Tweet Classification for Crisis Response

Reem AlRashdi

Doctor of Philosophy

University of York

Computer Science

June 2022

Dedication

To my children

Abstract

Tweet classification for crisis response is a text classification task that aims at identifying whether a tweet is related to a specific crisis event or not. Humanitarian organisations that intend to respond to people in need in the early hours of a crisis suffer from monitoring the massive number of tweets posted in real time. Therefore, the main objective of tweet classification models for crisis response is to filter the crisis-related tweets to simplify the work for these organisations. Still, crisis events have different characteristics, which prevents current models trained on past events from generalising in identifying tweets from new disasters, which is infeasible to be manually labelled at the crisis onset. This thesis introduces frameworks under the umbrella of distant supervision and domain adaptation to minimize the gap or maximize the similarities between training and testing data from disaster events. The contributions demonstrate the effectiveness of using automatically labelled training data from past or emerging events in tweet classification tasks for English and Arabic crisis tweets. To this end, we propose an automatically labelling framework that utilises distant supervision via an external knowledge base. Then, we introduce an approach that unifies our framework and adaptation techniques which automatically labels incoming tweets from an emerging incident. This approach can be seen as a robust method to classify unseen English tweets from current events. However, it has its restrictions when applied to tweets from other languages, especially if the language comes with limited resources, different text structures, and different people's behavior in posting tweets such as Arabic. Hence, we adapt our framework with significant changes to suit Arabic user-generated posts. Our results for both English and Arabic tweets show that our original and adaptive approaches continuously improve the classifier's performance compared with existing labelling techniques in different adaptation methods.

Contents

List of Figures

List of Tables

List of Algorithms

Acknowledgement

Firstly, I would like to express my sincere gratitude to my supervisor Dr. Simon O'Keefe for his continuous support, patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study.I thank Suresh Manandhar for giving me this opportunity, by accepting me as a PhD student. I also thank Dimitar Kazakov for all his advises during the TAP meetings.

I want to thank my friends in the Artificial Intelligence group for providing the moral support and a friendly work environment in the Computer Science department. I also thank Mr. Saleh AlMahmoud for providing his GPU during my last experiments in this thesis and for the valuable conversations and advices.

I would like to thank my family: my parents, my brothers and sisters for supporting me spiritually throughout writing this thesis and my life in general.

Last but not the least, I thank my partner for his unconditional, unequivocal, and loving support for the entire thesis process and everyday. Without his support, this thesis would not have been possible. Thank you for the strength you gave me. I also thank my children Mohammed, Abdullah and Yasmeen for being there for me the whole time. My accomplishments and success are because they believed in me. To my children, I give everything, including this.

Reem ALRashdi

June 2022, York, England

Declaration of author

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

The following papers resulted from the research in the direction to improve tweet classification models for crisis response.

Reem ALrashdi and Simon O'Keefe, "Deep learning and word embeddings for tweet classification for crisis response", In The 3rd National Computing Colleges Conference, 08-09 Oct 2018.

Above paper [16] forms the basis for the Chapter 3.

Reem Alrashdi and Simon O'Keefe. "Automatic Labeling of Tweets for Crisis Response Using Distant Supervision". In Companion Proceedings of the Web Conference 2020 (WWW'20). ACM, 2020. p. 418-425. Tapia, Taiwan.

Above paper [17] forms the basis for the Chapter 5.

Reem Alrashdi and Simon O'Keefe. "Robust Domain Adaptation Approach for Tweet Classification for Crisis Response". In International Conference Europe Middle East and North Africa Information Systems and Technologies to Support Learning (EMENA-ISTL 2019). Springer, Cham. p. 124-134. Marrakech, Morocco.

This paper [18] forms the basis for the Chapter 6.

Chapter 1

Introduction and Motivation

A crisis is a tragic large-scale event with a specific time and location [10] that is experienced by a large number of people whose daily lives are affected when one or more occur [99]. During crises, people disseminate the news on Twitter and share valuable, real-time and on-topic information, such as their statuses, information about injured or dead people, and infrastructural damage [130]. They also tweet to ask for help or to offer help to others. As a result, Twitter has become a dominant platform for organisations and individuals to disseminate or gather information during many natural and human-made crises in recent years [92], such as earthquakes [105], floods [118], wildfires [129] and nuclear disasters [123]. For example, in 2011, 177 million crisis-related tweets were shared on a single day during an earthquake in Japan [38]. Another example is when a haze hit Singapore in 2013, where people posted more than 23 million informative tweets [103].

The author in [130] states that people-generated tweets could significantly enhance situational awareness: relevant tweets can be used by large-scale disaster response organisations to respond to people during disasters, make better decisions and respond quickly [93]. For example, humanitarian organisations such as the UN can use tweets containing rich information to recognise medical emergencies. However, these organisations cannot manually observe, process and convert the enormous volume of data into actionable information [50]. Surveys of staff dealing with emergencies in seven European countries [107] and a survey of staff who manage and control emergencies in the US [102] have shown that emergency teams believe that social media (including Twitter) is a potent and valuable information source. However, they also believe that various issues delay the application of these data

in real-time operations. These issues include management and technical problems, such as the number of staff, reliability of social media data, and information overload. Thus, these organisations do not widely use social media data in their disaster response operations [122].

