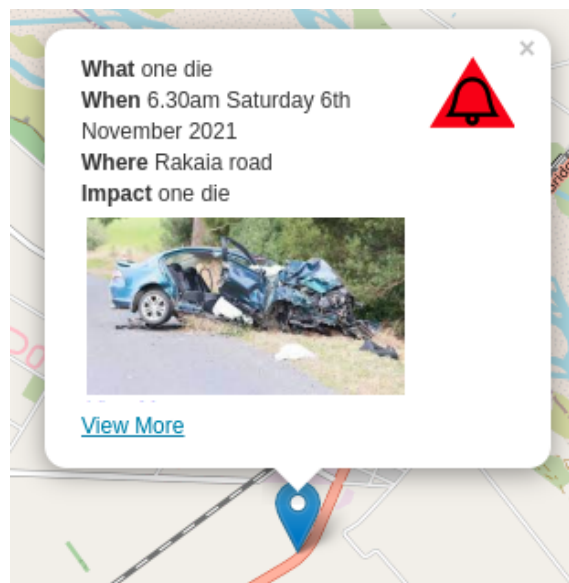
### 10.3.2 Contribution to practice

During this research, the overarching aim was to validate SM content by integrating multi-source multimodal data to improve the SA of disaster responders. As defined by Endsley, SA is the dynamic understanding of the “big picture” in space and time [98]. Moreover, Endsley defines SA as a three-level model of information processing. Level 1 is the perception of relevant elements in the environment; Level 2 is the comprehension of the significance of the elements related to task goals; and Level 3 is the projection of future actions of the elements in the environment. The following sections describe how the software prototype can support end users of the system to increase their SA across all three levels.

- **Level 1 SA: Perception of the Elements in the Environment**

According to Endsley, The first step in achieving SA is to perceive the status, attributes, and dynamics of relevant elements in the environment. For instance, a pilot must perceive the features of essential elements such as other aircraft, terrain, system condition, and warning lights with their characteristics [99].

Figure 10.1 shows the initial screen of the software prototype. The screen provides the event template, having only a summary of the event to make the responders aware of the event.



**Figure 10.1** Initial screen of the software prototype

- **Level 2 SA: Comprehension of the Current Situation**

The comprehension of the situation is based on the synthesis of independent Level 1 elements. As a result, level 2 SA extends beyond simply being aware of the elements present to include an understanding of their relevance in light of one’s goals. Moreover, the operators combine Level 1 data to create a complete picture of the environment, including understanding the significance of objects and events. For example, suppose a pilot sees warning lights indicating a problem during takeoff. In that case, they must quickly determine the severity of the problem in terms of the aircraft’s immediate airworthiness and combine this with knowledge of the number of runaways remaining to determine whether it is an abort situation [99]. Figure 10.2 shows the second screen of the software prototype. The responder can examine photographs

associated with the event by clicking the "view more" button on screen 1. Additionally, extra details from SM and online news sources are also included. This aids responders in assessing the severity of the incident to take necessary actions.

**News Headline:**

- [One dead after vehicle hits pedestrian on SH1 in Canterbury.](#)
- [One dead after crash near Rakaia](#)

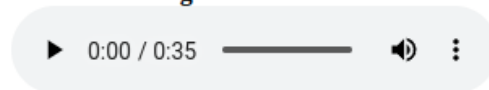
**Tweets:**

- I still don't believe that I saw the poor guy was crushed by the car near Pendarves Rakaia road
- First eye witness crash. Chertsey line and Pendarves Rakaia roads. He was an innocent pedestrian but died on the spot

**Images:**



**Voice Recording:**



[View Traffic Data](#)

**Figure 10.2** View more screen of the software prototype

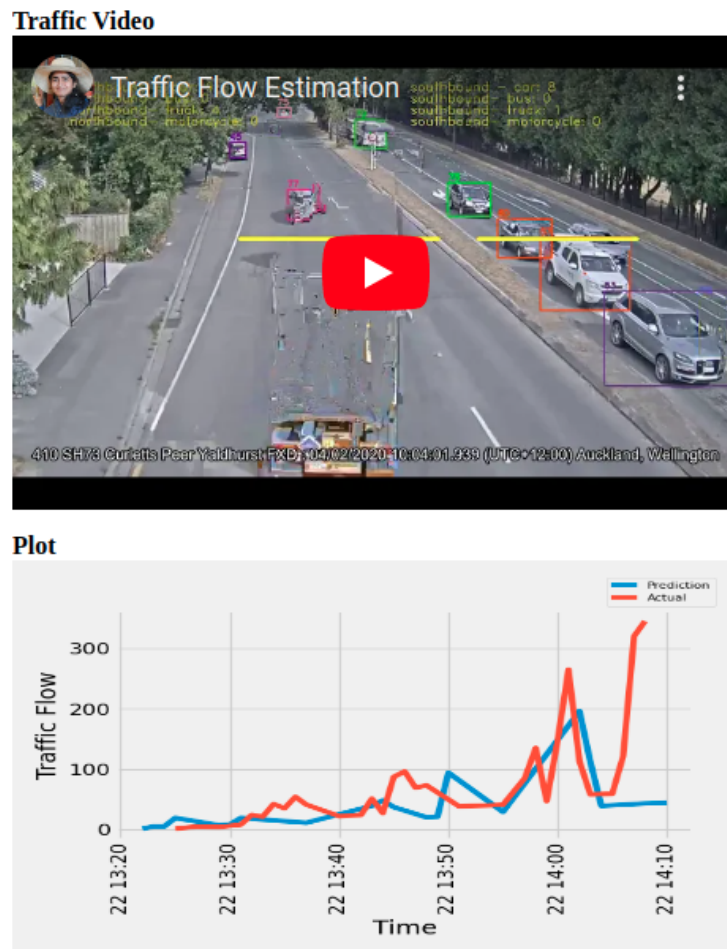
• **Level 3 SA: Projection of Future Status**

The third and highest degree of SA is the ability to predict the future activities of the elements in the environment, at least in the very near future. This is accomplished by understanding the circumstance, the dynamics of the elements, and their current status (both Level 1 and Level 2 SA). For instance, an Army commander may need to use his knowledge to integrate Level 2 and Level 1 information in order to estimate the direction from which enemy soldiers will approach, and the likely results of his own actions [99].

Projection of future decisions requires more details about the incidents. The final screen of the system provides live traffic counts from CCTV cameras. Moreover, a live plot shows the deviation of the live traffic count from the prediction. Therefore, the responders can decide whether they need to control the traffic lights, allocate detours or make any other management decisions.

In conclusion, the software prototype produced during this doctoral thesis can support all levels of SA, as summarised in Figure 10.4.

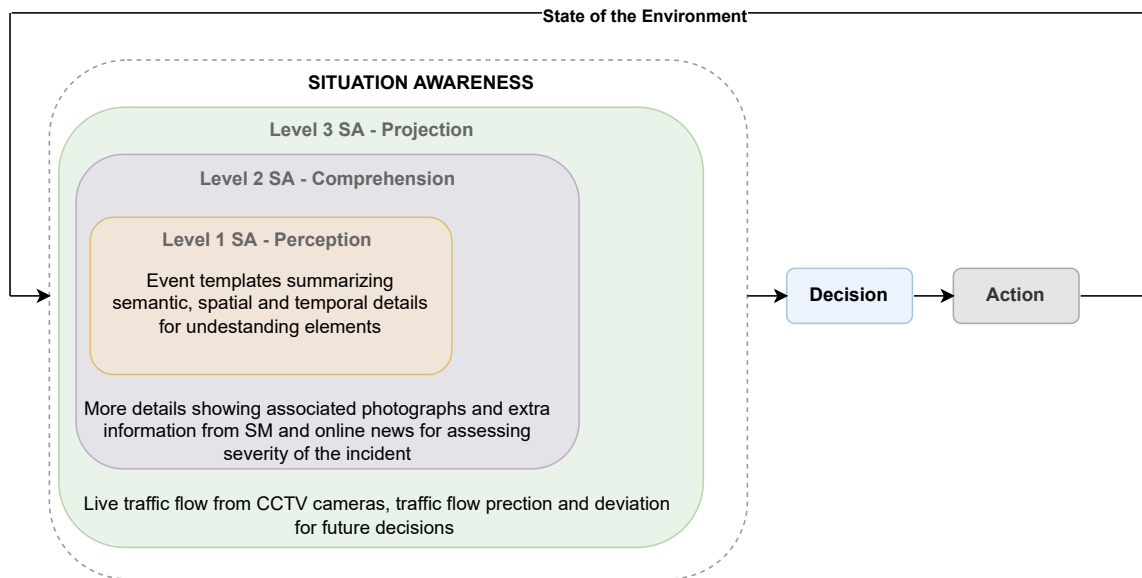
Additionally, the proposed software can support responders and developers in multiple ways,



**Figure 10.3** Final screen of the software prototype

as listed below.

- The real-time traffic flow estimation from CCTV images and footage can be used at TOCs for daily job tasks, such as understanding traffic flow patterns, predicting traffic flow at a given time, understanding traffic anomalies, and making management decisions.
- The novel DL algorithm proposed for short-term traffic flow prediction can be adopted at the TOCs for traffic prediction tasks to make daily decisions.
- The proposed system avoids the requirement for a person to monitor CCTV cameras 24 hours which currently happens.
- Most importantly, the extensibility of software artefact is an added advantage where the responders can modify it to identify any disaster event and associated communication in social media channels.
- The proposed system provides a collection of information in one place, and therefore, the responders are made easy to find and prepare reports for their briefings and lessons learned sessions.
- A vast number of studies relating to event extraction have been published only in the academic context, and there is no step-by-step guide provided for the practical implementation of such methods. However, the development of the software artefact is fully documented, and most



**Figure 10.4** An overview of how the software prototype supports all three levels of SA.

of the components, including algorithms, annotated datasets, and pre-processing techniques, are publicly made available for re-implementation. Therefore, the developers can use them as an implementation guideline.

## 10.4 Research Impacts

This research has been presented on several research platforms including a journal publication and two conference publications. The researcher won the “Delegates Choice Award – Best Poster” award for the poster titled “Real-time traffic flow estimation based on Deep Learning using CCTV videos” at the 2020 eResearch Australasia Conference. Apart from that, the research was presented as posters, abstracts, and lightning talks in the following forums.

- 05/03/2019 - 07/03/2019 - 2019 Disastrous Doctorate (3MT style presentation)  
Title: Multimedia Data Fusion and Analytics in Emergency Management to Inform Disaster Responders
- 02/09/2019 - 05/09/2019 - 2019 QuakeCoRE Annual Meeting (Poster presentation)  
Title: Identifying Research Gaps and Opportunities in the use of Multimodal Deep Learning for Emergency Management
- 19/08/2019 - Resilience to Natures Challenges: Urban theme Smart Resilient Cities workshop (3MT style presentation)
- 22/11/2019 - 2019 RNC2 Urban Theme Annual Research (Presentation)  
Title: Multi-source multimodal deep learning supporting urban traffic management
- 23/05/2020 - 2020 ISCRAM Conference (Presentation)  
Title: Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images
- 29/08/2020 - 2020 IEEE NZ Central Section Postgraduate Symposium

Title: Traffic Flow Estimation Based on Deep Learning for Emergency Traffic Management Using CCTV Images

- 10/09/2020 - 2020 New Zealand Research Software Engineering Conference - NZRSE 2020 (Lightning talk)

Title: Traffic Flow Estimation based on Deep Learning using CCTV Images

- 20/10/2020 - 2020 eResearch Australasia Conference (Lightning talk)

Title: Real-time traffic flow estimation based on Deep Learning using CCTV videos

- 30/11/2020 - 2020 RNC2 Urban Theme Annual Research (3MT style presentation)

Title: Real-time traffic flow estimation based on Deep Learning using CCTV videos

- 07/12/2020 - 10/12/2020 - 2020 QuakeCoRE Annual Meeting (Poster presentation)

Title: Real-time disaster event extraction from unstructured text sources

- 01/02/2021 - 03/02/2021 - 2021 Disastrous Doctorate (3MT style presentation)

Title: Disaster Tweet Classification

- 2021-05-24 - 2021 ISCRAM Doctoral Symposium (Poster presentation)

Title: Multi-source Multimodal Deep Learning to Improve Situation Awareness: An Application of Emergency Traffic Management

- 2021-05-25 - 2021 ISCRAM Conference (Presentation)

Title: Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage

- 2021-05-27 - 2021 ISCRAM Conference (Presentation)

Title: Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison

- 2021-05-28 - 2021 TechWeek Smart Resilience Cities Showcase (Lightning Talk)

Title: How AI can support future traffic management in Aotearoa New Zealand

- 2022-08-30 - NZ Geospatial Research Conference 2022 (Lightning Talk)

Title: Location reference extraction from unstructured web text

- 2022-11-08 - ISRAM Asia Pacific 2022 conference (Poster)

Title: Location reference extraction from disaster-related web text

In addition to the conferences, several invited presentations on this doctoral study were made to outside stakeholders, including the Christchurch Civil Defence and Emergency Management (CDEM) joint intelligence team and GNS science <sup>6</sup>. The research received highly positive feedback from the gatherings. For instance, one participant brought up how the New Zealand police had attempted to do the same with a complex method but failed. Moreover, they were delighted to implement the prototype developed into an actual software system and have already started initial discussions.

---

<sup>6</sup>GNS Science website: <https://www.gns.cri.nz/>

## 10.5 Research Implications, limitations and future work

During the requirements capturing interviews, it was revealed that the use of technology to assist disaster responders is minimal in the New Zealand context. For example, a recently published news article discusses outdated technology utilized in NZTA [280]. According to the news report, many critical Information Technology (IT) systems such as National Incident and Event Management System (NIEMS), Traffic Road Event Information System (TREIS) and national ticketing system remain years overdue or not performing as expected [280]. Therefore, it is evident that responders struggle with their SA and waste time manually gathering information due to limited access to IT systems. The findings of this study demonstrate how the most up-to-date social media data be successfully integrated to support all three levels of SA. An implication of this is the possibility that the software prototype can be developed into an actual system to be utilized at TOCs. Moreover, the evidence from this study suggests that introducing additional data sources and modalities for event extraction results in a comprehensive event extraction system. NZTA has access to a wide range of official reports, and if the system is created internally, these sources can also be effectively used to increase the accuracy of event extraction. Taken together, the complete architecture of the system outlined during this doctoral research can successfully be developed into a real-time system that shows live traffic, traffic deviations and events in order to assist responder SA during traffic emergencies (A demonstration is available here: <https://rangikanilani.github.io/events.html>).

The generalizability of the software artefact is one of the most important aspects to emphasise. Although the artefact's design, development, and evaluation were confined to emergency traffic management, it was built in a way that allows for extensions. As described in Chapter 7, the keyword-based rule matcher used in the event extraction system can be modified to limit the extraction for a single event or to add additional event types. Therefore, the extraction system can be used by responders for their SA during any disaster (A demonstration is available here: <https://rangikanilani.github.io/events1.html>).

Despite this doctoral thesis's research and practical contributions, several challenges and limitations must also be acknowledged. One of the significant challenges for this research was to obtain multimodal disaster datasets. The real-time nature of most of the parts of this research required live disaster data for evaluation purposes, which is practically hard to obtain. Therefore, the researcher confined the scope to emergency traffic management to overcome this challenge, as traffic emergencies are becoming more frequent. Moreover, since the research involved extensive training of DL models, specifically with visual datasets, it was challenging to find a computing environment with special hardware capabilities. As a result, the researcher applied for a high-performance computing facility at New Zealand eScience Infrastructure (NESI) and conducted the experiments. Furthermore, the circumstances created by the COVID-19 outbreak hindered the recruiting and interviewing of participants for this PhD project. For instance, the COVID-19 response plan severely slowed down the recruiting and interviewing of participants since individuals had to adjust to working from home. In addition, interview techniques had to be modified to accommodate the unpredictability and dynamism of the COVID-19 environment. Therefore, the majority of interviews were performed online as opposed to in-person.

A number of important limitations of this doctoral research need to be mentioned. First, the proposed software gets activated when there is a traffic flow deviation. However, as traffic systems



are dynamic, an extensive historical dataset is required to identify the traffic deviation function for a particular location. As it takes a considerable amount of time, this was considered beyond the PhD research scope. Second, the current study considered integrating only text, image, and audio data. However, video data has recently been more frequently shared through social media platforms, bringing more information than other modalities. Therefore, further work is required to identify methodologies to fuse video data. Third, the software prototype considered only extracting answers for *What (semantic)*, *Where (spatial)*, and *When (temporal)* (3W) questions. However, the system would be more beneficial if it could provide real-time event templates that answered 5W1H questions (*Who did (participant) What (semantic)*, *Where (spacial)* and *When (temporal)*, *Why (causal)*, and *How (method)*). Therefore, in future work, this software artefact can be extended to a full system capable of providing responders with real-time event templates covering the full spectrum of the event. Finally, one main concern of the users was the capability of integrating the system with existing applications. Therefore, further work is needed to develop an Application Programming Interface (API) which could provide the results of the artefact to any other system.

## 10.6 Conclusion

This project was undertaken to explore how DL approaches can be utilized to integrate multi-source multimodal data to support disaster response activities while cross-validating SM data. Emergency traffic management was selected as an application area to evaluate developed algorithms. This study has found that, generally, DL approaches proposed in research are only pursued academically and less implemented practically. Therefore, the researcher made available most of the annotated datasets, pre-processing steps, model details, and trained weights for future research. The findings of this study indicate that multimodal data extracted from several sources considerably improve information extraction compared to single sources. Furthermore, a software artefact was created employing the DSR method framework to illustrate the applicability of developed DL approaches throughout the project. It was demonstrated that this software prototype could support all three levels of SA: perception, comprehension, and projection. The present study makes several noteworthy contributions to both research and practice and has thrown up many questions that need further investigation.

# Appendices

## Appendix A

# Deep Learning - A short overview

---

# Deep Learning : A short overview

Nilani Algiriyage · Raj Prasanna · Kristin Stock · Emma  
E H Doyle · David Johnston

## 1 Deep Learning

AI is a field of study enabling machines to demonstrate the characteristics of human intelligence. Machine Learning (ML), a subset of AI, is a collection of algorithms that improve automatically through experience. They have been used to classify and cluster data to solve problems, such as spam detection, product recommendation, and online fraud detection, to name a few. Some of these ML algorithms need to be trained before they are used for research and are known as “supervised algorithms”. To train the algorithms, the researcher must collect a large set of labelled data relevant to the problem. On the other hand, “unsupervised” algorithms do not require labelled training data [14]. There are also “semi-supervised” algorithms where learning is based on partially labelled data sets. DL is a subset of ML (see Fig. 1). The main difference between traditional ML and DL is how the features are extracted. Traditional ML approaches use handcrafted features by applying a variety of feature extraction algorithms and then apply learning algorithms. In contrast, the features are learned automatically by the DL algorithm and are interpreted hierarchically in multiple levels [14].

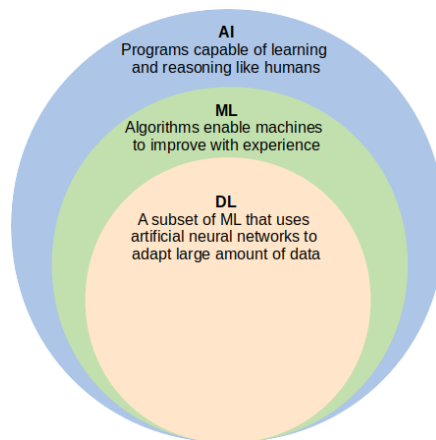


Fig. 1: The relationship between AI, ML and DL.

---

Nilani Algiriyage  
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.  
E-mail: r.nilani@massey.ac.nz

Raj Prasanna  
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.  
E-mail: r.prasanna@massey.ac.nz

Kristin Stock  
Institute of Natural and Mathematical Sciences, Massey University, Auckland, New Zealand.  
E-mail: k.stock@massey.ac.nz

Emma E H Doyle  
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.  
E-mail: e.e.hudson-doyle@massey.ac.nz

David Johnston  
Joint Centre for Disaster Research, Massey University, Wellington, New Zealand.  
E-mail: d.m.johnston@massey.ac.nz

The idea behind Artificial Neural Networks (ANNs), also known as Neural Networks (NNs), was inspired by the functioning of brain neurons. Generally, the brain can be represented as an interconnected set of nodes that can be organised in different layers; each layer generates outputs given certain inputs. DL algorithms are commonly identified under Deep Neural Networks (DNNs). A DNN is an ANN that has more than one layer of hidden nodes between its inputs and outputs (see Figures 2 and 3) [6]. DL algorithms can also be supervised: requiring labelled training data; unsupervised: not requiring labelled data for the training; or semi-supervised: partially requiring labelled training data.

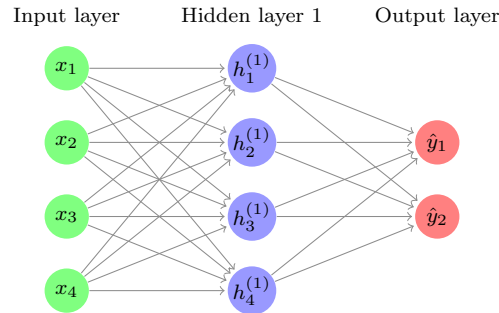


Fig. 2: Hidden layer of nodes in ANN.

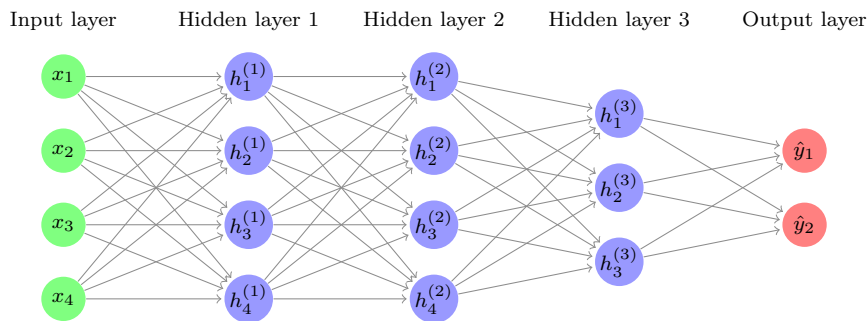


Fig. 3: Hidden layers of nodes in DNN.

As with the rapid increase in heterogeneous data sources and the multimodal data they provide, more sophisticated methods are required to analyse them. The main advantage of using DNNs is their ability to learn joint representations by correlating features in hidden layers. However, there exist different challenges in understanding multimodal representations, such as how to integrate data from heterogeneous sources, how to adapt to different levels of noise, and how to handle missing data [2]. The following section discusses the supervised DL techniques and the context of usage and their advantages and disadvantages.

### 1.0.1 Supervised Deep Learning Techniques

A supervised DL learning algorithm learns from the labelled training data and helps predict outcomes for unexpected data. We have identified Convolutional Neural Networks and Recurrent Neural Networks as more dominant supervised DL algorithms in many fields, including disaster research [19, 23].

*Convolutional Neural Network (CNN)*: CNN, alternatively known as ConvNet, is a type of DNN, largely applied for object recognition in *computer vision* research. Typically, the layers that form the CNN architecture are known as the convolutional layer(s), the pooling layer(s), and the fully connected layer. CNNs are mainly used for image processing tasks and recognise the patterns across space (for example, they first recognise the lines and curves and then the full object in an image). Each convolutional layer operates a set of learnable parameters, called filters, that have smaller dimensions than the input image.

During the training process, these filters go through the whole input volume (for example, in the case of a colour image, the input volume consists of width \* height \* number of the RGB (red, green, and blue) channels, that is 3) and calculate an inner product of the input volume and the filter. This computation over the whole input leads to a feature map of the filter. The objective of the pooling layer is to progressively reduce the

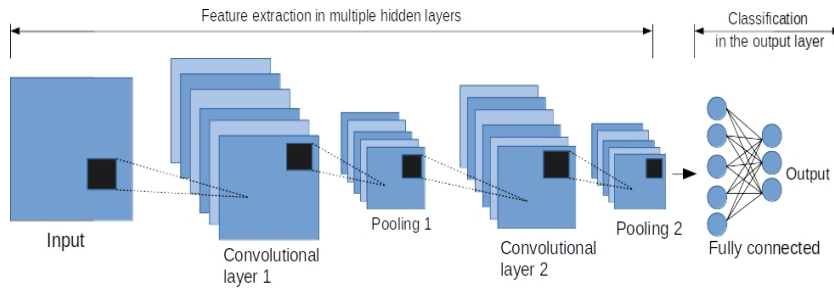


Fig. 4: The architecture of a CNN.

spatial size of the representation to reduce the number of parameters and computations in the network. The fully connected layer in the CNN represents the feature vector for the input. This feature vector is then further used for classification or translation to any other type of output [16, 18]. Over the years, variants of CNN architectures have been developed, such as AlexNet, DenseNet, VGGNet, and ResNet. More details of these architectures can be found in the survey paper by Khan et al. [11]. CNN has become very popular in multiple computer vision tasks, such as bidirectional images and sentence retrieval [10], emotion recognition [36], event recognition [32], and visual classification [16]. Fig. 4 shows the layout of a CNN.

*Recurrent Neural Networks (RNN):* RNNs are called recurrent, as they perform the same task for each element in a sequence. The input to an RNN consists of both the current sample and the previously observed sample. For example, the output of an RNN at time step  $t - 1$  affects the output at time step  $t$ . Each neuron is equipped with a feedback loop that returns the current output as an input for the next step. This structure can be expressed in such a way that each neuron in the RNN has an internal memory that keeps the information on the computations from the previous input [18].

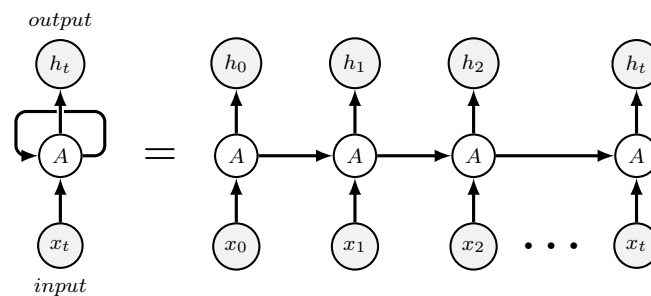


Fig. 5: The structure of a RNN.

The Long Short-Term Memory (LSTM) network is a variant of RNN. Both RNN and LSTM networks have been extensively used for analysing varying length sequences, such as videos, audio streams, and sentences [29, 30, 31, 34]. Fig. 5 depicts the structure of an RNN.

The following section discusses the unsupervised DL techniques and the context of usage, and their advantages and disadvantages.

### 1.0.2 Unsupervised Deep Learning Techniques

Unsupervised learning is a technique where we do not need to supervise the model. Instead, we allow the model to work on its own to discover patterns. Therefore, it mainly deals with the unlabelled data. Deep Belief Networks, Deep Boltzmann Machine, and Autoencoder are well-known unsupervised DL algorithms and are discussed below.

*Deep Belief Networks (DBN):* DBNs are a graphical representation that is essentially generative in nature as it produces all possible values that can be generated for the case at hand. DBNs consist of multiple layers with values. However, there is a relation between the layers but not the values. The main aim is to help the system classify the data into different categories. Srivastava et al. [25] introduced Multimodal Deep Belief Networks (DBN) to learn a joint density model over multimodal input space. In their multimodal DBM setting, two separate DBNs for text and image are trained in a completely unsupervised fashion and joined. Multimodal

DBNs have been applied in Audio-Visual Speech Recognition (AVSR) [8] and in gesture recognition [33], given the audio and skeleton features that assist the co-learning.

*Deep Boltzmann Machine (DBM)*: DBM is a generative, stochastic model that can graphically represent as a set of interconnected visible and hidden nodes. Srivastava et al. [26] introduced a Multimodal-DBM. The applications of the algorithm include gesture recognition [33] and AVSR [8]. The key idea is to learn a distribution over multimodal inputs and fill the missing modalities using the conditional distribution of them given the observed distribution. The Multimodal-DBM model is capable of obtaining the joint representation even in the absence of some modalities, and the joint representation can imply real-world concepts. Also, it is possible to fill in the missing modalities given the observed ones due to their generative nature. DBMs have difficulty in training, high computational costs, and the need to use approximate variational training techniques [26]. The architectures of DBM and DBN are shown in Fig. 6.

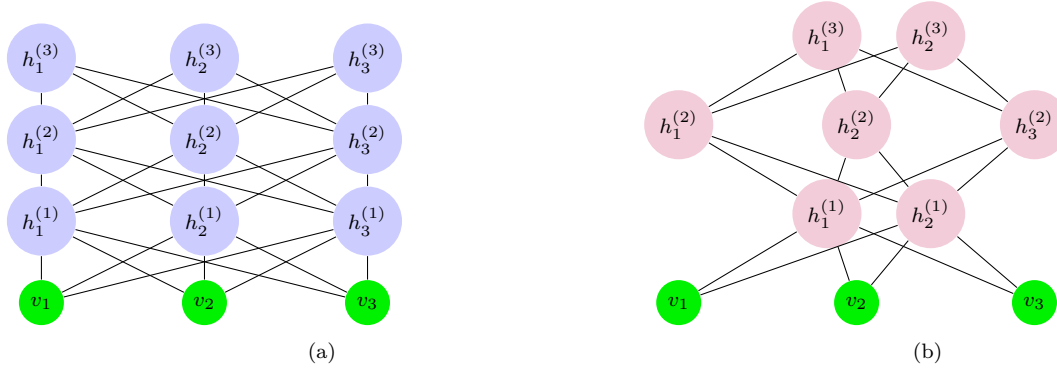


Fig. 6: The architectures of (a) DBM and (b) DBN.

*Autoencoder (AE)*: Autoencoders (AEs) are a type of DNN, which have input, hidden, and output layers. However, the input layer is forced to be identical to the output layer. This network aims to reconstruct the input by transforming inputs into outputs in the simplest possible way such that it does not distort the input very much. AEs are used for dimensionality reduction. The work by Jiquan Ngiam et al. [21] introduced the use of AEs in a multimodal context. They used denoising autoencoders for each modality and fused them into a multimodal representation using another autoencoder layer. AEs have been successfully used in video retrieval [17], and video-based human pose recognition [7]. The main advantage is that the model does not need any labelled data for the pre-training. However, it cannot handle missing data. Fig. 7 illustrates the structure of a typical AE.

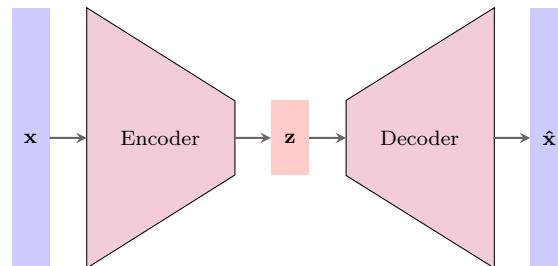


Fig. 7: The idea of an Autoencoder.

Table 1 summarises the supervised and unsupervised DL techniques and their applications. The survey paper by Zhang et al. [35] discusses more details of using DL models for BDA.

Table 1: Supervised and unsupervised DL techniques.

Application	Supervised DL Algorithm	Un-supervised DL Algorithm
1. Affect recognition <ul style="list-style-type: none"> <li>• Emotion recognition</li> <li>• Personality trait recognition</li> </ul>	CNN [36]	DBM [33], AE [7]
2. Event recognition <ul style="list-style-type: none"> <li>• Human action &amp; event recognition</li> </ul>	CNN [32]	
3. Media description <ul style="list-style-type: none"> <li>• Visual captioning</li> <li>• Visual Question Answering</li> </ul>	RNN/LSTM [29]	
4. Multimedia retrieval <ul style="list-style-type: none"> <li>• Bi-directional visual sentence search</li> </ul>	CNN [10], RNN/LSTM	AE [17]
5. Speech recognition	CNN [27]	DBM , DBN [8]
6. Visual classification	CNN [16]	DBN

### 1.1 Semi-supervised Deep Learning Techniques

Semi-supervised deep learning is a class of DL algorithms that are able to learn from partially labelled data sets. Generative Adversarial Networks (GAN) and Domain-Adversarial Neural Networks (DANN) are used as a semi-supervised learning technique. Additionally, RNNs, including LSTM, are also used for semi-supervised learning.

*Generative Adversarial Networks (GAN):* GANs consist of two neural networks, namely the generative and discriminative networks, which work together to produce high-quality data [4]. The generator is responsible for producing new data after learning the data distribution from the training data set. The discriminator discriminates between actual data (coming from training data) and fake input data (coming from the generator). The objective function in GANs is based on minimax theory so that one network seeks to maximise the value function while the other network wants to minimise it. In each step, the generator, willing to fool the discriminator, produces sample data from random noise. The discriminator receives several real data examples from the training set along with samples from the generator. Then, the discriminator determines how good the generated samples are. The output of the discriminator helps the generator to optimise the generated data for the next round. The idea of a GAN is shown in Fig. 8. Having been inspired by GAN, Ganin et al.[3] proposed a Domain-Adversarial Neural Network (DANN) for semi-supervised problems that includes a component that explicitly aims to reduce the shift between a source and a target.

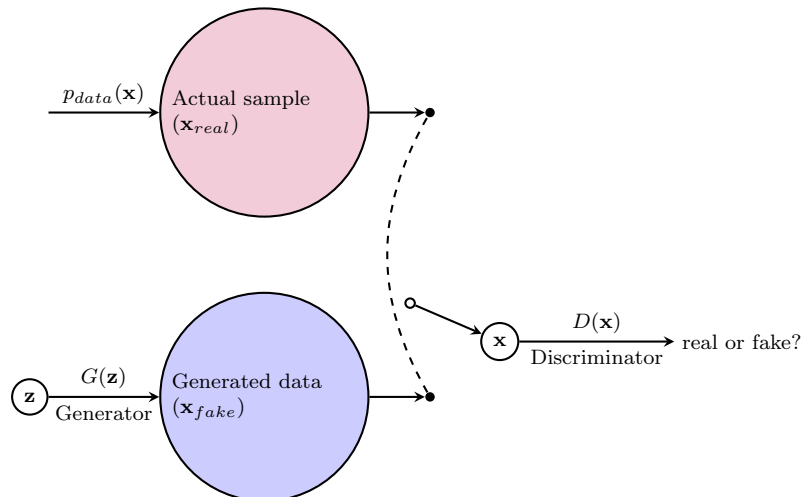


Fig. 8: The idea of a generative adversarial network.



It is a tedious task to train DL algorithms from scratch due to multiple reasons, such as dataset labelling and the computational power required for training. As a result, researchers have adopted multiple techniques such as transfer learning and domain adaptation to reuse already trained dataset.

## 1.2 Transfer Learning

Transfer learning is a highly adapted technique in DL research when there is insufficient data for a new domain. In transfer learning, a pre-trained DL model on a large data set (source) is applied to a new data set (target). The pre-trained network can be customised for 1) feature extraction and 2) fine-tuning to further train to make the model more relevant for the specific task. There are different sub-settings of transfer learning, such as inductive, transductive, and unsupervised transfer learning [24]. Table 2 shows the different transfer learning settings.

Table 2: Transfer learning settings.

Transfer learning setting	Source Domain Labels	Target Domain Labels
Inductive Transfer Learning	Available or unavailable	Available
Transductive Transfer Learning	Available	Unavailable
Unsupervised Transfer Learning	Unavailable	Unavailable

*Domain Adaptation* is a sub-class of transfer learning where data from a source domain is used to predict a target domain, under the assumption that the source and target domains have different distributions but share some similar patterns. It is related to transductive transfer learning, where the labels of the source domain data are available while the labels of the target domain data are not available. The task is to learn a classifier for the target data, using the labelled source data and the unlabelled target data. This technique has a high potential to be applied in the field of disaster research, given the lack of data just after a disaster [15]. Pan et al. [24] provide a comprehensive analysis of different transfer learning approaches. Table 3 shows the pre-trained DL models used in the papers we analysed.

Table 3: Transfer learning models used by the surveyed papers.

Author	Pre-trained Model	Data Modality	Description
[1, 13, 20, 22]	VGG-16	visual	A CNN trained on more than one million images from the ImageNet <sup>1</sup> database
[5, 15]	VGG-19	visual	A CNN trained on more than one million images from the ImageNet database
[9, 12]	GloVe	text	An unsupervised learning algorithm for obtaining vector representations for words
[28]	SoundNet	audio	A pre-trained model of natural sound representations using 2 million videos
[28]	Inception-v3	visual	A CNN trained on more than one million images from the ImageNet database

<sup>1</sup> <http://www.image-net.org/>

[19, 23]	AlexNet	visual	A CNN trained on more than one million images from the ImageNet database
----------	---------	--------	--

## References

- Attari, N., Ofli, F., Awad, M., Lucas, J., Chawla, S.: Nazr-cnn: Fine-grained classification of uav imagery for damage assessment. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 50–59. IEEE (2017). doi:10.1109/DSAA.2017.72
- Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(2), 423–443 (2019)
- Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* **17**(1), 2096–2030 (2016). doi:10.1007/978-3-319-58347-1\_10
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in neural information processing systems*, pp. 2672–2680 (2014)
- Hezaveh, M.M., Kanan, C., Salvaggio, C.: Roof damage assessment using deep learning. In: 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 6403–6408. IEEE (2017). doi:10.1109/AIPR.2017.8457946
- Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Kingsbury, B., et al.: Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine* **29** (2012)
- Hong, C., Yu, J., Wan, J., Tao, D., Wang, M.: Multimodal deep autoencoder for human pose recovery. *IEEE Transactions on Image Processing* **24**(12), 5659–5670 (2015). doi:10.1109/TIP.2015.2487860
- Huang, J., Kingsbury, B.: Audio-visual deep learning for noise robust speech recognition. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 7596–7599. IEEE (2013). doi:10.1109/ICASSP.2013.6639140
- Kabir, M.Y., Madria, S.: A deep learning approach for tweet classification and rescue scheduling for effective disaster management. In: *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 269–278 (2019). doi:10.1145/3347146.3359097
- Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: *Advances in neural information processing systems*, pp. 1889–1897 (2014)
- Khan, A., Sohail, A., Zahoor, U., Qureshi, A.S.: A survey of the recent architectures of deep convolutional neural networks. *arXiv preprint arXiv:1901.06032* (2019). doi:10.1007/s10462-020-09825-6
- Kumar, A., Singh, J.P.: Location reference identification from tweets during emergencies: A deep learning approach. *International journal of disaster risk reduction* **33**, 365–375 (2019). doi:10.1016/j.ijdrr.2018.10.021
- Kumar, A., Singh, J.P., Dwivedi, Y.K., Rana, N.P.: A deep multi-modal neural network for informative twitter content classification during emergencies. *Annals of Operations Research* pp. 1–32 (2020). doi:10.1007/s10479-020-03514-x
- LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *nature* **521**(7553), 436 (2015). doi:10.1038/nature14539
- Li, X., Caragea, D., Caragea, C., Imran, M., Ofli, F.: Identifying disaster damage images using a domain adaptation approach. In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, Valencia, Spain. Academic Press (2019)
- Li, Y., Ye, S., Bartoli, I.: Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning. *Journal of Applied Remote Sensing* **12**(4), 045008 (2018). doi:10.1117/1.JRS.12.045008
- Liu, Y., Feng, X., Zhou, Z.: Multimodal video classification with stacked contractive autoencoders. *Signal Processing* **120**, 761–766 (2016). doi:10.1016/j.sigpro.2015.01.001
- Mohammadi, M., Al-Fuqaha, A., Sorour, S., Guizani, M.: Deep learning for iot big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials* **20**(4), 2923–2960 (2018). doi:10.1109/COMST.2018.2844341
- Muhammad, K., Ahmad, J., Baik, S.W.: Early fire detection using convolutional neural networks during surveillance for effective disaster management. *Neurocomputing* **288**, 30–42 (2018). doi:10.1016/j.neucom.2017.04.083
- Naga Anitha, A.M.: Detection of disaster affected regions based on change detection using deep architecture. *Procedia Computer Science* **8** (2019)
- Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696 (2011)

22. Nguyen, D.T., Ofli, F., Imran, M., Mitra, P.: Damage assessment from social media imagery data during disasters. In: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, pp. 569–576 (2017). doi:10.1145/3110025.3110109
23. Pamuncak, A., Guo, W., Soliman Khaled, A., Laory, I.: Deep learning for bridge load capacity estimation in post-disaster and-conflict zones. *Royal Society open science* **6**(12), 190227 (2019). doi:10.1098/rsos.190227
24. Pan, S.J., Yang, Q.: A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**(10), 1345–1359 (2009). doi:10.1109/TKDE.2009.191
25. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: International conference on machine learning workshop, vol. 79 (2012)
26. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in neural information processing systems, pp. 2222–2230 (2012)
27. Tatulli, E., Hueber, T.: Feature extraction using multimodal convolutional neural networks for visual speech recognition. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2971–2975. IEEE (2017). doi:10.1109/ICASSP.2017.7952701
28. Tian, H., Zheng, H.C., Chen, S.C.: Sequential deep learning for disaster-related video classification. In: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 106–111. IEEE (2018). doi:10.1109/MIPR.2018.00026
29. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K.: Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014). doi:10.3115/v1/N15-1173
30. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156–3164 (2015). doi:10.1109/CVPR.2015.7298935
31. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence* **39**(4), 652–663 (2017). doi:10.1109/TPAMI.2016.2587640
32. Wang, L., Wang, Z., Du, W., Qiao, Y.: Object-scene convolutional neural networks for event recognition in images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 30–35 (2015). doi:10.1109/CVPRW.2015.7301333
33. Wu, D., Shao, L.: Multimodal dynamic networks for gesture recognition. In: Proceedings of the 22nd ACM international conference on Multimedia, pp. 945–948. ACM (2014). doi:10.1145/2647868.2654969
34. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp. 2048–2057 (2015)
35. Zhang, Q., Yang, L.T., Chen, Z., Li, P.: A survey on deep learning for big data. *Information Fusion* **42**, 146–157 (2018). doi:10.1016/j.inffus.2017.10.006
36. Zhang, S., Zhang, S., Huang, T., Gao, W.: Multimodal deep convolutional neural network for audio-visual emotion recognition. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 281–284. ACM (2016). doi:10.1145/2911996.2912051

## Appendix B

# Information Sheet

## ***Multi-source Multimodal Deep Learning to Improve Situation Awareness: An Application of Emergency Traffic Management***

### **INFORMATION SHEET**

#### **Researcher(s) Introduction**

My name is Rangika Nilani, a PhD student at the Joint Centre for Disaster Research – Massey University.

#### **Project Description and Invitation**

This research focuses on fusing multiple modalities of data such as text and visuals using deep learning in real-time to provide structured information (events) for disaster response. I have created a prototype system to demonstrate the capabilities of our proposed system.

The purpose of the interview is to evaluate the software prototype to further improve it.

#### **Project Procedures**

Interview will ideally be held in an office or meeting room in Massey University, Wellington Campus or any other location convenient for the participant. (The interviews will be held using zoom due to the current Covid-19 level restrictions.)

Participation will involve individual face-to-face audio-recorded interview with the researcher which can take approximately between 40 minutes to one hour.

Your name and identity will be held in confidence.

The participant will be asked to sign a written consent form to confirm his/her agreement to take part in the interview to be audio recorded prior to the start of the interview.

Once the scoping of the project is completed participants will be given the opportunity to indicate if they would like to receive a summary of the results at the end of the project.

#### **Data Management**

The data will be used for gather user ideas for the proposed project and will be securely stored for a period of five years after which time the files will be destroyed. The audio files and notes will be stored separately from the consent forms.

#### **Participant's Rights**

You are under no obligation to accept this invitation. If you decide to participate, you have the right to:

- decline to answer any particular question;
- withdraw from the study (specify timeframe);
- ask any questions about the study at any time during participation;

- provide information on the understanding that your name will not be used unless you give permission to the researcher;
- be given access to a summary of the project findings when it is concluded.
- ask for the recorder to be turned off at any time during the interview.

### **Project Contacts**

If you would like more information about the research please contact Rangika Nilani or her primary supervisor Dr Raj Prasanna.

Rangika Nilani  
Joint Centre for Disaster Research  
School of Psychology  
Wellington Campus  
Massey University  
Phone: [REDACTED]  
Email: R.Nilani@massey.ac.nz

Dr Raj Prasanna  
Joint Centre for Disaster Research  
School of Psychology  
Wellington Campus  
Massey University  
Phone: 04 801 5799 ext. 62169  
Email: R.Prasanna@massey.ac.nz

“This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University’s Human Ethics Committees. The researcher(s) named above are responsible for the ethical conduct of this research.

If you have any concerns about the conduct of this research that you wish to raise with someone other than the researcher(s), please contact Prof Craig Johnson, Director, Research Ethics, telephone 06 356 9099 x 85271, email [humanethics@massey.ac.nz](mailto:humanethics@massey.ac.nz)”.

## Appendix C

# Consent Form

***Multi-source Multimodal Deep Learning to Improve Situation Awareness: An Application of Emergency Traffic Management***

**PARTICIPANT CONSENT FORM - INDIVIDUAL**

I have read, or have had read to me in my first language, and I understand the Information Sheet attached as Appendix I. I have had the details of the study explained to me, any questions I had have been answered to my satisfaction, and I understand that I may ask further questions at any time. I have been given sufficient time to consider whether to participate in this study and I understand participation is voluntary and that I may withdraw from the study at any time.

1. I agree/do not agree to the interview being sound recorded. (if applicable include this statement)
2. I wish/do not wish to have my recordings returned to me. (if applicable include this statement)
3. I wish/do not wish to have data placed in an official archive. (if applicable include this statement)
4. I agree to participate in this study under the conditions set out in the Information Sheet.

**Declaration by Participant:**

I \_\_\_\_\_ hereby consent to take part in this study.

**Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_



## Appendix D

# Prototype Evaluation User Guide and Interview Questions - Round 1

# Prototype Evaluation: User Guide

\*Algiriyage, N., Prasanna, R., Doyle, E. E., Stock, K., Johnston, D.  
Massey University  
\*R.Nilani@massey.ac.nz

August 2021

This prototype demonstrates the functionality of our proposed Multi-Source Multimodal Event Extraction System for Disaster Response (**M<sub>2</sub>E<sub>2</sub>S** for DR). We conducted multiple interviews with relevant end-users and stakeholders before developing the system. All captured requirements are listed in Appendix 1. The system extracts live events from online news and tweets. We collect online news from three main online news providers in New Zealand namely, rnz news <sup>1</sup>, nzherald news <sup>2</sup> and stuff news <sup>3</sup> through their rss feeds. The tweets are extracted in real-time from the Twitter streaming Application Programming Interface (API) using python tweepy library <sup>4</sup>. We set the geographical boundary only to New Zealand to collect tweets generated by users within New Zealand.

Figure 1 illustrates the interface of the home screen. The home screen has four components as follows.

- Component 01 : Real-time Traffic Flow Counting from CCTV Footage
- Component 02: Real-time Traffic Flow vs Traffic Flow Prediction
- Component 03: Multi-Source Multimodal Event Extraction for Disaster Response
- Component 04: Geo-location

The link between different components are illustrated in Figure 2. The next sections describe each component in detail.

---

<sup>1</sup>rnz news, <https://www.rnz.co.nz/>

<sup>2</sup>nzherald news, <https://www.nzherald.co.nz/>

<sup>3</sup>stuff news, <https://www.stuff.co.nz/>

<sup>4</sup>Python tweepy library version 3.9.0, <https://pypi.org/project/tweepy/>

# Multi-source Multimodal Deep Learning to Improve Situation Awareness: An Application of Emergency Traffic Management



Nilani Agrigayage, The Joint Centre for Disaster Research, Massey University ©2021

Figure 1: Multi-source Multimodal Event Extraction System ( $M_2E_2S$ )

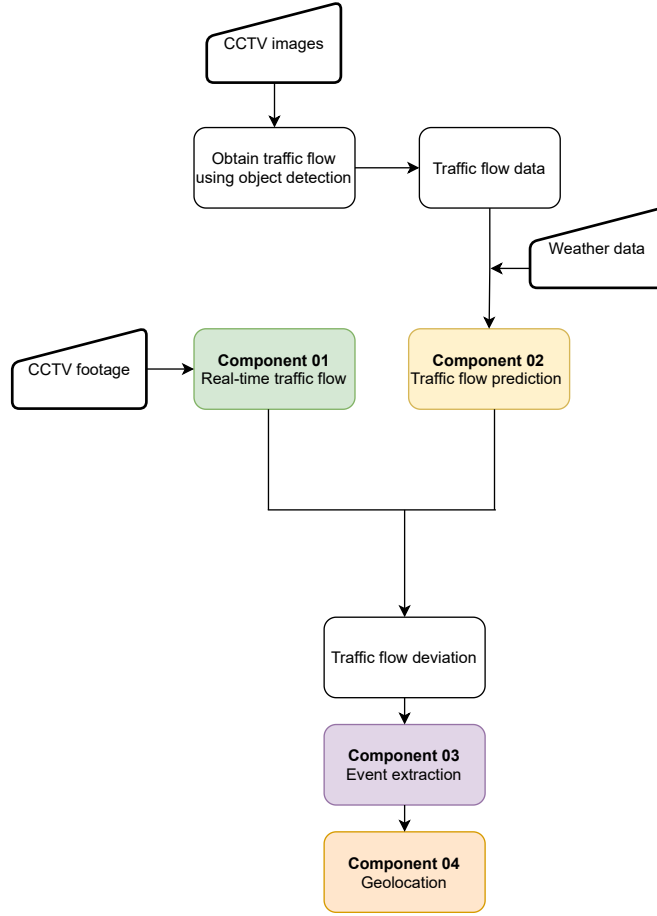


Figure 2: The links between different components of  $M_2E_2S$

### 0.1 Component 01 : Real-time Traffic Flow Counting from CCTV Footage

Component 01 demonstrates Real-time Traffic Flow Counting from CCTV Footage. The user can integrate the CCTV tracking system here (see Figure 3). We have trained the Yolo-v4 algorithm to detect vehicle objects for a traffic dataset collected in Christchurch, New Zealand. Furthermore, we introduced an algorithm to count the number of vehicles based on class such as car, bus, truck, van and bike and movement direction. Therefore, Component 01 deals with counting live traffic from the integrated CCTV camera footage. More details of our traffic flow estimation system are available in the following publications.

“Nilani Algiriyage, Raj Prasanna, Emma E H Doyle, Kristin Stock, & David Johnston. (2020). **Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images**. In Amanda Hughes, Fiona McNeill, & Christopher W. Zobel (Eds.), *ISCRAM 2020 Conference Proceedings – 17th International Conference on Information Systems for Crisis Response and Management* (pp. 100–109). Blacksburg, VA (USA): Virginia Tech.”

“Nilani Algiriyage, Raj Prasanna, Emma E H Doyle, Kristin Stock, & David Johnston. (2021). **Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage.** In Amanda Hughes, Fiona McNeill, & Christopher W. Zobel (Eds.), *ISCRAM 2021 Conference Proceedings – 18th International Conference on Information Systems for Crisis Response and Management* (pp. 40–48). Blacksburg, VA (USA): Virginia Tech.”



Figure 3: Component 01 - Real-time Traffic Flow Counting from CCTV Footage

## 0.2 Component 02: Real-time Traffic Flow vs Traffic Flow Prediction

Component 02 of our system visualizes live traffic flow and predicted traffic flow in a single plot. Live traffic data are from the Component 01. We have trained a Deep Learning algorithm to make a short term traffic flow prediction using both traffic and weather data. The prediction model predicts the traffic flow for the next 20 minutes. Figure 4 illustrates the screen for live traffic flow and predicted traffic flow visualization.

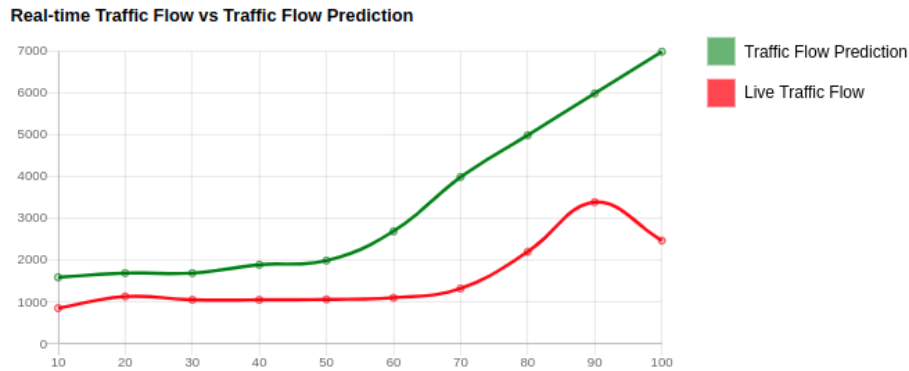


Figure 4: Component 02 - Real-time Traffic Flow vs Traffic Flow Prediction

### 0.3 Component 03: Multi-Source Multimodal Event Extraction for Disaster Response

Component 03 of our system extracts events in real-time from online news and tweets if there is a significant deviation between the real-time traffic flow and the prediction. The event extraction system comes up with event templates answering *what*, *when* and *where* questions. Additionally, it will provide detailed links for the online news and tweets. The user can click on the links to access the related news articles or tweets of the event. Moreover, all related images extracted from online news and tweets are available for the user. Furthermore, the event extraction system can be run separately if the user wants to identify events relating to any ongoing events (e.g., flooding, armed conflict and storms). Figure 5 shows the screen for event extraction component.

**Multi-Source Multimodal Event Extraction for Disaster Response**

What	One die
When	3.15 p.m (20th April 2020)
Where	Bay of Plenty
More Details	[Online News] <a href="#">Article 1</a>
	[Online News] <a href="#">Article 2</a>
More Details	[Tweets] <a href="#">Tweet 1</a>
	[Tweets] <a href="#">Tweet 2</a>

**Related Images**

[View More Images](#)

Figure 5: Component 03 - Multi-Source Multimodal Event Extraction for Disaster Response

## 0.4 Component 04: Geo-location

The fourth component will show the Geo-location of the incident in a map. This geo-location is extracted from the event extracted in Component 03. This will allow the user to see the exact location in a live map where the incident has occurred. Figure 6 shows the screen for geo-location.

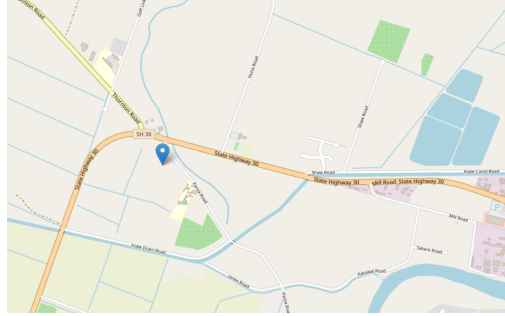


Figure 6: Component 04 - Geo-location

All the events collected from the system will be stored in a database. Additionally, the collected traffic flow data will be used to improve the traffic flow prediction algorithm. The functions of  $M_2E_2S$  are summarized in Table 1.

ID	Function
01	Real-time traffic flow count from footage by direction
02	Real-time traffic flow count from footage by vehicle class
03	Traffic flow prediction using traffic and weather data
04	Plot live-traffic vs prediction for next 30 minutes
05	Extract text data in real-time using tweets and online news
06	Cluster text data group them
07	Extract event templates answering what, when and where questions
08	Geo-locate the event in a live map

Table 1: Functions of the  $M_2E_2S$

# 1 Appendix

## 1.1 Requirements Captured during Phase 1 Interviews

---

Requirement ID	Requirement
R.1	The system shall be able to keep history data
R.2	The system shall be able to summarize information
R.3	The users shall be able to track the progress of the tasks
R.4	The system shall be able to prioritize tasks
R.5	The system shall be able to provide comprehensive details of the events
R.6	The system shall be able to make predictions based on history data
R.7	The system shall be able to used for Business As Usual (BAU) purposes
R.8	The system shall be able to integrate multiple data sources
R.9	The system shall be able to count live traffic flow from footage
R.10	The system shall be able to extract live events
R.11	The system shall be able to provide the location of the event

---



## 2 Interview Script

### 2.1 Preliminary questions

1. Designation?
2. Based location?
3. Short description of the current job role?
4. What sort of Information Systems (ISs) are you using for your daily job tasks?
  - Do you use any real-time systems?
  - Do you use social media for your daily job tasks?
5. What roles you have played in your previous careers?
  - What sort of Information Systems (ISs) have you used in your previous careers?
  - What sort of training / prior training do you have of ISs and other types of real-time information platforms?
  - Have you used any IS(s) for real-time events?

### 2.2 Functionality of M<sub>2</sub>E<sub>2</sub>S

#### 2.2.1 Repeat questions for component 1 to 4

1. How well will the component be able to support your daily decision making tasks?
2. What functional improvements would you like see so that the system will be more useful for you?
3. Any final comments?

#### 2.2.2 Questions for the whole system

1. In a scale of 1 to 4 how do you prioritize the functions of each component in M<sub>2</sub>E<sub>2</sub>S?
  - Briefly explain the reasons for your rating?
2. How well will the system be able to support your daily decision making tasks?
3. What functional improvements would you like see so that the system will be more useful for you?
4. Any final comments?

### 2.3 Usability of M<sub>2</sub>E<sub>2</sub>S

1. Do you think the information is clear, concise, and informative?
  - What comments or issues do you have?
  - How can the system be more improved to make it clear, concise, and informative?
2. Do you like how the information is presented?
  - What comments or issues do you have?
3. Do you think the component/system is easy to use?
  - What comments or issues do you have?

- How can the system be more improved to make it easy to use?
4. Do you think the component/system is easy to learn for a novice user?
    - What comments or issues do you have?
    - How can the system be more improved to make it easy to learn?
  5. What more information you would like to see?
    - Why do you want to see this information?
  6. Any final comments?

## Appendix E

# Interview Questions - Round 2

# Software Prototype Evaluation (Round 02)- M<sub>2</sub>E<sub>2</sub>S

February 2022

This questionnaire evaluates the functionality, usability, reliability and performance of our proposed Multi-Source Multimodal Event Extraction System for Disaster Response (M<sub>2</sub>E<sub>2</sub>S). Currently, is prototype level. It extracts live news and tweets. News is collected from three main online news providers in New Zealand: rnz news, nzherald news, and stuff news through their RSS feed. The tweets are extracted in real-time from the Twitter streaming Application Programming Interface (API) using the python tweepy library. The geographical boundary is set to New Zealand to collect tweets generated by users within New Zealand.

## 0.1 Functions

M<sub>2</sub>E<sub>2</sub>S is expected to generate event templates answering (W3) questions, what, when and where. Apart from the News and Tweets, the system can integrate call-center recordings if available for event extraction. In addition, the new version of the prototype can include the impact information such as Severe, Moderate and Minor.

## 0.2 This section evaluates the M<sub>2</sub>E<sub>2</sub>S through different aspects of its functionality.

1. M<sub>2</sub>E<sub>2</sub>S has all the necessary main features for its intended tasks
  - Strongly Agree
  - Agree
  - Neutral
  - Disagree
  - Strongly Disagree
2. M<sub>2</sub>E<sub>2</sub>S supports decision-makers to undertake activities and make decisions
  - Strongly Agree
  - Agree
  - Neutral
  - Disagree
  - Strongly Disagree
3. M<sub>2</sub>E<sub>2</sub>S can be generalized for any events and not specific or limited
  - Strongly Agree
  - Agree
  - Neutral
  - Disagree
  - Strongly Disagree
4. M<sub>2</sub>E<sub>2</sub>S can collect appropriate history data

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

5. M<sub>2</sub>E<sub>2</sub>S can be combined to an existing software

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

6. Any final comments regarding the functionality of M<sub>2</sub>E<sub>2</sub>S?

.....

**0.3 This section evaluates the M<sub>2</sub>E<sub>2</sub>S through different aspects of its usability.**

1. M<sub>2</sub>E<sub>2</sub>S is simple to use

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

2. The sequence of screens are clear

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

3. The information presented in screens is clear, concise, and informative

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

4. The information presented using drill-down approach avoids information overload

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

5. M<sub>2</sub>E<sub>2</sub>S is easy to learn for a novice user

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

6. Users can solve real-world problems using this M<sub>2</sub>E<sub>2</sub>S in an acceptable way

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

7. Users can easily reuse the M<sub>2</sub>E<sub>2</sub>S after some time of not having used it, without having to learn everything all over again

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

8. Any final comments regarding the usability of M<sub>2</sub>E<sub>2</sub>S?

.....

#### **0.4 This section evaluates the M<sub>2</sub>E<sub>2</sub>S through different aspects of its reliability.**

1. The number of errors that users can generate are minimal

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

2. The generated event templates are accurate

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

3. Warning dialogues are generated where necessary

- Strongly Agree
- Agree
- Neutral

- Disagree
- Strongly Disagree

4. Error messages are helpful

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

5. Any final comments regarding the reliability of M<sub>2</sub>E<sub>2</sub>S?

.....

### **0.5 This section evaluates the M<sub>2</sub>E<sub>2</sub>S through different aspects of its performance.**

1. The response time of the M<sub>2</sub>E<sub>2</sub>S is acceptable

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

2. The speed of event generation by M<sub>2</sub>E<sub>2</sub>S is acceptable (Average time taken to download 1000 tweets : 5.45 secs, Average time taken to download news : 0.02 sec, Average time taken to generate event templates : 2.39 sec)

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

3. Any final comments regarding the performance of M<sub>2</sub>E<sub>2</sub>S?

.....

### **0.6 Questions regarding the improvements.**

1. Are you happy with the improvements done?

- Strongly Agree
- Agree
- Neutral
- Disagree
- Strongly Disagree

2. Would you recommend to develop M<sub>2</sub>E<sub>2</sub>S as a software?

- Strongly Agree
- Agree

- Neutral
- Disagree
- Strongly Disagree

3. What improvements do you expect in the final software application?

.....

4. Any final comments regarding M<sub>2</sub>E<sub>2</sub>S?

.....

**Thank you very much for your participation!**



## Appendix F

# Human Ethics Notification



Date: 19 December 2018

Dear Nilani Algiriyage Hewa Algiriyage

Re: Ethics Notification - **4000020443** - **Cross-Domain Data Fusion and Analytics of Social Media in Disaster Management**

Thank you for your notification which you have assessed as Low Risk.

Your project has been recorded in our system which is reported in the Annual Report of the Massey University Human Ethics Committee.

The low risk notification for this project is valid for a maximum of three years.

If situations subsequently occur which cause you to reconsider your ethical analysis, please contact a Research Ethics Administrator.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

**A reminder to include the following statement on all public documents:**

*"This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research."*

*If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director - Ethics, telephone 06 3569099 ext 85271, email [humanethics@massey.ac.nz](mailto:humanethics@massey.ac.nz)."*

Please note, if a sponsoring organisation, funding authority or a journal in which you wish to publish requires evidence of committee approval (with an approval number), you will have to complete the application form again, answering "yes" to the publication question to provide more information for one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

Yours sincerely

**Research Ethics Office, Research and Enterprise**

Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand **T** 06 350 5573; 06 350 5575 **F** 06 355 7973

**E** [humanethics@massey.ac.nz](mailto:humanethics@massey.ac.nz) **W** <http://humanethics.massey.ac.nz>

Human Ethics Low Risk notification

A handwritten signature in blue ink, appearing to read 'C Johnson', on a light-colored background.

Professor Craig Johnson  
Chair, Human Ethics Chairs' Committee and Director (Research Ethics)

**Research Ethics Office, Research and Enterprise**

Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand **T** 06 350 5573; 06 350 5575 **F** 06 355 7973  
**E** [humanethics@massey.ac.nz](mailto:humanethics@massey.ac.nz) **W** <http://humanethics.massey.ac.nz>

## Appendix G

### DRC 16 Forms

## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Rangika Nilani Hewa Algiriyage
Name and title of main supervisor:	Associate Professor Raj Prasanna
In which chapter is the manuscript/published work?	Two
What percentage of the manuscript/published work was contributed by the student?	85%

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.

Please select one of the following three options:

- The manuscript/published work is published or in press**  
Please provide the full reference of the research output:  
Algiriyage, N., Prasanna, R., Stock, K., Doyle, E. E., & Johnston, D. (2022). Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review. SN Computer Science, 3(1), 1-29.
- The manuscript is currently under review for publication**  
Please provide the name of the journal:
- It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:		Main supervisor's signature:	<b>Raj Prasanna</b> Digitally signed by Raj Prasanna Date: 2023.04.16 19:08:31 +12'00'
----------------------	---	------------------------------	--

*This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.*

## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Rangika Nilani Hewa Algiriyage
Name and title of main supervisor:	Associate Professor Raj Prasanna
In which chapter is the manuscript/published work?	Five
What percentage of the manuscript/published work was contributed by the student?	85%

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.

Please select one of the following three options:


- The manuscript/published work is published or in press**  
Please provide the full reference of the research output:  
Nilani Algiriyage, Raj Prasanna, Emma E H Doyle, Kristin Stock, & David Johnston. (2020). Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images. In Amanda Hughes, Fiona McNeill, & Christopher W. Zobel (Eds.), ISCRAM 2020 Conference Proceedings – 17th Internati
- The manuscript is currently under review for publication**  
Please provide the name of the journal:
- It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:		Main supervisor's signature:	<b>Raj Prasanna</b> Digitally signed by Raj Prasanna Date: 2023.04.16 19:09:16 +12'00'
----------------------	---	------------------------------	--

*This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.*

## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Rangika Nilani Hewa Algiriyage		
Name and title of main supervisor:	Associate Professor Raj Prasanna		
In which chapter is the manuscript/published work?	Five		
What percentage of the manuscript/published work was contributed by the student?	85%		
Describe the contribution that the student has made to the manuscript/published work: The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.			
Please select one of the following three options:			
<input checked="" type="radio"/>	<b>The manuscript/published work is published or in press</b> Please provide the full reference of the research output: Nilani Algiriyage, Raj Prasanna, Kristin Stock, Emma Hudson-Doyle, David Johnston, Minura Punchihewa, et al. (2021). Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage. In Anouck Adrot, Rob Grace, Kathleen Moore, & Christopher W. Zobel (Eds.), ISCRAM 2021 Conference Proceedi		
<input type="radio"/>	<b>The manuscript is currently under review for publication</b> Please provide the name of the journal:		
<input type="radio"/>	<b>It is intended that the manuscript will be published, but it has not yet been submitted to a journal</b>		
Student's signature:		Main supervisor's signature:	<b>Raj Prasanna</b> Digitally signed by Raj Prasanna Date: 2023.04.16 19:09:53 +12'00'

*This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.*

## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Rangika Nilani Hewa Algiriyage		
Name and title of main supervisor:	Associate Professor Raj Prasanna		
In which chapter is the manuscript/published work?	Six		
What percentage of the manuscript/published work was contributed by the student?	85%		

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.

Please select one of the following three options:

- The manuscript/published work is published or in press**  
Please provide the full reference of the research output:
- The manuscript is currently under review for publication**  
Please provide the name of the journal:  
International Journal of Intelligent Transportation Systems Research
- It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:		Main supervisor's signature:	<b>Raj Prasanna</b> Digitally signed by Raj Prasanna Date: 2023.04.16 19:10:52 +12'00'
----------------------	---	------------------------------	--

*This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.*



## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Rangika Nilani Hewa Algiriyage	
Name and title of main supervisor:	Associate Professor Raj Prasanna	
In which chapter is the manuscript/published work?	Seven	
What percentage of the manuscript/published work was contributed by the student?	85%	

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.

Please select one of the following three options:

- The manuscript/published work is published or in press**  
Please provide the full reference of the research output:  
Algiriyage, N., Prasanna, R., Stock, K. et al. DEES: a real-time system for event extraction from disaster-related web text. Soc. Netw. Anal. Min. 13, 6 (2023). <https://doi.org/10.1007/s13278-022-01007-2>
- The manuscript is currently under review for publication**  
Please provide the name of the journal:
- It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:		Main supervisor's signature:	<b>Raj Prasanna</b> Digitally signed by Raj Prasanna Date: 2023.04.16 19:11:33 +12'00'
----------------------	---	------------------------------	--

*This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.*

## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Rangika Nilani Hewa Algiriyage
Name and title of main supervisor:	Associate Professor Raj Prasanna
In which chapter is the manuscript/published work?	Seven
What percentage of the manuscript/published work was contributed by the student?	85%

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.

Please select one of the following three options:



**The manuscript/published work is published or in press**

Please provide the full reference of the research output:

Nilani Algiriyage, Rangana Sampath, Raj Prasanna, Kristin Stock, Emma Hudson-Doyle, & David Johnston. (2021). Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison. In Anouck Adrot, Rob Grace, Kathleen Moore, & Christopher W. Zobel (Eds.), ISCRAM 2021 Conference Proceedings – 18th Internati



**The manuscript is currently under review for publication**

Please provide the name of the journal:



**It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:



Main supervisor's signature:

**Raj  
Prasanna**

Digitally signed by  
Raj Prasanna  
Date: 2023.04.16  
19:13:08 +12'00'

*This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.*

## STATEMENT OF CONTRIBUTION DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Rangika Nilani Hewa Algiriyage		
Name and title of main supervisor:	Associate Professor Raj Prasanna		
In which chapter is the manuscript/published work?	Eight		
What percentage of the manuscript/published work was contributed by the student?	85%		

Describe the contribution that the student has made to the manuscript/published work:

The candidate conducted the data collection and analysis, drafted the manuscript and made subsequent revisions based on the supervisors' feedback.

Please select one of the following three options:

- The manuscript/published work is published or in press**  
Please provide the full reference of the research output:
- The manuscript is currently under review for publication**  
Please provide the name of the journal:
- It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:		Main supervisor's signature:	<b>Raj Prasanna</b> Digitally signed by Raj Prasanna Date: 2023.04.16 19:13:58 +12'00'
----------------------	---	------------------------------	--

*This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis.*

# References

- [1] Mahdi Abavisani et al. “Multimodal categorization of crisis events in social media”. In: *arXiv* (2020). ISSN: 23318422.
- [2] Fabian Abel et al. “Twitcident: fighting fire with information from social web streams”. In: *Proceedings of the 21st International Conference on World Wide Web*. ACM. 2012, pp. 305–308.
- [3] Chadia Abras, Diane Maloney-Krichmar, Jenny Preece, et al. “User-centered design”. In: *Bainbridge, W. Encyclopedia of Human-Computer Interaction*. Thousand Oaks: Sage Publications 37.4 (2004), pp. 445–456.
- [4] Flavia Sofia Acerbo and Claudio Rossi. “Filtering informative tweets during emergencies: A machine learning approach”. In: *I-TENDER 2017 - Proceedings of the 2017 1st CoNEXT Workshop on ICT Tools for Emergency Networks and DisastEr Relief* (2017), pp. 1–6. DOI: 10.1145/3152896.3152897.
- [5] Flavia Sofia Acerbo and Claudio Rossi. “Filtering informative tweets during emergencies: a machine learning approachfacerbo2017filtering”. In: *Proceedings of the First CoNEXT Workshop on ICT Tools for Emergency Networks and DisastEr Relief*. 2017, pp. 1–6. DOI: 10.1145/3152896.3152897.
- [6] Anant Agarwal et al. “Efficient Traffic Density Estimation Using Convolutional Neural Network”. In: *2020 6th International Conference on Signal Processing and Communication (ICSC)*. IEEE. 2020, pp. 96–100. DOI: 10.1109/ICSC48311.2020.9182718.
- [7] Mansi Agarwal et al. “Crisis-DIAS: Towards Multimodal Damage Analysis - Deployment, Challenges and Assessment”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.01 (2020), pp. 346–353. ISSN: 2159-5399. DOI: 10.1609/aaai.v34i01.5369.
- [8] Kashif Ahmad et al. “Automatic detection of passable roads after floods in remote sensed and social media data”. In: *Signal Processing: Image Communication* 74.December 2018 (2019), pp. 110–118. ISSN: 09235965. DOI: 10.1016/j.image.2019.02.002. arXiv: 1901.03298. URL: <https://doi.org/10.1016/j.image.2019.02.002>.
- [9] Tariq Ahmad and Allan Ramsay. “Linking tweets to news: Is all news of interest?” In: *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer. 2016, pp. 151–161.
- [10] Alan Aipe et al. “Deep learning approach towards multi-label classification of crisis related tweets”. In: *Proceedings of the 15th ISCRAM Conference*. 2018.

- [11] Alan Aipe et al. “Linguistic Feature Assisted Deep Learning Approach towards Multi-label Classification of Crisis Related Tweets”. In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. Ed. by Kees Boersma and Brian M. Tomaszewski. ISCRAM Association, 2018. URL: [http://idl.iscram.org/files/alanaipe/2018/1592%5C\\_AlanAipe%5C\\_etal2018.pdf](http://idl.iscram.org/files/alanaipe/2018/1592%5C_AlanAipe%5C_etal2018.pdf).
- [12] Shahriar Akter and Samuel Fosso Wamba. “Big data and disaster management: a systematic review and agenda for future research”. In: *Annals of Operations Research* (2017), pp. 1–21.
- [13] Shahriar Akter and Samuel Fosso Wamba. “Big data and disaster management: a systematic review and agenda for future research”. In: *Annals of Operations Research* (2017), pp. 1–21. ISSN: 15729338. DOI: 10.1007/s10479-017-2584-2.
- [14] Firoj Alam, Muhammad Imran, and Ferda Ofli. “Image4Act: Online social media image processing for disaster response”. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017* (2017), pp. 601–604. DOI: 10.1145/3110025.3110164.
- [15] Firoj Alam, Shafiq Joty, and Muhammad Imran. “Domain adaptation with adversarial training and graph embeddings”. In: *arXiv preprint arXiv:1805.05151* (2018).
- [16] Firoj Alam, Shafiq Joty, and Muhammad Imran. “Graph Based Semi-Supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets”. In: *Twelfth International AAI Conference on Web and Social Media*. 2018.
- [17] Firoj Alam, Shafiq R. Joty, and Muhammad Imran. “Graph Based Semi-Supervised Learning with Convolution Neural Networks to Classify Crisis Related Tweets”. In: *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*. AAAI Press, 2018, pp. 556–559. URL: <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17815>.
- [18] Firoj Alam, Ferda Ofli, and Muhammad Imran. “CrisisMMD: Multimodal Twitter Datasets from Natural Disasters”. In: *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*. AAAI Press, 2018, pp. 465–473. URL: <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/2017816>.
- [19] Firoj Alam, Ferda Ofli, and Muhammad Imran. “Crisismmd: Multimodal twitter datasets from natural disasters”. In: *arXiv preprint arXiv:1805.00713* (2018).
- [20] Firoj Alam et al. “Flood detection via twitter streams using textual and visual features”. In: *arXiv* (2020), pp. 4–6. ISSN: 23318422. arXiv: 2011.14944.
- [21] MD Jahedul Alam et al. “Evaluation of the traffic impacts of mass evacuation of Halifax: flood risk and dynamic traffic microsimulation modeling”. In: *Transportation research record* 2672.1 (2018), pp. 148–160. DOI: 10.1177/0361198118799169.
- [22] David E Alexander. “Social media in disaster risk reduction and crisis management”. In: *Science and engineering ethics* 20.3 (2014), pp. 717–733.
- [23] Nilani Algiriyage et al. *Data Analysis Details of the Systematic Literature Review of DL for DR*. <https://github.com/mu-clab/DLforDR>. Accessed: 2021-09-18.

- [24] Nilani Algiriyage et al. “Identifying Disaster-related Tweets: A Large-Scale Detection Model Comparison”. In: (2021). Ed. by Anouck Adrot et al., pp. 731–743. URL: [http://idl.iscram.org/files/nilanialgiriyage/2021/2368\\_NilaniAlgiriyage\\_etal2021.pdf](http://idl.iscram.org/files/nilanialgiriyage/2021/2368_NilaniAlgiriyage_etal2021.pdf).
- [25] Nilani Algiriyage et al. “Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review”. In: *SN Comput. Sci.* 3.1 (2022), p. 92. DOI: 10.1007/s42979-021-00971-4. URL: <https://doi.org/10.1007/s42979-021-00971-4>.
- [26] Nilani Algiriyage et al. “Multi-source Multimodal Data and Deep Learning for Disaster Response: A Systematic Review”. In: *SN Comput. Sci.* 3.1 (2022), p. 92. DOI: 10.1007/s42979-021-00971-4. URL: <https://doi.org/10.1007/s42979-021-00971-4>.
- [27] Nilani Algiriyage et al. “Towards Real-time Traffic Flow Estimation using YOLO and SORT from Surveillance Video Footage”. In: ().
- [28] Nilani Algiriyage et al. “Traffic Flow Estimation based on Deep Learning for Emergency Traffic Management using CCTV Images”. In: *17th International Conference on Information Systems for Crisis Response and Management* (2020), pp. 100–109.
- [29] Ebtesam Alomari, Rashid Mehmood, and Iyad Katib. “Road traffic event detection using twitter data, machine learning, and apache spark”. In: *2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*. IEEE. 2019, pp. 1888–1895.
- [30] Reem ALRashdi and Simon O’Keefe. “Deep Learning and Word Embeddings for Tweet Classification for Crisis Response”. In: *arXiv preprint arXiv:1903.11024* (2019).
- [31] Nasser Alsaedi, Pete Burnap, and Omer Rana. “Can we predict a riot? Disruptive event detection using Twitter”. In: *ACM Transactions on Internet Technology (TOIT)* 17.2 (2017), pp. 1–26.
- [32] Siti Nor Khuzaimah Binti Amit and Yoshimitsu Aoki. “Disaster detection from aerial imagery with convolutional neural network”. In: *Proceedings - International Electronics Symposium on Knowledge Creation and Intelligent Computing, IES-KCIC 2017* 2017-January.July 2018 (2017), pp. 239–245. DOI: 10.1109/KCIC.2017.8228593.
- [33] M Anbarasan et al. “Detection of flood disaster system based on IoT, big data and convolutional deep neural network”. In: *Computer Communications* 150 (2020), pp. 150–157. DOI: 10.1016/j.comcom.2019.11.022.
- [34] Arif et al. “Visual attention-based comparative study on disaster detection from social media images”. In: *Innovations in Systems and Software Engineering* 16.3-4 (2020), pp. 309–319. ISSN: 16145054. DOI: 10.1007/s11334-020-00368-1.
- [35] Octavio Arriaga, Paul Plöger, and Matias Valdenegro-Toro. “Image Captioning and Classification of Dangerous Situations”. In: *CoRR* abs/1711.02578 (2017). arXiv: 1711.02578. URL: <http://arxiv.org/abs/1711.02578>.
- [36] Zahra Ashktorab et al. “Tweedr: Mining twitter to inform disaster response.” In: *ISCRAM*. 2014.

- [37] Nazia Attari et al. “Nazr-CNN: Fine-Grained classification of UAV imagery for damage assessment”. In: *Proceedings - 2017 International Conference on Data Science and Advanced Analytics, DSAA 2017* 2018-Janua (2017), pp. 50–59. DOI: 10.1109/DSAA.2017.72. arXiv: 1611.06474.
- [38] Gerhard Backfried et al. “Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project”. In: *2013 European Intelligence and Security Informatics Conference, Uppsala, Sweden, August 12-14, 2013*. IEEE, 2013, pp. 143–146. DOI: 10.1109/EISIC.2013.31. URL: <https://doi.org/10.1109/EISIC.2013.31>.
- [39] Anas Bassam Al-Badareen et al. “Software quality models: A comparative study”. In: *International Conference on Software Engineering and Computer Systems*. Springer. 2011, pp. 46–55.
- [40] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal machine learning: A survey and taxonomy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443.
- [41] Saksham Bansal. “A Mutli-Task Mutlimodal Framework for Tweet Classification Based on CNN (Grand Challenge)”. In: *Proceedings - 2020 IEEE 6th International Conference on Multimedia Big Data, BigMM 2020* (2020), pp. 456–460. DOI: 10.1109/BigMM50055.2020.00075.
- [42] Johan Barthélemy et al. “Edge-computing video analytics for real-time traffic monitoring in a smart city”. In: *Sensors* 19.9 (2019), p. 2048. DOI: 10.3390/s19092048.
- [43] Moumita Basu, Kripabandhu Ghosh, and Saptarshi Ghosh. “Information Retrieval from Microblogs During Disasters: In the Light of IRMiDis Task”. In: *SN Computer Science* 1.1 (2020), pp. 1–10. DOI: 10.1007/s42979-020-0065-1.
- [44] Moumita Basu et al. “Extracting Resource Needs and Availabilities from Microblogs for Aiding Post-Disaster Relief Operations”. In: *IEEE Transactions on Computational Social Systems* 6.3 (2019), pp. 604–618. ISSN: 2329924X. DOI: 10.1109/TCSS.2019.2914179.
- [45] Moumita Basu et al. “Extracting resource needs and availabilities from microblogs for aiding post-disaster relief operations”. In: *IEEE Transactions on Computational Social Systems* 6.3 (2019), pp. 604–618.
- [46] Mesay Belete Bejiga et al. “A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery”. In: *Remote Sensing* 9.2 (2017). ISSN: 20724292. DOI: 10.3390/rs9020100.
- [47] Alex Bewley et al. “Simple online and realtime tracking”. In: *2016 IEEE International Conference on Image Processing (ICIP)*. 2016, pp. 3464–3468. DOI: 10.1109/ICIP.2016.7533003.
- [48] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. “YOLOv4: Optimal Speed and Accuracy of Object Detection”. In: *arXiv preprint arXiv:2004.10934* (2020).

- [49] Barry W. Boehm, John R. Brown, and M. Lipow. “Quantitative Evaluation of Software Quality”. In: *Proceedings of the 2nd International Conference on Software Engineering, San Francisco, California, USA, October 13-15, 1976*. Ed. by Raymond T. Yeh and C. V. Ramamoorthy. IEEE Computer Society, 1976, pp. 592–605. URL: <http://dl.acm.org/citation.cfm?id=807736>.
- [50] Piotr Bojanowski et al. “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. DOI: 10.1162/tacl\_a\_00051.
- [51] Virginia Braun and Victoria Clarke. “Using thematic analysis in psychology”. In: *Qualitative research in psychology* 3.2 (2006), pp. 77–101.
- [52] Virginia Braun and Victoria Clarke. “What can “thematic analysis” offer health and well-being researchers?” In: *International journal of qualitative studies on health and well-being* 9 (2014).
- [53] Khac-Hoai Nam Bui et al. “Video-Based Traffic Flow Analysis for Turning Volume Estimation at Signalized Intersections”. In: *Asian Conference on Intelligent Information and Database Systems*. Springer. 2020, pp. 152–162. DOI: 10.1007/978-3-030-42058-1\_13.
- [54] Grégoire Burel and Harith Alani. “Crisis event extraction service (CREES) – Automatic detection and classification of crisis-related content on social media”. In: *Proceedings of the International ISCRAM Conference 2018-May* (2018), pp. 597–608. ISSN: 24113387.
- [55] Grégoire Burel and Harith Alani. “Crisis event extraction service (CREES) – Automatic detection and classification of crisis-related content on social media”. In: *Proceedings of the International ISCRAM Conference 2018-May* (2018), pp. 597–608. ISSN: 24113387.
- [56] Grégoire Burel, Hassan Saif, and Harith Alani. “Semantic wide and deep learning for detecting crisis-information categories on social media”. In: *International Semantic Web Conference*. Springer. 2017, pp. 138–155.
- [57] Gibson Burrell and Gareth Morgan. *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life*. Routledge, 2017.
- [58] Cornelia Caragea, Adrian Silvescu, and Andrea H Tapia. “Identifying informative messages in disaster events using convolutional neural networks”. In: *International Conference on Information Systems for Crisis Response and Management*. 2016, pp. 137–147.
- [59] Cornelia Caragea et al. “Classifying text messages for the haiti earthquake.” In: *ISCRAM*. Citeseer. 2011.
- [60] S Kay Carpender et al. “Urban evacuations and rural America: lessons learned from Hurricane Rita”. In: *Public Health Reports* 121.6 (2006), pp. 775–779.
- [61] L Carver and Murray Turoff. “The human and computer as a team in emergency management information systems”. In: *CACM* 50.3 (2007), pp. 33–38.
- [62] Carlos Castillo. *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press, 2016.
- [63] Young Jin Cha, Wooram Choi, and Oral Büyüköztürk. “Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks”. In: *Computer-Aided Civil and Infrastructure Engineering* 32.5 (2017), pp. 361–378. ISSN: 14678667. DOI: 10.1111/mice.12263.



- [64] Pranamesh Chakraborty et al. “Traffic Congestion Detection from Camera Images using Deep Convolution Neural Networks”. In: *Transportation Research Record* 2672.45 (2018), pp. 222–231. ISSN: 21694052. DOI: 10.1177/0361198118777631.
- [65] Neha Chaudhuri and Indranil Bose. “Exploring the role of deep neural networks for post-disaster decision support”. In: *Decision Support Systems* 130.July 2019 (2020), p. 113234. ISSN: 01679236. DOI: 10.1016/j.dss.2019.113234. URL: <https://doi.org/10.1016/j.dss.2019.113234>.
- [66] Fang Chen and Bo Yu. “Earthquake-Induced Building Damage Mapping Based on Multi-Task Deep Learning Framework”. In: *IEEE Access* 7 (2019), pp. 181396–181404. DOI: 10.1109/ACCESS.2019.2958983.
- [67] Fang Chen and Bo Yu. “Earthquake-Induced Building Damage Mapping Based on Multi-Task Deep Learning Framework”. In: *IEEE Access* 7 (2019), pp. 181396–181404. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2958983.
- [68] Chih Shen Cheng, Amir H. Behzadan, and Arash Noshadravan. “Deep learning for post-hurricane aerial damage assessment of buildings”. In: *Computer-Aided Civil and Infrastructure Engineering* (2021), pp. 1–16. ISSN: 14678667. DOI: 10.1111/mice.12658.
- [69] Shaif Choudhury, Soumyo Priyo Chattopadhyay, and Tapan Kumar Hazra. “Vehicle detection and counting using haar feature-based classifier”. In: *2017 8th Industrial Automation and Electromechanical Engineering Conference, IEMECON 2017* (2017), pp. 106–109. DOI: 10.1109/IEMECON.2017.8079571.
- [70] *Climate Change: How Do We Know?* <https://climate.nasa.gov/evidence/>. Accessed: 2021-10-29.
- [71] Nico Colic and Fabio Rinaldi. “Improving spaCy dependency annotation and PoS tagging web service using independent NER services”. In: *Genomics & informatics* 17.2 (2019).
- [72] Jill Collis and Roger Hussey. *Business research: A practical guide for undergraduate and postgraduate students*. Macmillan International Higher Education, 2013.
- [73] Aleksa Corovic et al. “The Real-Time Detection of Traffic Participants Using YOLO Algorithm”. In: *2018 26th Telecommunications Forum, TELFOR 2018 - Proceedings* (2018). DOI: 10.1109/TELFOR.2018.8611986.
- [74] Stefano Cresci et al. “A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages”. In: *Proceedings of the 24th International Conference on World Wide Web*. 2015, pp. 1195–1200. DOI: 10.1145/2740908.2741722.
- [75] John W Creswell. “The selection of a research approach”. In: *Research design: Qualitative, quantitative, and mixed methods approaches* (2014), pp. 3–24.
- [76] John W Creswell et al. “Best practices for mixed methods research in the health sciences”. In: *Bethesda (Maryland): National Institutes of Health* 2013 (2011), pp. 541–545.
- [77] Radu George Cretulescu et al. “DBSCAN Algorithm for Document Clustering”. In: *International Journal of Advanced Statistics and IT&C for Economics and Life Sciences* 9.1 (2019).
- [78] Michael Crotty. *The foundations of social research: Meaning and perspective in the research process*. Routledge, 2020.

- [79] Zhiyong Cui et al. “Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values”. In: *Transportation Research Part C: Emerging Technologies* 118 (2020), p. 102674.
- [80] Donna L. Cuomo and Charles D. Bowen. “Understanding Usability Issues Addressed by Three User-System Interface Evaluation Techniques”. In: *Interact. Comput.* 6.1 (1994), pp. 86–108. DOI: 10.1016/0953-5438(94)90006-X. URL: [https://doi.org/10.1016/0953-5438\(94\)90006-X](https://doi.org/10.1016/0953-5438(94)90006-X).
- [81] Jacob Danovitch. “Linking Social Media Posts to News with Siamese Transformers”. In: *arXiv preprint arXiv:2001.03303* (2020).
- [82] Swagatam Das, Shounak Datta, and Bidyut B. Chaudhuri. “Handling data irregularities in classification: Foundations, trends, and future challenges”. In: *Pattern Recognit.* 81 (2018), pp. 674–693. DOI: 10.1016/j.patcog.2018.03.008. URL: <https://doi.org/10.1016/j.patcog.2018.03.008>.
- [83] Leon Derczynski et al. “Helping crisis responders find the informative needle in the tweet haystack”. In: *arXiv preprint arXiv:1801.09633* (2018).
- [84] Heather Desurvire, Jim Kondziela, and Michael E Atwood. “What is gained and lost when using methods other than empirical testing”. In: *Posters and short talks of the 1992 SIGCHI conference on Human factors in computing systems*. 1992, pp. 125–126.
- [85] Ashwin Devaraj, Dhiraj Murthy, and Aman Dontula. “Machine-learning methods for identifying social media-based requests for urgent help during hurricanes”. In: *International Journal of Disaster Risk Reduction* 51 (2020), p. 101757. ISSN: 22124209. DOI: 10.1016/j.ijdr.2020.101757. URL: <https://doi.org/10.1016/j.ijdr.2020.101757>.
- [86] Nikhil Dhavase and AM Bagade. “Location identification for crime & disaster events by geoparsing Twitter”. In: *International Conference for Convergence for Technology-2014*. IEEE. 2014, pp. 1–3.
- [87] Barbara DiCicco-Bloom and Benjamin F Crabtree. “The qualitative research interview”. In: *Medical education* 40.4 (2006), pp. 314–321.
- [88] Bob Dick. “Postgraduate programs using action research”. In: *The learning organization* 9.4 (2002), pp. 159–170.
- [89] *Disaster*. <https://www.undrr.org/terminology/disaster>. Accessed: 2022-10-01.
- [90] *Disasters 2018: Year in Review*. <https://www.cred.be/publications>. Accessed: 2019-05-15.
- [91] *DOCTORAL THESIS WITH PUBLICATIONS GUIDELINES*. <https://bit.ly/38Qtzhc>. Accessed: 2021-09-15.
- [92] Chunjiao Dong et al. “Kalman filter algorithm for short-term jam traffic prediction based on traffic parameter correlation”. In: *Journal of Southeast University (Natural Science Edition)* 44.2 (2014), pp. 413–419.
- [93] RG Dromey. “Concerning the Chimera- software quality”. In: *IEEE Software* 13.1 (1996), pp. 33–43.

- [94] D Duarte et al. “SATELLITE IMAGE CLASSIFICATION OF BUILDING DAMAGES USING AIRBORNE AND SATELLITE IMAGE SAMPLES IN A DEEP LEARNING APPROACH.” In: *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences* 4.2 (2018).
- [95] Amnon H. Eden. “Three Paradigms of Computer Science”. In: *Minds Mach.* 17.2 (2007), pp. 135–167. DOI: 10.1007/s11023-007-9060-8. URL: <https://doi.org/10.1007/s11023-007-9060-8>.
- [96] Azadeh Emami, Majid Sarvi, and Saeed Asadi Bagloee. “Using Kalman filter algorithm for short-term traffic flow prediction in a connected vehicle environment”. In: *Journal of Modern Transportation* 27.3 (2019), pp. 222–232.
- [97] *Emergency Response and Recovery*. [https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/253488/Emergency\\_Response\\_and\\_Recovery\\_5th\\_edition\\_October\\_2013.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/253488/Emergency_Response_and_Recovery_5th_edition_October_2013.pdf). Accessed: 2021-04-30. 2013.
- [98] Mica R Endsley. “Toward a theory of situation awareness in dynamic systems”. In: *Human factors* 37.1 (1995), pp. 32–64.
- [99] Mica R Endsley, Betty Bolté, and Debra G Jones. *Designing for situation awareness: An approach to user-centered design*. CRC press, 2003.
- [100] Martin Ester et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.
- [101] Roger D Evered. “An Assessment of the Scientific Merits of Action Research Gerald 1. Susman and”. In: *Administrative science quarterly* 23.4 (1978), pp. 582–603.
- [102] Mark Everingham et al. “The pascal visual object classes challenge: A retrospective”. In: *International journal of computer vision* 111.1 (2015), pp. 98–136. DOI: 10.1007/s11263-014-0733-5.
- [103] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. “From Data Mining to Knowledge Discovery in Databases”. In: *AI Mag.* 17.3 (1996), pp. 37–54. DOI: 10.1609/aimag.v17i3.1230. URL: <https://doi.org/10.1609/aimag.v17i3.1230>.
- [104] Aleksandr Fedorov et al. “Traffic flow estimation with data from a video surveillance camera”. In: *Journal of Big Data* 6.1 (2019), p. 73. DOI: 10.1186/s40537-019-0234-z.
- [105] Xinxin Feng et al. “Adaptive multi-kernel SVM with spatial-temporal correlation for short-term traffic flow prediction”. In: *IEEE Transactions on Intelligent Transportation Systems* 20.6 (2018), pp. 2001–2013.
- [106] Jean-Christophe Filliâtre. “Formal proof of a program: Find”. In: *Sci. Comput. Program.* 64.3 (2007), pp. 332–340. DOI: 10.1016/j.scico.2006.10.002. URL: <https://doi.org/10.1016/j.scico.2006.10.002>.
- [107] Gian Luca Foresti, Manuela Farinosi, and Marco Vernier. “Situational awareness in smart environments: socio-mobile and sensor data fusion for emergency response to disasters”. In: *J. Ambient Intell. Humaniz. Comput.* 6.2 (2015), pp. 239–257. DOI: 10.1007/s12652-014-0227-x. URL: <https://doi.org/10.1007/s12652-014-0227-x>.
- [108] Huiji Gao, Geoffrey Barbier, and Rebecca Goolsby. “Harnessing the crowdsourcing power of social media for disaster relief”. In: *IEEE Intelligent Systems* 26.3 (2011), pp. 10–14.

- [109] Wang Gao et al. “Detecting Disaster-Related Tweets Via Multimodal Adversarial Neural Network”. In: *IEEE Multimedia* 27.4 (2020), pp. 28–37. ISSN: 19410166. DOI: 10.1109/MMUL.2020.3012675.
- [110] Windu Gata et al. “Informative Tweet Classification of the Earthquake Disaster Situation In Indonesia”. In: *2019 5th International Conference on Computing Engineering and Design (ICCED)*. IEEE, 2019, pp. 1–6.
- [111] Kent Gauen et al. “Comparison of visual datasets for machine learning”. In: *2017 IEEE International Conference on Information Reuse and Integration (IRI)*. IEEE, 2017, pp. 346–355. DOI: 10.1109/IRI.2017.59.
- [112] Saman Ghaffarian et al. “Post-disaster building database updating using automated deep learning: An integration of pre-disaster OpenStreetMap and multi-temporal satellite data”. In: *Remote Sensing* 11.20 (2019), pp. 1–20. ISSN: 20724292. DOI: 10.3390/rs11202427.
- [113] Shalmoli Ghosh et al. “Identifying Multi-Dimensional Information from Microblogs During Epidemics”. In: *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. 2019, pp. 224–230. DOI: 10.1145/3297001.3297030.
- [114] Tarutal Ghosh Mondal et al. “Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance”. In: *Structural Control and Health Monitoring* 27.4 (2020), pp. 1–15. ISSN: 15452263. DOI: 10.1002/stc.2507.
- [115] Ross Girshick. “Fast r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1440–1448. DOI: 10.1109/ICCV.2015.169.
- [116] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587. DOI: 10.1109/CVPR.2014.81.
- [117] Christopher Gomez and Heather Purdie. “UAV-based photogrammetry and geocomputing for hazards and disaster risk monitoring—a review”. In: *Geoenvironmental Disasters* 3.1 (2016), p. 23.
- [118] David Graf et al. “Cross-domain informativeness classification for disaster situations”. In: *Proceedings of the 10th international conference on management of digital ecosystems*. 2018, pp. 183–190. DOI: 10.1145/3281375.3281385.
- [119] Maria J Grant and Andrew Booth. “A typology of reviews: an analysis of 14 review types and associated methodologies”. In: *Health Information & Libraries Journal* 26.2 (2009), pp. 91–108.
- [120] Weiwei Guo et al. “Linking tweets to news: A framework to enrich short text data in social media”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2013, pp. 239–249.
- [121] Anchal Gupta, Monika Rani, and Sakshi Kaushal. “Disaster Event Detection from Text: A Survey”. In: *Computational Intelligence in Data Mining*. Springer, 2022, pp. 281–293.
- [122] Dhruv Gupta, Jannik Strötgen, and Klaus Berberich. “Eventminer: Mining events from annotated documents”. In: *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*. 2016, pp. 261–270.

- [123] Hyunsoo Ha and Byung-Yeon Hwang. “Keyword filtering about disaster and the method of detecting area in detecting real-time event using twitter”. In: *KIPS Transactions on Software and Data Engineering* 5.7 (2016), pp. 345–350.
- [124] Mohamed Hagra, Ghada Hassan, and Nadine Farag. “Towards natural disasters detection from Twitter using topic modelling”. In: *2017 European Conference on Electrical Engineering and Computer Science (EECS)*. IEEE, 2017, pp. 272–279.
- [125] Felix Hamborg, Corinna Breiting, and Bela Gipp. “Giveme5W1H: A Universal System for Extracting Main Events from News Articles”. In: *arXiv preprint arXiv:1909.02766* (2019).
- [126] Felix Hamborg, Corinna Breiting, and Bela Gipp. “Giveme5W1H: A Universal System for Extracting Main Events from News Articles”. In: *Proceedings of the 7th International Workshop on News Recommendation and Analytics in conjunction with 13th ACM Conference on Recommender Systems, INRA@RecSys 2019, Copenhagen, Denmark, September 20, 2019*. Ed. by Özlem Özgöbek et al. Vol. 2554. CEUR Workshop Proceedings. CEUR-WS.org, 2019, pp. 35–43. URL: [http://ceur-ws.org/Vol-2554/paper%5C\\_06.pdf](http://ceur-ws.org/Vol-2554/paper%5C_06.pdf).
- [127] Felix Hamborg et al. “Giveme5W: Main Event Retrieval from News Articles by Extraction of the Five Journalistic W Questions”. In: *Transforming Digital Worlds - 13th International Conference, iConference 2018, Sheffield, UK, March 25-28, 2018, Proceedings*. Ed. by Gobinda Chowdhury et al. Vol. 10766. Lecture Notes in Computer Science. Springer, 2018, pp. 356–366. DOI: 10.1007/978-3-319-78105-1\_39. URL: [https://doi.org/10.1007/978-3-319-78105-1\\_39](https://doi.org/10.1007/978-3-319-78105-1_39).
- [128] Felix Hamborg et al. “Giveme5W: main event retrieval from news articles by extraction of the five journalistic w questions”. In: *International Conference on Information*. Springer, 2018, pp. 356–366.
- [129] Xuehua Han and Juanle Wang. “Earthquake Information Extraction and Comparison from Different Sources Based on Web Text”. In: *ISPRS International Journal of Geo-Information* 8.6 (2019), p. 252.
- [130] Xuehua Han and Juanle Wang. “Using social media to mine and analyze public sentiment during a disaster: A case study of the 2018 Shouguang city flood in china”. In: *ISPRS International Journal of Geo-Information* 8.4 (2019), p. 185.
- [131] Haiyan Hao and Yan Wang. “Leveraging multimodal social media data for rapid disaster damage assessment”. In: *International Journal of Disaster Risk Reduction* 51 (2020), p. 101760.
- [132] Haiyan Hao and Yan Wang. “Leveraging multimodal social media data for rapid disaster damage assessment”. In: *International Journal of Disaster Risk Reduction* 51.July (2020), p. 101760. ISSN: 22124209. DOI: 10.1016/j.ijdr.2020.101760. URL: <https://doi.org/10.1016/j.ijdr.2020.101760>.
- [133] Haiyan Hao and Yan Wang. “Leveraging multimodal social media data for rapid disaster damage assessment”. In: *International Journal of Disaster Risk Reduction* 51.March (2020), p. 101760. ISSN: 22124209. DOI: 10.1016/j.ijdr.2020.101760. URL: <https://doi.org/10.1016/j.ijdr.2020.101760>.

- [134] Dean Rizki Hartawan, Tito Waluyo Purboyo, and Casi Setianingsih. “Disaster victims detection system using convolutional neural network (CNN) method”. In: *Proceedings - 2019 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology, IAICT 2019* (2019), pp. 105–111. DOI: 10.1109/ICIAICT.2019.8784782.
- [135] Elizabeth Harten et al. “Evaluation of traffic mitigation strategies for pre-hurricane emergency evacuations”. In: *2018 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 2018, pp. 214–219.
- [136] Douglas M Hawkins. “The problem of overfitting”. In: *Journal of chemical information and computer sciences* 44.1 (2004), pp. 1–12.
- [137] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [138] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969. DOI: 10.1109/TPAMI.2018.2844175.
- [139] Lisa Anne Hendricks et al. “Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 1–10. DOI: 10.1109/CVPR.2016.8. URL: <https://doi.org/10.1109/CVPR.2016.8>.
- [140] Aldo Hernandez-Suarez et al. “Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation”. In: *Sensors (Switzerland)* 19.7 (2019). ISSN: 14248220. DOI: 10.3390/s19071746.
- [141] Marlien Herselman and Adele Botha. “Evaluating an Artifact in Design Science Research”. In: *Proceedings of the 2015 Annual Research Conference on South African Institute of Computer Scientists and Information Technologists, SAICSIT '15, Stellenbosch, South Africa, September 28-30, 2015*. Ed. by Richard J. Barnett et al. ACM, 2015, 21:1–21:10. DOI: 10.1145/2815782.2815806. URL: <https://doi.org/10.1145/2815782.2815806>.
- [142] Alan R Hevner. “A three cycle view of design science research”. In: *Scandinavian journal of information systems* 19.2 (2007), p. 4.
- [143] Alan R Hevner and Salvatore T March. “The information systems research cycle”. In: *Computer* 36.11 (2003), pp. 111–113.
- [144] Mahshad Mahdavi Hezaveh, Christopher Kanan, and Carl Salvaggio. “Roof damage assessment using deep learning”. In: *Proceedings - Applied Imagery Pattern Recognition Workshop 2017-Octob* (2018), pp. 6403–6408. ISSN: 21642516. DOI: 10.1109/AIPR.2017.8457946.
- [145] Starr Roxanne Hiltz and Linda Plotnick. “Dealing with information overload when using social media for emergency management: Emerging solutions”. In: *10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013*. Ed. by Tina Comes et al. ISCRAM Association, 2013. URL: [http://idl.iscram.org/files/hiltz/2013/583%5C\\_Hiltz+Plotnick2013.pdf](http://idl.iscram.org/files/hiltz/2013/583%5C_Hiltz+Plotnick2013.pdf).
- [146] Wei-Chiang Hong et al. “Hybrid evolutionary algorithms in a SVR traffic flow forecasting model”. In: *Applied Mathematics and Computation* 217.15 (2011), pp. 6733–6747.

- [147] Yue Hou, Zhiyuan Deng, and Hanke Cui. “Short-Term Traffic Flow Prediction with Weather Conditions: Based on Deep Learning Algorithms and Data Fusion”. In: *Complex*. 2021 (2021), 6662959:1–6662959:14. DOI: 10.1155/2021/6662959. URL: <https://doi.org/10.1155/2021/6662959>.
- [148] *Households, communities stunned by storm and flood damage*. <https://www.rnz.co.nz/news/national/473223/households-communities-stunned-by-storm-and-flood-damage>. Accessed: 2022-10-12.
- [149] Wenbin Hu et al. “Pso-svr: A hybrid short-term traffic flow forecasting method”. In: *2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE. 2015, pp. 553–561.
- [150] Jing Huang and Brian Kingsbury. “Audio-visual deep learning for noise robust speech recognition”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 7596–7599.
- [151] Jonathan Huang et al. “Speed/accuracy trade-offs for modern convolutional object detectors”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7310–7311.
- [152] Xiao Huang et al. “A visual-textual fused approach to automated tagging of flood-related tweets during a flood event”. In: *International Journal of Digital Earth* 12.11 (2019), pp. 1248–1264. ISSN: 17538955. DOI: 10.1080/17538947.2018.1523956.
- [153] Xiao Huang et al. “Identifying disaster related social media for rapid response: a visual-textual fused CNN architecture”. In: *International Journal of Digital Earth* 13.9 (2020), pp. 1017–1039. ISSN: 17538955. DOI: 10.1080/17538947.2019.1633425.
- [154] Earnest Paul Ijjina et al. “Computer Vision-based Accident Detection in Traffic Surveillance”. In: *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE. 2019, pp. 1–6. DOI: 10.1109/ICCCNT45670.2019.8944469.
- [155] Hyeongsun Im et al. “Bigdata analytics on CCTV images for collecting traffic information”. In: *2016 International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2016, pp. 525–528.
- [156] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. “Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. Ed. by Nicoletta Calzolari et al. European Language Resources Association (ELRA), 2016. URL: <http://www.lrec-conf.org/proceedings/lrec2016/summaries/842.html>.
- [157] Muhammad Imran, Prasenjit Mitra, and Jaideep Srivastava. “Cross-Language Domain Adaptation for Classifying Crisis-Related Short Messages”. In: *13th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Rio de Janeiro, Brasil, May 22-25, 2016*. ISCRAM Association, 2016. URL: [http://idl.iscram.org/files/muhammadimran/2016/%201396%5C\\_MuhammadImran%5C\\_etal2016.pdf](http://idl.iscram.org/files/muhammadimran/2016/%201396%5C_MuhammadImran%5C_etal2016.pdf).
- [158] Muhammad Imran et al. “AIDR: Artificial intelligence for disaster response”. In: *Proceedings of the 23rd International Conference on World Wide Web*. ACM. 2014, pp. 159–162.

- [159] Muhammad Imran et al. “Extracting information nuggets from disaster- Related messages in social media”. In: *10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013*. Ed. by Tina Comes et al. ISCRAM Association, 2013. URL: [http://idl.iscram.org/files/imran/2013/613%5C\\_Imran%5C\\_etal2013.pdf](http://idl.iscram.org/files/imran/2013/613%5C_Imran%5C_etal2013.pdf).
- [160] Muhammad Imran et al. “Extracting information nuggets from disaster-Related messages in social media.” In: *Iscram*. 2013.
- [161] Muhammad Imran et al. “Practical extraction of disaster-relevant information from social media”. In: *Proceedings of the 22nd International Conference on World Wide Web*. 2013, pp. 1021–1024.
- [162] Muhammad Imran et al. “Processing Social Media Messages in Mass Emergency: A Survey”. In: *ACM Comput. Surv.* 47.4 (2015), 67:1–67:38. DOI: 10.1145/2771588. URL: <https://doi.org/10.1145/2771588>.
- [163] Muhammad Imran et al. “Processing social media messages in mass emergency: A survey”. In: *ACM Computing Surveys (CSUR)* 47.4 (2015), p. 67.
- [164] Muhammad Imran et al. “Rapid Damage Assessment Using Social Media Images by Combining Human and Machine Intelligence”. In: *arXiv preprint arXiv:2004.06675* (2020).
- [165] Muhammad Imran et al. “Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions”. In: *Inf. Process. Manag.* 57.5 (2020), p. 102261. DOI: 10.1016/j.ipm.2020.102261. URL: <https://doi.org/10.1016/j.ipm.2020.102261>.
- [166] Muhammad Imran et al. *Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions*. 2020.
- [167] K Indira, KV Mohan, and Theegalapally Nikhilashwary. “Automatic license plate recognition”. In: *Recent Trends in Signal and Image Processing*. Springer, 2019, pp. 67–77. DOI: 10.1007/978-981-10-8863-6\_8.
- [168] Roberto Interdonato, Jean-Loup Guillaume, and Antoine Doucet. “A lightweight and multilingual framework for crisis information extraction from Twitter data”. In: *Social Network Analysis and Mining* 9.1 (2019), p. 65.
- [169] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.
- [170] Iso Iso and IEC Std. “9126 Software product evaluation–quality characteristics and guidelines for their use”. In: *ISO/IEC Standard 9126* (2001).
- [171] H Jabbar and Rafiqul Zaman Khan. “Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study)”. In: *Computer Science, Communication and Instrumentation Devices* (2015), pp. 163–172.
- [172] Beakcheol Jang, Inhwan Kim, and Jong Wook Kim. “Word2vec convolutional neural networks for classification of news articles and tweets”. In: *PloS one* 14.8 (2019), e0220976.
- [173] Nathalie Japkowicz. “The class imbalance problem: Significance and strategies”. In: *Proc. of the Int’l Conf. on Artificial Intelligence*. Vol. 56. Citeseer. 2000.



- [174] Robin Jeffries et al. “User interface evaluation in the real world: a comparison of four techniques”. In: *Conference on Human Factors in Computing Systems, CHI 1991, New Orleans, LA, USA, April 27 - May 2, 1991, Proceedings*. Ed. by Scott P. Robertson, Gary M. Olson, and Judith S. Olson. ACM, 1991, pp. 119–124. DOI: 10.1145/108844.108862. URL: <https://doi.org/10.1145/108844.108862>.
- [175] Paul Johannesson and Erik Perjons. *An Introduction to Design Science*. Springer, 2014. ISBN: 978-3-319-10631-1. DOI: 10.1007/978-3-319-10632-8. URL: <https://doi.org/10.1007/978-3-319-10632-8>.
- [176] Scott Jones and Jafar Saniie. “Using Deep Learning and Satellite Imagery to Assess the Damage to Civil Structures After Natural Disasters”. In: *2019 IEEE International Conference on Electro Information Technology (EIT)*. IEEE. 2019, pp. 189–193. DOI: 10.1109/EIT.2019.8833724.
- [177] Scott Jones and Jafar Saniie. “Using deep learning and satellite imagery to assess the damage to civil structures after natural disasters”. In: *IEEE International Conference on Electro Information Technology 2019-May (2019)*, pp. 189–193. ISSN: 21540373. DOI: 10.1109/EIT.2019.8833724.
- [178] Jay F. Nunamaker Jr., Minder Chen, and Titus D. M. Purdin. “Systems Development in Information Systems Research”. In: *J. Manag. Inf. Syst.* 7.3 (1991), pp. 89–106. URL: <http://www.jmis-web.org/articles/731>.
- [179] Md Kabir, Sanjay Madria, et al. “A Deep Learning Approach for Tweet Classification and Rescue Scheduling for Effective Disaster Management”. In: *arXiv preprint arXiv:1908.01456* (2019).
- [180] Yasin Kabir and Sanjay Madria. “A Deep Learning Approach for Tweet Classification and Rescue Scheduling for Effective Disaster Management (Industrial)”. In: *arXiv* (2019). ISSN: 23318422.
- [181] Janani Kalyanam et al. “Prediction and characterization of high-activity events in social media triggered by real-world news”. In: *PloS one* 11.12 (2016), e0166694.
- [182] Yiannis Kamarianakis and Poulicos Prastacos. “Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches”. In: *Transportation Research Record* 1857.1 (2003), pp. 74–84.
- [183] Danqing Kang, Yisheng Lv, and Yuan-yuan Chen. “Short-term traffic flow prediction with LSTM recurrent neural network”. In: *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*. IEEE. 2017, pp. 1–6.
- [184] Nayomi Kankanamge et al. “Can volunteer crowdsourcing reduce disaster risk? A systematic review of the literature”. In: *International Journal of Disaster Risk Reduction* (2019), p. 101097.
- [185] Amir Karami et al. “Twitter speaks: A case of national disaster situational awareness”. In: *Journal of Information Science* 46.3 (2020), pp. 313–324.
- [186] Marc-André Kaufhold, Markus Bayer, and Christian Reuter. “Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning”. In: *Information Processing & Management* 57.1 (2020), p. 102132. DOI: 10.1016/j.ipm.2019.102132.

- [187] Ruimin Ke et al. “Real-time traffic flow parameter estimation from UAV video based on ensemble classifier and optical flow”. In: *IEEE Transactions on Intelligent Transportation Systems* 20.1 (2018), pp. 54–64. DOI: 10.1109/TITS.2018.2797697.
- [188] Jaana Kekäläinen and Kalervo Järvelin. “Using graded relevance assessments in IR evaluation”. In: *J. Assoc. Inf. Sci. Technol.* 53.13 (2002), pp. 1120–1129. DOI: 10.1002/asi.10137. URL: <https://doi.org/10.1002/asi.10137>.
- [189] Salman H. Khan et al. “Forest Change Detection in Incomplete Satellite Images with Deep Neural Networks”. In: *IEEE Transactions on Geoscience and Remote Sensing* 55.9 (2017), pp. 5407–5423. ISSN: 01962892. DOI: 10.1109/TGRS.2017.2707528.
- [190] Sameer Khan and Suet-Peng Yong. “A comparison of deep learning and hand crafted features in medical image modality classification”. In: *2016 3rd International Conference on Computer and Information Sciences (ICCOINS)*. IEEE, 2016, pp. 633–638.
- [191] Prashant Khare, Grégoire Burel, and Harith Alani. “Classifying crises-information relevancy with semantics”. In: *European Semantic Web Conference*. Springer, 2018, pp. 367–383. DOI: 10.1007/978-3-319-93417-4\_24.
- [192] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. “M2m: Imbalanced Classification via Major-to-Minor Translation”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 13893–13902. DOI: 10.1109/CVPR42600.2020.01391.
- [193] Yoon Kim. “Convolutional Neural Networks for Sentence Classification”. In: *arXiv preprint arXiv:1408.5882* (2014). DOI: 10.3115/v1/D14-1181.
- [194] Arief Koedwiady, Ridha Soua, and Fakhreddine Karray. “Improving Traffic Flow Prediction With Weather Information in Connected Cars: A Deep Learning Approach”. In: *IEEE Trans. Veh. Technol.* 65.12 (2016), pp. 9508–9517. DOI: 10.1109/TVT.2016.2585575. URL: <https://doi.org/10.1109/TVT.2016.2585575>.
- [195] Mehrdad Koohikamali and Dan J. Kim. “Rumor and Truth Spreading Patterns on Social Network Sites During Social Crisis: Big Data Analytics Approach”. In: *E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life - 15th Workshop on e-Business, WEB 2015, Fort Worth, Texas, USA, December 12, 2015, Revised Selected Papers*. Ed. by Vijayan Sugumaran, Victoria Y. Yoon, and Michael J. Shaw. Vol. 258. Lecture Notes in Business Information Processing. Springer, 2015, pp. 166–170. DOI: 10.1007/978-3-319-45408-5\_15. URL: [https://doi.org/10.1007/978-3-319-45408-5\\_15](https://doi.org/10.1007/978-3-319-45408-5_15).
- [196] Mehrdad Koohikamali and Dan J. Kim. “Rumor and Truth Spreading Patterns on Social Network Sites During Social Crisis: Big Data Analytics Approach”. In: *E-Life: Web-Enabled Convergence of Commerce, Work, and Social Life - 15th Workshop on e-Business, WEB 2015, Fort Worth, Texas, USA, December 12, 2015, Revised Selected Papers*. Ed. by Vijayan Sugumaran, Victoria Y. Yoon, and Michael J. Shaw. Vol. 258. Lecture Notes in Business Information Processing. Springer, 2015, pp. 166–170. DOI: 10.1007/978-3-319-45408-5\_15. URL: [https://doi.org/10.1007/978-3-319-45408-5\\_15](https://doi.org/10.1007/978-3-319-45408-5_15).
- [197] Anna Kruspe, Jens Kersten, and Friederike Klan. “Detection of informative tweets in crisis events”. In: *Natural Hazards and Earth System Sciences Discussions* (2020), pp. 1–18.

- [198] Anna Kruspe, Jens Kersten, and Friederike Klan. “Detection of informative tweets in crisis events”. In: *Natural Hazards and Earth System Sciences (NHESS)* (2021).
- [199] Abhinav Kumar and Jyoti Prakash Singh. “Disaster severity prediction from Twitter images”. In: *Advances in Intelligent Systems and Computing* 1279. December 2020 (2021), pp. 65–73. ISSN: 21945365. DOI: 10.1007/978-981-15-9290-4\_7.
- [200] Abhinav Kumar and Jyoti Prakash Singh. “Location reference identification from tweets during emergencies: A deep learning approach”. In: *International Journal of Disaster Risk Reduction* 33 (2019), pp. 365–375. ISSN: 22124209. DOI: 10.1016/j.ijdr.2018.10.021. arXiv: 1901.08241.
- [201] Abhinav Kumar, Jyoti Prakash Singh, and Sunil Saumya. “A Comparative Analysis of Machine Learning Techniques for Disaster-Related Tweet Classification”. In: *2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129)*. IEEE. 2019, pp. 222–227. DOI: 10.1109/R10-HTC47129.2019.9042443.
- [202] Abhinav Kumar et al. *A deep multi-modal neural network for informative Twitter content classification during emergencies*. 0123456789. Springer US, 2020. ISBN: 0123456789. DOI: 10.1007/s10479-020-03514-x. URL: <https://doi.org/10.1007/s10479-020-03514-x>.
- [203] Pakhee Kumar et al. “Detection of disaster-affected cultural heritage sites from social media images using deep learning techniques”. In: *Journal on Computing and Cultural Heritage* 13.3 (2020). ISSN: 15564711. DOI: 10.1145/3383314.
- [204] Sakitha P Kumarage et al. “Traffic Flow Estimation for Urban Roads Based on Crowdsourced Data and Machine Learning Principles”. In: *First International Conference on Intelligent Transport Systems*. Springer. 2017, pp. 263–273. DOI: 10.1007/978-3-319-93710-6\_27.
- [205] Felix Kunde et al. “Traffic Prediction using a Deep Learning Paradigm.” In: *EDBT/ICDT Workshops*. 2017.
- [206] Shamik Kundu, P. K. Srijith, and Maunendra Sankar Desarkar. “Classification of short-texts generated during disasters: A deep neural network based approach”. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018* (2018), pp. 790–793. DOI: 10.1109/ASONAM.2018.8508695.
- [207] Jason Kurniawan, Sensa GS Syahra, Chandra K Dewa, et al. “Traffic Congestion Detection: Learning from CCTV Monitoring Images using Convolutional Neural Network”. In: *Procedia computer science* 144 (2018), pp. 291–297. DOI: 10.1016/j.procs.2018.10.530.
- [208] Samia Laghrabli, Loubna Benabbou, and Abdelaziz Berrado. “A new methodology for literature review analysis using association rules mining”. In: *10th International Conference on Intelligent Systems: Theories and Applications, SITA 2015, Rabat, Morocco, October 20-21, 2015*. IEEE, 2015, pp. 1–6. DOI: 10.1109/SITA.2015.7358394. URL: <https://doi.org/10.1109/SITA.2015.7358394>.
- [209] Rayson Laroca et al. “A robust real-time automatic license plate recognition based on the YOLO detector”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–10. DOI: 10.1109/IJCNN.2018.8489629.
- [210] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “Deep learning”. In: *nature* 521.7553 (2015), p. 436.

- [211] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. “Deep learning”. In: *Nat.* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539. URL: <https://doi.org/10.1038/nature14539>.
- [212] Kyungeun Lee et al. “Short-Term Traffic Prediction With Deep Neural Networks: A Survey”. In: *IEEE Access* 9 (2021), pp. 54739–54756. DOI: 10.1109/ACCESS.2021.3071174. URL: <https://doi.org/10.1109/ACCESS.2021.3071174>.
- [213] Sangsoo Lee and Daniel B Fambro. “Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting”. In: *Transportation Research Record* 1678.1 (1999), pp. 179–188.
- [214] Moshe Levin and Yen-Der Tsao. “On forecasting freeway occupancies and volumes (abridgment)”. In: *Transportation Research Record* 773 (1980).
- [215] Clayton Lewis and John Rieman. “Task-centered user interface design”. In: *A practical introduction* (1993).
- [216] Hongmin Li et al. “Twitter Mining for Disaster Response: A Domain Adaptation Approach.” In: *ISCRAM*. 2015.
- [217] Xukun Li and Doina Caragea. “Improving Disaster-related Tweet Classification with a Multimodal Approach”. In: *Social Media for Disaster Response and Resilience Proceedings of the 17th ISCRAM Conference* May (2020), pp. 893–902.
- [218] Xukun Li et al. “Identifying disaster damage images using a domain adaptation approach”. In: *Proceedings of the International ISCRAM Conference 2019-May*. May 2019 (2019), pp. 633–645. ISSN: 24113387.
- [219] Xukun Li et al. “Localizing and quantifying damage in social media images”. In: *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018* (2018), pp. 194–201. DOI: 10.1109/ASONAM.2018.8508298. arXiv: 1806.07378.
- [220] Yingjie Li et al. “Sympathy Detection in Disaster Twitter Data”. In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response And Management* (2019).
- [221] Yundong Li, Shi Ye, and Ivan Bartoli. “Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning”. In: *Journal of Applied Remote Sensing* 12.4 (2018), p. 045008.
- [222] Yundong Li, Shi Ye, and Ivan Bartoli. “Semisupervised classification of hurricane damage from postevent aerial imagery using deep learning”. In: *Journal of Applied Remote Sensing* 12.04 (2018), p. 1. ISSN: 1931-3195. DOI: 10.1117/1.jrs.12.045008.
- [223] Xiao Liang. “Image-based post-disaster inspection of reinforced concrete bridge systems using deep learning with Bayesian optimization”. In: *Computer-Aided Civil and Infrastructure Engineering* 34.5 (2019), pp. 415–430. ISSN: 14678667. DOI: 10.1111/mice.12425.

- [224] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*. Ed. by David J. Fleet et al. Vol. 8693. Lecture Notes in Computer Science. Springer, 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1\_48. URL: [https://doi.org/10.1007/978-3-319-10602-1\\_48](https://doi.org/10.1007/978-3-319-10602-1_48).
- [225] Tsung-Yi Lin et al. “Microsoft coco: Common objects in context”. In: *European conference on computer vision*. Springer. 2014, pp. 740–755. DOI: 10.1007/978-3-319-10602-1\_48.
- [226] Xian Yao Ling et al. “Short-term traffic flow prediction with optimized multi-kernel support vector machine”. In: *2017 IEEE Congress on Evolutionary Computation (CEC)*. IEEE. 2017, pp. 294–300.
- [227] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37. DOI: 10.1007/978-3-319-46448-0\_2.
- [228] Ying Liu and Linzhi Wu. “Geological Disaster Recognition on Optical Remote Sensing Images Using Deep Learning”. In: *Procedia Computer Science* 91.Itqm (2016), pp. 566–575. ISSN: 18770509. DOI: 10.1016/j.procs.2016.07.144. URL: <http://dx.doi.org/10.1016/j.procs.2016.07.144>.
- [229] Ying Liu and Linzhi Wu. “High Performance Geological Disaster Recognition using Deep Learning”. In: *Procedia Computer Science* 139 (2018), pp. 529–536. ISSN: 18770509. DOI: 10.1016/j.procs.2018.10.237. URL: <https://doi.org/10.1016/j.procs.2018.10.237>.
- [230] Yipeng Liu et al. “Short-term traffic flow prediction with Conv-LSTM”. In: *2017 9th International Conference on Wireless Communications and Signal Processing, WCSP 2017 - Proceedings 2017-Janua* (2017), pp. 1–6. DOI: 10.1109/WCSP.2017.8171119.
- [231] Karen D Locke. *Grounded theory in management research*. Sage, 2000.
- [232] Kanishk Lohumi and Sudip Roy. “Automatic Detection of Flood Severity Level from Flood Videos using Deep Learning Models”. In: *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*. IEEE. 2018, pp. 1–7.
- [233] Kanishk Lohumi and Sudip Roy. “Automatic Detection of Flood Severity Level from Flood Videos using Deep Learning Models”. In: *2018 5th International Conference on Information and Communication Technologies for Disaster Management, ICT-DM 2018* (2019), pp. 1–7. DOI: 10.1109/ICT-DM.2018.8636373.
- [234] Markus Lucking et al. “A video-based vehicle counting system using an embedded device in realistic traffic conditions”. In: *IEEE World Forum on Internet of Things, WF-IoT 2020 - Symposium Proceedings* (2020), pp. 1–6. DOI: 10.1109/WF-IoT48130.2020.9221094.
- [235] Sergio Luna and Michael J Pennock. “Social media applications and emergency management: A literature review and research agenda”. In: *International journal of disaster risk reduction* 28 (2018), pp. 565–577.
- [236] Yisheng Lv et al. “Traffic flow prediction with big data: a deep learning approach”. In: *IEEE Transactions on Intelligent Transportation Systems* 16.2 (2014), pp. 865–873.

- [237] Sreenivasulu Madichetty and Sridevi Muthukumarasamy. “Detection of situational information from Twitter during disaster using deep learning models”. In: *Sadhana - Academy Proceedings in Engineering Sciences* 45.1 (2020), pp. 1–13. ISSN: 09737677. DOI: 10.1007/s12046-020-01504-0. URL: <https://doi.org/10.1007/s12046-020-01504-0>.
- [238] Sreenivasulu Madichetty, Sridevi Muthukumarasamy, and P. Jayadev. “Multi-modal classification of Twitter data during disasters for humanitarian response”. In: *Journal of Ambient Intelligence and Humanized Computing* 0123456789 (2021). ISSN: 18685145. DOI: 10.1007/s12652-020-02791-5. URL: <https://doi.org/10.1007/s12652-020-02791-5>.
- [239] Sreenivasulu Madichetty and M Sridevi. “Detecting informative tweets during disaster using deep neural networks”. In: *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE. 2019, pp. 709–713. DOI: 10.1109/COMSNETS.2019.8711095.
- [240] Sreenivasulu Madichetty and M. Sridevi. “A stacked convolutional neural network for detecting the resource tweets during a disaster”. In: *Multimedia Tools and Applications* 80.3 (2021), pp. 3927–3949. ISSN: 15737721. DOI: 10.1007/s11042-020-09873-8.
- [241] Sreenivasulu Madichetty and M. Sridevi. “Classifying informative and non-informative tweets from the twitter by adapting image features during disaster”. In: *Multimedia Tools and Applications* 79.39-40 (2020), pp. 28901–28923. ISSN: 15737721. DOI: 10.1007/s11042-020-09343-1.
- [242] Sreenivasulu Madichetty and M. Sridevi. “Detecting Informative Tweets during Disaster using Deep Neural Networks”. In: *2019 11th International Conference on Communication Systems and Networks, COMSNETS 2019* 2061 (2019), pp. 709–713. DOI: 10.1109/COMSNETS.2019.8711095.
- [243] Tomas Majtner, Sule Yildirim-Yayilgan, and Jon Yngve Hardeberg. “Combining deep learning and hand-crafted features for skin lesion classification”. In: *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2016, pp. 1–6.
- [244] Sujith Mangalathu and Henry V. Burton. “Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions”. In: *International Journal of Disaster Risk Reduction* 36.1538866 (2019). ISSN: 22124209. DOI: 10.1016/j.ijdr.2019.101111.
- [245] Salvatore T. March and Gerald F. Smith. “Design and natural science research on information technology”. In: *Decis. Support Syst.* 15.4 (1995), pp. 251–266. DOI: 10.1016/0167-9236(94)00041-2. URL: [https://doi.org/10.1016/0167-9236\(94\)00041-2](https://doi.org/10.1016/0167-9236(94)00041-2).
- [246] Adrian Thornhill Mark NK Saunders Philip Lewis. *Research methods for business students, 7/e*. Pearson Education India, 2015.
- [247] Jim A McCall, Paul K Richards, and Gene F Walters. *Factors in software quality. volume i. concepts and definitions of software quality*. Tech. rep. GENERAL ELECTRIC CO SUNNYVALE CA, 1977.
- [248] Mary L McHugh. “Interrater reliability: the kappa statistic”. In: *Biochemia medica: Biochemia medica* 22.3 (2012), pp. 276–282.

- [249] Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. “Building a large-scale corpus for evaluating event detection on twitter”. In: *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2013, pp. 409–418. DOI: 10.1145/2505515.2505695.
- [250] Jesús Mena-Oreja and Javier Gozávez. “A Comprehensive Evaluation of Deep Learning-Based Techniques for Traffic Prediction”. In: *IEEE Access* 8 (2020), pp. 91188–91212. DOI: 10.1109/ACCESS.2020.2994415. URL: <https://doi.org/10.1109/ACCESS.2020.2994415>.
- [251] Christian Meurisch et al. “Enhanced Detection of Crisis-Related Microblogs by Spatiotemporal Feedback Loops”. In: *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*. Vol. 1. IEEE. 2019, pp. 507–512. DOI: 10.1109/COMPSAC.2019.00078.
- [252] Jose P. Miguel, David Mauricio, and Glen Rodriguez. “A Review of Software Quality Models for the Evaluation of Software Products”. In: *CoRR* abs/1412.2977 (2014). arXiv: 1412.2977. URL: <http://arxiv.org/abs/1412.2977>.
- [253] Marko Mijač. “Evaluation of Design Science instantiation artifacts in Software engineering research”. In: *Central European Conference on Information and Intelligent Systems*. Faculty of Organization and Informatics Varazdin. 2019, pp. 313–321.
- [254] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.
- [255] Tomas Mikolov et al. “Efficient estimation of word representations in vector space”. In: *arXiv preprint arXiv:1301.3781* (2013).
- [256] Zeeshan Hameed Mir and Fethi Filali. “An adaptive Kalman filter based traffic prediction algorithm for urban road network”. In: *2016 12th International Conference on Innovations in Information Technology (IIT)*. IEEE. 2016, pp. 1–6.
- [257] Hiroyuki Miura, Tomohiro Aridome, and Masashi Matsuoka. “Deep learning-based identification of collapsed, non-collapsed and blue tarp-covered buildings from post-disaster aerial images”. In: *Remote Sensing* 12.12 (2020). ISSN: 20724292. DOI: 10.3390/rs12121924.
- [258] S Moechammad, R Cahya, and A N A Berkah. “Detecting body parts from natural disaster victims using You Only Look Once (YOLO)”. In: *IOP Conference Series: Materials Science and Engineering* 1073.1 (2021), p. 012062. ISSN: 1757-8981. DOI: 10.1088/1757-899x/1073/1/012062.
- [259] Somya D. Mohanty et al. “A multi-modal approach towards mining social media data during natural disasters - A case study of Hurricane Irma”. In: *International Journal of Disaster Risk Reduction* 54.July 2020 (2021), p. 102032. ISSN: 22124209. DOI: 10.1016/j.ijdr.2020.102032. arXiv: 2101.00480. URL: <https://doi.org/10.1016/j.ijdr.2020.102032>.
- [260] Alireza Mostafizi, Haizhong Wang, and Shangjia Dong. “Understanding the multimodal evacuation behavior for a near-field tsunami”. In: *Transportation research record* 2673.11 (2019), pp. 480–492. DOI: 10.1177/0361198119837511.
- [261] Hussein Mouzannar, Yara Rizk, and Mariette Awad. “Damage Identification in Social Media Posts Using Multimodal Deep Learning”. In: *Proceedings of the International ISCRAM Conference 2018-May*. May (2018), pp. 529–543. ISSN: 24113387.

- [262] Khan Muhammad, Jamil Ahmad, and Sung Wook Baik. “Early fire detection using convolutional neural networks during surveillance for effective disaster management”. In: *Neurocomputing* 288 (2018), pp. 30–42. ISSN: 18728286. DOI: 10.1016/j.neucom.2017.04.083. URL: <https://doi.org/10.1016/j.neucom.2017.04.083>.
- [263] Khan Muhammad, Jamil Ahmad, and Sung Wook Baik. “Early fire detection using convolutional neural networks during surveillance for effective disaster management”. In: *Neurocomputing* 288 (2018), pp. 30–42.
- [264] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. “Generative Adversarial Minority Oversampling”. In: *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 1695–1704. DOI: 10.1109/ICCV.2019.00178. URL: <https://doi.org/10.1109/ICCV.2019.00178>.
- [265] Chevuru Naga Pavan Srivathsav et al. “Detection of disaster affected regions based on change detection using deep architecture”. In: *International Journal of Innovative Technology and Exploring Engineering* 8.5 (2019), pp. 124–128. ISSN: 22783075.
- [266] Khac Hoai Nam Bui, Hongsuk Yi, and Jiho Cho. “A multi-class multi-movement vehicle counting framework for traffic analysis in complex areas using CCTV systems”. In: *Energies* 13.8 (2020). ISSN: 19961073. DOI: 10.3390/en13082036.
- [267] Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. “Deep Neural Networks versus Naive Bayes Classifiers for Identifying Informative Tweets during Disasters”. In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. Ed. by Kees Boersma and Brian M. Tomaszewski. ISCRAM Association, 2018. URL: [http://idl.iscram.org/files/venkatakishoreneppalli/2018/1589%5C\\_VenkataKishoreNeppalli%5C\\_etal2018.pdf](http://idl.iscram.org/files/venkatakishoreneppalli/2018/1589%5C_VenkataKishoreNeppalli%5C_etal2018.pdf).
- [268] Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. “Deep neural networks versus naïve bayes classifiers for identifying informative tweets during disasters”. In: *Proceedings of the International ISCRAM Conference 2018-May*. May (2018), pp. 677–686. ISSN: 24113387.
- [269] Allen Newell and Herbert A Simon. “Computer science as empirical inquiry: Symbols and search”. In: *Communications of the Association for Computing Machinery* 19 (1981).
- [270] Francesco Nex et al. “Structural building damage detection with deep learning: Assessment of a state-of-the-art CNN in operational conditions”. In: *Remote Sensing* 11.23 (2019). ISSN: 20724292. DOI: 10.3390/rs11232765.
- [271] Jiquan Ngiam et al. “Multimodal deep learning”. In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 689–696.
- [272] Dat T. Nguyen et al. “Damage assessment from social media imagery data during disasters”. In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2017* (2017), pp. 569–576. DOI: 10.1145/3110025.3110109.
- [273] Dat Tien Nguyen et al. “Applications of Online Deep Learning for Crisis Response Using Social Media Information”. In: *CoRR* abs/1610.01030 (2016). arXiv: 1610.01030. URL: <http://arxiv.org/abs/1610.01030>.



- [274] Dat Tien Nguyen et al. “Applications of online deep learning for crisis response using social media information”. In: *arXiv preprint arXiv:1610.01030* (2016).
- [275] Dat Tien Nguyen et al. “Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks.” In: *ICWSM 31.3* (2017), pp. 632–635.
- [276] Van Quan Nguyen, Tien Nguyen Anh, and Hyung Jeong Yang. “Real-time event detection using recurrent neural network in social sensors”. In: *International Journal of Distributed Sensor Networks* 15.6 (2019). ISSN: 15501477. DOI: 10.1177/1550147719856492.
- [277] Xiaodong Ning et al. “Source-Aware Crisis-Relevant Tweet Identification and Key Information Summarization”. In: *ACM Transactions on Internet Technology (TOIT)* 19.3 (2019), pp. 1–20. DOI: 10.1145/3300229.
- [278] Brian Keith Norambuena, Michael Horning, and Tanushree Mitra. “Evaluating the Inverted Pyramid Structure through Automatic 5W1H Extraction and Summarization”. In: *Computational Journalism Symposium*. 2020.
- [279] Vimala Nunavath and Morten Goodwin. “The Role of Artificial Intelligence in Social Media Big data Analytics for Disaster Management-Initial Results of a Systematic Literature Review”. In: *2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*. IEEE, 2018, pp. 1–4. DOI: 10.1109/ICT-DM.2018.8636388.
- [280] *NZTA’s ‘critical’ IT risks to cost more than \$50m to fix*. <https://www.rnz.co.nz/news/national/470368/nzta-s-critical-it-risks-to-cost-more-than-50m-to-fix?s=08>. Accessed: 2022-09-05.
- [281] Ferda Offi, Firoj Alam, and Muhammad Imran. “Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response”. In: 1.May 2020 (2020). arXiv: 2004.11838. URL: <http://arxiv.org/abs/2004.11838>.
- [282] Gabriel Oltean et al. “Towards Real Time Vehicle Counting using YOLO-Tiny and Fast Motion Estimation”. In: *SIITME 2019 - 2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging, Proceedings* October (2019), pp. 240–243. DOI: 10.1109/SIITME47687.2019.8990708.
- [283] Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. “What to expect when the unexpected happens: Social media communications across crises”. In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 2015, pp. 994–1009.
- [284] Alexandra Olteanu et al. “Crisislex: A lexicon for collecting and filtering microblogged communications in crises”. In: *Eighth international AAAI conference on weblogs and social media*. 2014.
- [285] Abraham Naftali Oppenheim. *Questionnaire design, interviewing and attitude measurement*. Bloomsbury Publishing, 2000.
- [286] Asil Oztekin. “A decision support system for usability evaluation of web-based information systems”. In: *Expert Systems with Applications* 38.3 (2011), pp. 2110–2118.
- [287] Swati Padhee et al. “Clustering of social media messages for humanitarian aid response during crisis”. In: *arXiv* (2020). ISSN: 23318422. arXiv: 2007.11756.

- [288] Ketan R Pandhare and Medha A Shah. “Real time road traffic event detection using Twitter and spark”. In: *2017 International conference on inventive communication and computational technologies (ICICCT)*. IEEE, 2017, pp. 445–449.
- [289] Kishore Papineni et al. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: <https://aclanthology.org/P02-1040/>.
- [290] Beverly Estephany Parilla-Ferrer, PL Fernandez, and JT Ballena. “Automatic classification of disaster-related tweets”. In: *Proc. International conference on Innovative Engineering Technologies (ICIET)*. Vol. 62. 2014.
- [291] Beverly Estephany Parilla-Ferrer, Proceso L Fernandez, and Jaime T Ballena. “Automatic classification of disaster-related tweets”. In: *Proc. International conference on Innovative Engineering Technologies (ICIET)*. Vol. 62. 2014.
- [292] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.
- [293] Lew Kan Peng, Chennupati K. Ramaiah, and Schubert Foo. “Heuristic-based user interface evaluation at Nanyang Technological University in Singapore”. In: *Program* 38.1 (2004), pp. 42–59. DOI: 10.1108/00330330410519198. URL: <https://doi.org/10.1108/00330330410519198>.
- [294] Jeffrey Pennington, Richard Socher, and Christopher D Manning. “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162.
- [295] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “Glove: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. ACL, 2014, pp. 1532–1543. DOI: 10.3115/v1/d14-1162. URL: <https://doi.org/10.3115/v1/d14-1162>.
- [296] MV Peppia et al. “Urban traffic flow analysis based on deep learning car detection from cctv image series”. In: *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* 42.4 (2018).
- [297] Fabio Petroni et al. “An extensible event extraction system with cross-media event resolution”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 626–635.
- [298] Yalong Pi, Nipun D. Nath, and Amir H. Behzadan. “Convolutional neural networks for object detection in aerial imagery for disaster response and recovery”. In: *Advanced Engineering Informatics* 43.October 2019 (2020), p. 101009. ISSN: 14740346. DOI: 10.1016/j.aei.2019.101009. URL: <https://doi.org/10.1016/j.aei.2019.101009>.
- [299] Yalong Pi, Nipun D. Nath, and Amir H. Behzadan. “Disaster impact information retrieval using deep learning object detection in crowdsourced drone footage”. In: *EG-ICE 2020 Workshop on Intelligent Computing in Engineering, Proceedings* (2020), pp. 134–143.

- [300] Jakub Piskorski et al. “Online news event extraction for global crisis surveillance”. In: *Transactions on computational collective intelligence V*. Springer, 2011, pp. 182–212.
- [301] *Please Send Help. Hurricane Harvey Victims Turn to Twitter and Facebook*. <https://time.com/4921961/hurricane-harvey-twitter-facebook-social-media/>. Accessed: 2020-11-10.
- [302] Nicolai Pogrebnyakov and Edgar Maldonado. “Identifying emergency stages in Facebook posts of police departments with convolutional and recurrent neural networks and support vector machines”. In: *arXiv* (2018), pp. 4343–4352. ISSN: 23318422.
- [303] Nicholas G Polson and Vadim O Sokolov. “Deep learning for short-term traffic flow prediction”. In: *Transportation Research Part C: Emerging Technologies* 79 (2017), pp. 1–17.
- [304] Peter G Polson et al. “Cognitive walkthroughs: a method for theory-based evaluation of user interfaces”. In: *International Journal of man-machine studies* 36.5 (1992), pp. 741–773.
- [305] Samira Pouyanfar et al. “A Survey on Deep Learning: Algorithms, Techniques, and Applications”. In: *ACM Comput. Surv.* 51.5 (2019), 92:1–92:36. DOI: 10.1145/3234150. URL: <https://doi.org/10.1145/3234150>.
- [306] Samira Pouyanfar et al. “Multimodal deep learning based on multiple correspondence analysis for disaster management”. In: *World Wide Web* (2018), pp. 1–19.
- [307] Raj Pranesh. “Exploring Multimodal Features and Fusion Strategies for Analyzing Disaster Tweets”. In: *Proceedings of the Eighth Workshop on Noisy User-generated Text, WNUT@COLING 2022, Gyeongju, Republic of Korea, October 12 - 17, 2022*. Association for Computational Linguistics, 2022, pp. 62–68. URL: <https://aclanthology.org/2022.wnut-1.6>.
- [308] Raj Prasanna and Thomas J Huggins. “Factors affecting the acceptance of information systems supporting emergency operations centres”. In: *Computers in Human Behavior* 57 (2016), pp. 168–181.
- [309] Raj Prasanna, Lili Yang, and Malcolm King. “Guidance for developing human–computer interfaces for supporting fire emergency response”. In: *Risk Management* 15.3 (2013), pp. 155–179.
- [310] Jan Pries-Heje, Richard L. Baskerville, and John R. Venable. “Strategies for Design Science Research Evaluation”. In: *16th European Conference on Information Systems, ECIS 2008, Galway, Ireland, 2008*. Ed. by Willie Golden et al. 2008, pp. 255–266. URL: <http://aisel.aisnet.org/ecis2008/87>.
- [311] Shalini Priya et al. “TAQE: Tweet Retrieval-Based Infrastructure Damage Assessment during Disasters”. In: *IEEE Transactions on Computational Social Systems* 7.2 (2020), pp. 389–403. ISSN: 2329924X. DOI: 10.1109/TCSS.2019.2957208.
- [312] *Public Feed API*. [https://developers.facebook.com/docs/public\\_feed/](https://developers.facebook.com/docs/public_feed/). Accessed: 2021-05-10.
- [313] Junaid Qadir et al. “Crisis analytics: big data-driven crisis response”. In: *Journal of International Humanitarian Action* 1.1 (2016), p. 12.
- [314] Sandy Q Qu and John Dumay. “The qualitative research interview”. In: *Qualitative research in accounting & management* 8.3 (2011), pp. 238–264.

- [315] Rafa E Al-Qutaish. “Quality models in software engineering literature: an analytical and comparative study”. In: *Journal of American Science* 6.3 (2010), pp. 166–175.
- [316] Maryam Rahnemoonfar et al. “Flooded area detection from UAV images based on densely connected recurrent neural networks”. In: *International Geoscience and Remote Sensing Symposium (IGARSS) 2018-July* (2018), pp. 1788–1791. DOI: 10.1109/IGARSS.2018.8517946.
- [317] Cyrus Rashtchian et al. “Collecting Image Annotations Using Amazon’s Mechanical Turk”. In: *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, Los Angeles, USA, June 6, 2010*. Ed. by Chris Callison-Burch and Mark Dredze. Association for Computational Linguistics, 2010, pp. 139–147. URL: <https://aclanthology.org/W10-0721/>.
- [318] Adnan Rawashdeh and Bassem Matakah. “A new software quality model for evaluating COTS components”. In: *Journal of Computer Science* 2.4 (2006), pp. 373–381.
- [319] Joseph Redmon and Ali Farhadi. “YOLO9000: better, faster, stronger”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 7263–7271. DOI: 10.1109/CVPR.2017.690.
- [320] Joseph Redmon and Ali Farhadi. “Yolov3: An incremental improvement”. In: *arXiv preprint arXiv:1804.02767* (2018).
- [321] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788. DOI: 10.1109/CVPR.2016.91.
- [322] Shaoqing Ren et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *Advances in neural information processing systems*. 2015, pp. 91–99. DOI: 10.1109/TPAMI.2016.2577031.
- [323] Centre for Research on the Epidemiology of Disasters (CRED). *The International Disaster Database*. <https://www.emdat.be/>. Accessed: 2021-09-01. 2021.
- [324] Leslie Rice, Eric Wong, and J. Zico Kolter. “Overfitting in adversarially robust deep learning”. In: *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 8093–8104. URL: <http://proceedings.mlr.press/v119/rice20a.html>.
- [325] John Rieman, Marita Franzke, and David F. Redmiles. “Usability evaluation with the cognitive walkthrough”. In: *Human Factors in Computing Systems, CHI '95 Conference Companion: Mosaic of Creativity, Denver, Colorado, USA, May 7-11, 1995*. Ed. by Jim Miller et al. ACM, 1995, pp. 387–388. DOI: 10.1145/223355.223735. URL: <https://doi.org/10.1145/223355.223735>.
- [326] Yara Rizk et al. “A computationally efficient multi-modal classification approach of disaster-related Twitter images”. In: *Proceedings of the ACM Symposium on Applied Computing Part F1477*. January (2019), pp. 2050–2059. DOI: 10.1145/3297280.3297481.
- [327] Brett W. Robertson et al. “Using a combination of human insights and ‘deep learning’ for real-time disaster communication”. In: *Progress in Disaster Science* 2 (2019), p. 100030. ISSN: 25900617. DOI: 10.1016/j.pdisas.2019.100030. URL: <https://doi.org/10.1016/j.pdisas.2019.100030>.

- [328] Colin Robson. *Real world research*. Vol. 3. Wiley Chichester, 2011.
- [329] Galina Rogova and Peter Scott. *Fusion Methodologies in Crisis Management: Higher Level Fusion and Decision Making*. Springer, 2016.
- [330] Jakob Rogstadius et al. “CrisisTracker: Crowdsourced social media curation for disaster awareness”. In: *IBM Journal of Research and Development* 57.5 (2013), pp. 4–1.
- [331] Yuji Roh, Geon Heo, and Steven Euijong Whang. “A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective”. In: *IEEE Trans. Knowl. Data Eng.* 33.4 (2021), pp. 1328–1347. DOI: 10.1109/TKDE.2019.2946162. URL: <https://doi.org/10.1109/TKDE.2019.2946162>.
- [332] Shuvendu Roy and Md Sakif Rahman. “Emergency Vehicle Detection on Heavy Traffic Road from CCTV Footage Using Deep Convolutional Neural Network”. In: *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2019, pp. 1–6.
- [333] M Sadeghzadeh Hemayati and H Rashidi. “Software quality models: A comprehensive review and analysis”. In: *Journal of Electrical and Computer Engineering Innovations (JECEI)* 6.1 (2017), pp. 59–76.
- [334] Amin Muhammad Sadiq, Huynsik Ahn, and Young Bok Choi. “Human Sentiment and Activity Recognition in Disaster Situations Using Social Media Images Based on Deep Learning”. In: *Sensors* 20.24 (2020), p. 7115. DOI: 10.3390/s20247115. URL: <https://doi.org/10.3390/s20247115>.
- [335] Bukhoree Sahoh and Anant Choksuriwong. “Smart Emergency Management Based on Social Big Data Analytics: Research Trends and Future Directions”. In: *Proceedings of the 2017 International Conference on Information Technology*. 2017, pp. 1–6. DOI: 10.1145/3176653.3176657.
- [336] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. “Earthquake shakes Twitter users: real-time event detection by social sensors”. In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 851–860.
- [337] Mark Saunders, Philip Lewis, and Adrian Thornhill. “Research methods”. In: *Business Students* (2007).
- [338] Mark Saunders, Philip Lewis, and Adrian Thornhill. “Understanding research philosophies and approaches”. In: *Research methods for business students* 4.1 (2009), pp. 106–135.
- [339] Anirban Sen, Koustav Rudra, and Saptarshi Ghosh. “Extracting situational awareness from microblogs during disaster events”. In: *2015 7th International Conference on Communication Systems and Networks (COMSNETS)*. IEEE, 2015, pp. 1–6.
- [340] *Sendai Framework for Disaster Risk Reduction 2015-2030*. <https://www.undrr.org/publication/sendai-framework-disaster-risk-reduction-2015-2030>. Accessed: 2022-10-01.
- [341] S. T. Seydi and H. Rastiveis. “A deep learning framework for roads network damage assessment using post-earthquake lidar data”. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives* 42.4/W18 (2019), pp. 955–961. ISSN: 16821750. DOI: 10.5194/isprs-archives-XLII-4-W18-955-2019.

- [342] Shayan Shams, Sayan Goswami, and Kisung Lee. “Deep learning-based spatial analytics for disaster-related tweets: An experimental study”. In: *Proceedings - IEEE International Conference on Mobile Data Management 2019-June.Mdm* (2019), pp. 337–342. ISSN: 15516245. DOI: 10.1109/MDM.2019.00-40.
- [343] Vido Shaweddy and Wahyono Wahyono. “Vehicle counting framework for intelligent traffic monitoring system”. In: *Proceedings - 2019 5th International Conference on Science and Technology, ICST 2019* (2019), pp. 1–5. DOI: 10.1109/ICST47872.2019.9166440.
- [344] Martin J. Shepperd. “Practical software metrics for project management and process improvement: R Grady Prentice-Hall (1992) £30.95 282 pp ISBN 0 13 720384 5”. In: *Inf. Softw. Technol.* 35.11-12 (1993), p. 701. DOI: 10.1016/0950-5849(93)90091-G. URL: [https://doi.org/10.1016/0950-5849\(93\)90091-G](https://doi.org/10.1016/0950-5849(93)90091-G).
- [345] Hoo-Chang Shin et al. “Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning”. In: *CoRR* abs/1602.03409 (2016). arXiv: 1602.03409. URL: <http://arxiv.org/abs/1602.03409>.
- [346] Prajol Shrestha, Christine Jacquin, and Béatrice Daille. “Clustering short text and its evaluation”. In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer. 2012, pp. 169–180.
- [347] Sima Siami-Namini, Neda Tavakoli, and Akbar Siami Namin. “The performance of LSTM and BiLSTM in forecasting time series”. In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE. 2019, pp. 3285–3292.
- [348] Herbert A Simon. *The sciences of the artificial*. MIT press, 1996.
- [349] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: <http://arxiv.org/abs/1409.1556>.
- [350] Muhammed Ali Sit, Caglar Koylu, and Ibrahim Demir. “Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma”. In: *International Journal of Digital Earth* 12.11 (2019), pp. 1205–1229. ISSN: 17538955. DOI: 10.1080/17538947.2018.1563219.
- [351] Michael J Spivey. “Redesigning our theories of human information processing”. In: *Information Design Journal* 15.3 (2007), pp. 261–265.
- [352] Nitish Srivastava and Ruslan R Salakhutdinov. “Multimodal learning with deep boltzmann machines”. In: *Advances in neural information processing systems*. 2012, pp. 2222–2230.
- [353] Nitish Srivastava et al. “Dropout: a simple way to prevent neural networks from overfitting”. In: *J. Mach. Learn. Res.* 15.1 (2014), pp. 1929–1958. URL: <http://dl.acm.org/citation.cfm?id=2670313>.
- [354] K Stewart Hornsby and W Wang. “Representing dynamic phenomena based on spatiotemporal information extracted from web documents”. In: *Extended abstracts, GIScience Conference 2010*. 2010.

- [355] Kevin Stowe et al. “Developing and evaluating annotation procedures for twitter data during hazard events”. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. 2018, pp. 133–143.
- [356] Kevin Stowe et al. “Identifying and categorizing disaster-related tweets”. In: *Proceedings of The fourth international workshop on natural language processing for social media*. 2016, pp. 1–6. DOI: 10.18653/v1/W16-6201.
- [357] Jérémie Sublime and Ekaterina Kalinicheva. “Automatic post-disaster damage mapping using deep-learning techniques for change detection: Case study of the Tohoku tsunami”. In: *Remote Sensing* 11.9 (2019). ISSN: 20724292. DOI: 10.3390/rs11091123.
- [358] Manoj Wadhwa Suman and MDU Rohtak. “A comparative study of software quality models”. In: *International Journal of Computer Science and Information Technologies* 5.4 (2014), pp. 5634–5638.
- [359] Wenjuan Sun, Paolo Bocchini, and Brian D. Davison. *Applications of artificial intelligence for disaster management*. 0123456789. Springer Netherlands, 2020. ISBN: 0123456789. DOI: 10.1007/s11069-020-04124-3. URL: <https://doi.org/10.1007/s11069-020-04124-3>.
- [360] Ying Sun and Paul B. Kantor. “Cross-Evaluation: A new model for information system evaluation”. In: *J. Assoc. Inf. Sci. Technol.* 57.5 (2006), pp. 614–628. DOI: 10.1002/asi.20324. URL: <https://doi.org/10.1002/asi.20324>.
- [361] Christian Szegedy et al. “Rethinking the Inception Architecture for Computer Vision”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826. DOI: 10.1109/CVPR.2016.308. URL: <https://doi.org/10.1109/CVPR.2016.308>.
- [362] Hristo Tanev, Jakub Piskorski, and Martin Atkinson. “Real-time news event extraction for global monitoring systems”. In: *Joint Research Center of the European Commission, Web and Language Technology Group of IPSC, TP 267* ().
- [363] Eric Tatulli and Thomas Hueber. “Feature extraction using multimodal convolutional neural networks for visual speech recognition”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2971–2975.
- [364] Charles Teddlie and Fen Yu. “Mixed methods sampling: A typology with examples”. In: *Journal of mixed methods research* 1.1 (2007), pp. 77–100.
- [365] Alberto Téllez Valero and Manuel Montes y Gómez. “Using machine learning for extracting information from natural disaster news reports”. In: (2009).
- [366] *The Learning Problem*. <http://work.caltech.edu/slides/slides01.pdf>. Accessed: 2021-03-10.
- [367] Haiman Tian, Hector Cen Zheng, and Shu-Ching Chen. “Sequential Deep Learning for Disaster-Related Video Classification”. In: *IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018, Miami, FL, USA, April 10-12, 2018*. IEEE, 2018, pp. 106–111. DOI: 10.1109/MIPR.2018.00026. URL: <http://doi.ieeecomputersociety.org/10.1109/MIPR.2018.00026>.

- [368] Haiman Tian, Hector Cen Zheng, and Shu-Ching Chen. “Sequential deep learning for disaster-related video classification”. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE. 2018, pp. 106–111.
- [369] Yan Tian et al. “LSTM-based traffic flow prediction with missing data”. In: *Neurocomputing* 318 (2018), pp. 297–305.
- [370] Hien To et al. “On identifying disaster-related tweets: Matching-based or learning-based?”. In: *2017 IEEE third international conference on multimedia big data (BigMM)*. IEEE. 2017, pp. 330–337. DOI: 10.1109/BigMM.2017.82.
- [371] Fujio Toriumi and Seigo Baba. “Real-time tweet classification in disaster situation”. In: *Proceedings of the 25th International Conference Companion on World Wide Web*. 2016, pp. 117–118. DOI: 10.1145/2872518.2889365.
- [372] *traffic monitoring for state highways*. <https://www.nzta.govt.nz/assets/resources/traffic-monitoring-state-hways/docs/traffic-monitoring-state-highways.pdf>. Accessed: 2021-12-31.
- [373] *TREC-Incident Streams*. [http://dcs.gla.ac.uk/~richardm/TREC\\_IS/](http://dcs.gla.ac.uk/~richardm/TREC_IS/). Accessed: 2021-04-01.
- [374] Jasper RR Uijlings et al. “Selective search for object recognition”. In: *International journal of computer vision* 104.2 (2013), pp. 154–171. DOI: 10.1007/s11263-013-0620-5.
- [375] Mascha Van Der Voort, Mark Dougherty, and Susan Watson. “Combining Kohonen maps with ARIMA time series models to forecast traffic flow”. In: *Transportation Research Part C: Emerging Technologies* 4.5 (1996), pp. 307–318.
- [376] John R. Venable, Jan Pries-Heje, and Richard L. Baskerville. “A Comprehensive Framework for Evaluation in Design Science Research”. In: *Design Science Research in Information Systems. Advances in Theory and Practice - 7th International Conference, DESRIST 2012, Las Vegas, NV, USA, May 14-15, 2012. Proceedings*. Ed. by Ken Peffers, Marcus A. Rothenberger, and William L. Kuechler Jr. Vol. 7286. Lecture Notes in Computer Science. Springer, 2012, pp. 423–438. DOI: 10.1007/978-3-642-29863-9\_31. URL: [https://doi.org/10.1007/978-3-642-29863-9\\_31](https://doi.org/10.1007/978-3-642-29863-9_31).
- [377] S Veni, R Anand, and B Santosh. “Road Accident Detection and Severity Determination from CCTV Surveillance”. In: *Advances in Distributed Computing and Machine Learning*. Springer, 2020, pp. 247–256. DOI: 10.1007/978-981-15-4218-3\_25.
- [378] Subhashini Venugopalan et al. “Captioning Images with Diverse Objects”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 1170–1178. DOI: 10.1109/CVPR.2017.130. URL: <https://doi.org/10.1109/CVPR.2017.130>.
- [379] Subhashini Venugopalan et al. “Translating videos to natural language using deep recurrent neural networks”. In: *arXiv preprint arXiv:1412.4729* (2014).
- [380] Rakesh Verma et al. “Newswire versus social media for disaster response and recovery”. In: *2019 Resilience Week (RWS)*. Vol. 1. IEEE. 2019, pp. 132–141.



- [381] Anand Vetrivel et al. “Disaster damage detection through synergistic use of deep learning and 3D point cloud features derived from very high resolution oblique aerial images, and multiple-kernel-learning”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 140 (2018), pp. 45–59. ISSN: 09242716. DOI: 10.1016/j.isprsjprs.2017.03.001. URL: <https://doi.org/10.1016/j.isprsjprs.2017.03.001>.
- [382] Oriol Vinyals et al. “Show and tell: A neural image caption generator”. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. IEEE Computer Society, 2015, pp. 3156–3164. DOI: 10.1109/CVPR.2015.7298935. URL: <https://doi.org/10.1109/CVPR.2015.7298935>.
- [383] Eleni I. Vlahogianni, John C. Golias, and Matthew G. Karlaftis. “Short-term traffic forecasting: Overview of objectives and methods”. In: *Transport Reviews* 24.5 (2004), pp. 533–557. DOI: 10.1080/0144164042000195072. URL: <https://doi.org/10.1080/0144164042000195072>.
- [384] Eleni I. Vlahogianni, Matthew G. Karlaftis, and John C. Golias. “Short-term traffic forecasting: Where we are and where we’re going”. In: *Transportation Research Part C: Emerging Technologies* 43 (2014). Special Issue on Short-term Traffic Flow Forecasting, pp. 3–19. ISSN: 0968-090X. DOI: <https://doi.org/10.1016/j.trc.2014.01.005>. URL: <https://www.sciencedirect.com/science/article/pii/S0968090X14000096>.
- [385] R Hevner Von Alan et al. “Design science in information systems research”. In: *MIS quarterly* 28.1 (2004), pp. 75–105.
- [386] Chien-yao Wang and Hong-yuan Mark Liao. “YOLOv4: Optimal Speed and Accuracy of Object Detection”. In: (2020). DOI: [arXiv:2004.10934v1](https://arxiv.org/abs/2004.10934).
- [387] Jingyuan Wang, Fei Hu, and Li Li. “Deep Bi-directional Long Short-Term Memory Model for Short-Term Traffic Flow Prediction”. In: *Neural Information Processing - 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part V*. Ed. by Derong Liu et al. Vol. 10638. Lecture Notes in Computer Science. Springer, 2017, pp. 306–316. DOI: 10.1007/978-3-319-70139-4\_31. URL: [https://doi.org/10.1007/978-3-319-70139-4\\_31](https://doi.org/10.1007/978-3-319-70139-4_31).
- [388] Limin Wang et al. “Object-scene convolutional neural networks for event recognition in images”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2015, pp. 30–35. DOI: 10.1109/CVPRW.2015.7301333.
- [389] Tianyi Wang et al. “Multi-Task Multimodal Learning for Disaster Situation Assessment”. In: *Proceedings - 3rd International Conference on Multimedia Information Processing and Retrieval, MIPR 2020* (2020), pp. 209–212. DOI: 10.1109/MIPR49039.2020.00050.
- [390] Wei Wang and Kathleen Stewart. “Spatiotemporal and semantic information extraction from Web news reports about natural hazards”. In: *Computers, environment and urban systems* 50 (2015), pp. 30–40.
- [391] Yuan Wang et al. “Enhancing transportation systems via deep learning: A survey”. In: *Transportation research part C: emerging technologies* (2018).
- [392] Zheyue Wang and Xinyue Ye. “Social media analytics for natural disaster management”. In: *International Journal of Geographical Information Science* 32.1 (2018), pp. 49–72.

- [393] Zheye Wang and Xinyue Ye. “Social media analytics for natural disaster management”. In: *Int. J. Geogr. Inf. Sci.* 32.1 (2018), pp. 49–72. DOI: 10.1080/13658816.2017.1367003. URL: <https://doi.org/10.1080/13658816.2017.1367003>.
- [394] Zheye Wang and Xinyue Ye. “Space, time, and situational awareness in natural hazards: A case study of Hurricane Sandy with social media data”. In: *Cartography and Geographic Information Science* 46.4 (2019), pp. 334–346.
- [395] Zheye Wang, Xinyue Ye, and Ming-Hsiang Tsou. “Spatial, temporal, and content analysis of Twitter for wildfire hazards”. In: *Natural Hazards* 83.1 (2016), pp. 523–540.
- [396] Napong Wanichayapong et al. “Social-based traffic information extraction and classification”. In: *2011 11th International Conference on ITS Telecommunications*. IEEE, 2011, pp. 107–112.
- [397] Cody Watson et al. “A systematic literature review on the use of deep learning in Software Engineering Research”. In: *arXiv* (2020). ISSN: 23318422. arXiv: 2009.06520.
- [398] Peter Wegner. “Research paradigms in computer science”. In: *Proceedings of the 2nd international Conference on Software Engineering*. Citeseer, 1976, pp. 322–330.
- [399] Mika Westerlund. “The emergence of deepfake technology: A review”. In: *Technology Innovation Management Review* 9.11 (2019).
- [400] Cathleen Wharton et al. “The cognitive walkthrough method: A practitioner’s guide”. In: *Usability inspection methods*. 1994, pp. 105–140.
- [401] Matti Wiegmann et al. “Analysis of Detection Models for Disaster-Related Tweets”. In: (2020).
- [402] Billy M Williams and Lester A Hoel. “Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results”. In: *Journal of transportation engineering* 129.6 (2003), pp. 664–672.
- [403] Si Si Mar Win and Than Nwe Aung. “Target oriented tweets monitoring system during natural disasters”. In: *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*. IEEE, 2017, pp. 143–148. DOI: 10.1109/ICIS.2017.7959984.
- [404] Peter C. Wright and Andrew F. Monk. “The use of think-aloud evaluation methods in design”. In: *ACM SIGCHI Bull.* 23.1 (1991), pp. 55–57. DOI: 10.1145/122672.122685. URL: <https://doi.org/10.1145/122672.122685>.
- [405] Bailin Yang et al. “Traffic flow prediction using LSTM with feature enhancement”. In: *Neurocomputing* 332 (2019), pp. 320–327. DOI: 10.1016/j.neucom.2018.12.016. URL: <https://doi.org/10.1016/j.neucom.2018.12.016>.
- [406] Tengfei Yang et al. “Social Media Big Data Mining and Spatio-Temporal Analysis on Public Emotions for Disaster Mitigation”. In: *ISPRS International Journal of Geo-Information* 8.1 (2019), p. 29. DOI: 10.3390/ijgi8010029.
- [407] Hongsuk Yi, Khac-Hoai Nam Bui, and HeeJin Jung. “Implementing A Deep Learning Framework for Short Term Traffic Flow Prediction”. In: *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics, WIMS 2019, Seoul, Republic of Korea, June 26-28, 2019*. Ed. by Rajendra Akerkar and Jason J. Jung. ACM, 2019, 7:1–7:8. DOI: 10.1145/3326467.3326492. URL: <https://doi.org/10.1145/3326467.3326492>.

- [408] Tan Yigitcanlar et al. “Can cities become smart without being sustainable? A systematic review of the literature”. In: *Sustainable cities and society* (2018).
- [409] Jie Yin et al. “Using Social Media to Enhance Emergency Situation Awareness: Extended Abstract”. In: *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. Ed. by Qiang Yang and Michael J. Wooldridge. AAAI Press, 2015, pp. 4234–4239. URL: <http://ijcai.org/Abstract/15/602>.
- [410] Peter Young et al. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Trans. Assoc. Comput. Linguistics* 2 (2014), pp. 67–78. URL: <https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/229>.
- [411] Manzhu Yu, Chaowei Yang, and Yun Li. “Big data in natural disaster management: a review”. In: *Geosciences* 8.5 (2018), p. 165.
- [412] Faxi Yuan and Rui Liu. “Mining social media data for rapid damage assessment during Hurricane Matthew: feasibility study”. In: *Journal of Computing in Civil Engineering* 34.3 (2020), p. 05020001.
- [413] Shengcheng Yuan et al. “An urban traffic evacuation model with decision-making capability”. In: *10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013*. Ed. by Tina Comes et al. ISCRAM Association, 2013. URL: [http://idl.iscram.org/files/yuan/2013/1136%5C\\_Yuan%5C\\_etal2013.pdf](http://idl.iscram.org/files/yuan/2013/1136%5C_Yuan%5C_etal2013.pdf).
- [414] Hongwon Yun. “Disaster events detection using twitter data”. In: *Journal of information and communication convergence engineering* 9.1 (2011), pp. 69–73.
- [415] Da Zhang and Mansur R. Kabuka. “Combining Weather Condition Data to Predict Traffic Flow: A GRU Based Deep Learning Approach”. In: *15th IEEE Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress, DASC/PiCom/DataCom/CyberSciTech 2017, Orlando, FL, USA, November 6-10, 2017*. IEEE Computer Society, 2017, pp. 1216–1219. DOI: 10.1109/DASC-PiCom-DataCom-CyberSciTec.2017.194. URL: <https://doi.org/10.1109/DASC-PiCom-DataCom-CyberSciTec.2017.194>.
- [416] FK Zhang, Feng Yang, and Ce Li. “Fast vehicle detection method based on improved YOLOv3 [J]”. In: *Computer Engineering and Applications* 55.2 (2019), pp. 12–20.
- [417] Jian-Shu Zhang, Jie Cao, and Bo Mao. “Application of deep learning and unmanned aerial vehicle technology in traffic flow monitoring”. In: *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*. Vol. 1. IEEE, 2017, pp. 189–194. DOI: 10.1109/ICMLC.2017.8107763.
- [418] Qi Xing Zhang et al. “Wildland Forest Fire Smoke Detection Based on Faster R-CNN using Synthetic Smoke Images”. In: *Procedia Engineering* 211 (2018), pp. 441–446. ISSN: 18777058. DOI: 10.1016/j.proeng.2017.12.034. URL: <https://doi.org/10.1016/j.proeng.2017.12.034>.

- [419] Qinghui Zhang, Xianing Chang, and Shanfeng Bian. “Vehicle-Damage-Detection Segmentation Algorithm Based on Improved Mask RCNN”. In: *IEEE Access* 8 (2020), pp. 6997–7004. DOI: 10.1109/ACCESS.2020.2964055.
- [420] Yongjie Zhang, Jian Wang, and Xin Yang. “Real-time vehicle detection and tracking in video based on faster R-CNN”. In: *Journal of Physics: Conference Series*. Vol. 887. 1. IOP Publishing. 2017, p. 012068. DOI: 10.1088/1742-6596/887/1/012068.
- [421] Fei Zhao and Chengcui Zhang. “Building Damage Evaluation from Satellite Imagery using Deep Learning”. In: *Proceedings - 2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science, IRI 2020* (2020), pp. 82–89. DOI: 10.1109/IRI49571.2020.00020.
- [422] Zheng Zhao et al. “LSTM network: a deep learning approach for short-term traffic forecast”. In: *IET Intelligent Transport Systems* 11.2 (2017), pp. 68–75.
- [423] Zhong-Qiu Zhao et al. “Object detection with deep learning: A review”. In: *IEEE transactions on neural networks and learning systems* (2019).
- [424] Xin Zheng and Aixin Sun. “Collecting event-related tweets from twitter stream”. In: *Journal of the Association for Information Science and Technology* 70.2 (2019), pp. 176–186. DOI: 10.1002/asi.24096.
- [425] Zibin Zheng et al. “Deep and Embedded Learning Approach for Traffic Flow Prediction in Urban Informatics”. In: *IEEE Trans. Intell. Transp. Syst.* 20.10 (2019), pp. 3927–3939. DOI: 10.1109/TITS.2019.2909904. URL: <https://doi.org/10.1109/TITS.2019.2909904>.