Challenges

Tweet classification for crisis response is a text classification task that aims to identify whether a tweet is related to a specific crisis event/type [32]. For example, the tweet "BREAKING: Nepal police official says at least 1,910 have died, including 721 in Kathmandu, in the quake" related to a Nepal earthquake event, while "So important! Hindu, Buddhist, Christian and Muslim leaders denounce #childmarriage in joint broadcast in Nepal" is irrelevant in the context of earthquakes. The main purpose of tweet classification for crisis response is to reduce the volume of tweets in real time, thereby simplifying the work of humanitarian organisations to respond to people in need in the early hours of a crisis. However, this task is challenging for two main reasons. First, tweets are informal, full of noise and limited to only 140 characters, meaning they are difficult to understand. Second, individuals' judgments of the corresponding crisis event of a given tweet are subjective [92].

Limitations

In addition, current tweet classification models suffer from three fundamental problems. The first is the lack of labelled training data [34], which prevents the models from reaching a generalised model [82], as tweets related to various crisis events have different features and social media responses [97]. Moreover, certain crises do not occur frequently enough to be collected, such as airplane crashes [93]. Second, producing labelled training data for every crisis event would be a time-consuming and expensive task that requires significant effort and money. In turn, the authors in [128] note that previous models trained on a source event cannot successfully generalise to a target event, even if the two events fall under the same crisis type (e.g., earthquake), because each event has its own distinct location and nature. Likewise, it is infeasible to manually annotate tweets for a crisis event in real time. Third, in accordance with these inherited issues, classification models for tweets from low-resource languages like Arabic are unable to reach a good level of performance for crisis data. Finetuning large, pre-trained models on data from the same language

causes performance to deteriorate if the downstream task has a minimal dataset [53] from Twitter data [68], which is the case with tweet classification for Arabic crisis data. Producing syntenic tweets with high confidence can be helpful when used for training cross-lingual classifiers in the setting of zero-shot learning (where the model makes predictions of unseen classes without being trained on data therein). However, producing such data is difficult, and the transferred domains should have a similar distribution of labels [68], which is not the case for Arabic crisis data.

Contributions

For the classification of texts, especially crisis-related tweets, a clear need exists for more well-labelled training data. Furthermore, the classification of tweets from low-resource languages needs to reach a robust model in real-time situations. Given these issues, approaches that automatically label tweets from new or current crisis events – and domain adaptation methods that use automatically labelled data from target disasters – are desirable to boost the performance of tweet classifiers. In addition, one of the most successful techniques for automated labelling for textual data is distant supervision. Crisis keywords also play an essential role in the annotations of tweets for crisis response [111]. Therefore, the work presented in this thesis aims to enhance the field of tweet classification for crisis response by utilising a novel distant supervision framework. This framework expands crisis keywords to automatically label crisis tweets to be incorporated into the training data under transfer learning settings, including domain adaptation.

The context of this contribution is to build tweet classifiers that are ready for humanitarian organisations to use when a crisis hits. We contribute to this research field by introducing a novel distant supervision framework under the umbrella of automated labelling of training data and domain adaptation for normal- and low-resource languages, English and Arabic, respectively. The work presented in this thesis aims to enhance the field of tweet classification for crisis response. The next sections briefly describe our research aims and contributions.

1.1 Research Aims of the Thesis

Based on the limitations discussed, this research aims to answer the following questions:

- Can we automatically generate labelled training data for tweet classification for crisis response with a competitive quality level compared to manually labelled training data?
- Can the tweet classification model be improved by employing distant supervision of unlabelled tweets from current (target) events to incorporate them into the training data in real time?
- Can the original distant supervised framework be modified to automatically label tweets in Arabic (low-resource languages) to be incorporated into the training data to successfully classify Arabic tweets from current events?

1.2 Thesis Contributions

To answer the research questions, we provide four contributions to the field of tweet classification for crisis response.

Deep learning and word embedding for tweet classification for crisis response

Chapter 3 shows our first contribution to studying the best deep learning architecture and word embedding to build a good tweet classifier for crisis response. We compare four tweet classification models using the CrisisNLP dataset with general-purpose and domain-specific word embeddings. The results indicate that general-purpose word embedding, such as Global Vectors for word representation (GloVe), can be used instead of domain-specific word embedding, especially with Bidirectional Long Short-Term Memory (BiLSTM), where the results reported the highest performance (0.6204 F1 score).

Automatic labelling using distant supervision

The tweet classification models currently used to enhance crisis response are based on supervised deep learning. They rely on the quality and quantity of human-labelled training data. However, the available training data are small and imbalanced in their coverage of crisis types, which prevents the models from generalisation. Such datasets are also expensive to produce due to the manual labelling. Chapter 5 presents our second contribution to overcoming this issue by introducing a distant supervision-based framework to automatically generate large-scale labelled data for tweet classification for crisis response. Experimental results on different crisis events from five crisis types show that our work can produce good-quality labelled data from past and recent events. Substituting automatically labelled training data for part of the manually labelled training data has minimal impact on model performance, indicating that automatically labelled data can be used when no hand-labelled data are available.

Domain adaptation for English crisis response

Deep learning algorithms can identify related tweets to reduce the information overload that prevents humanitarian organisations from using Twitter posts. However, they rely heavily on labelled data, which are unavailable for emerging crises. Because each crisis has its own features, such as location, time and social media response, current models are known to suffer from generalising to unseen disaster events when pre-trained on past ones. To solve this problem, Chapter 6 demonstrates our third contribution by introducing a novel domain adaptation approach that uses our distant supervision-based framework to label the unlabelled data from emerging crises. Pseudo-labelled target data and labelled data from similar past disasters are then used to build the target model. Finally, we investigate the model's performance for crisis-related data and compare it to the pseudo-labelling technique used in the crisis response literature in three adaptation methods. We evaluate our work on eight 2012–2015 crisis events from three crisis types (earthquake, floods and typhoons). Our results show that our approach can be considered a general robust method for classifying unseen tweets from current events.

Domain adaptation for Arabic crisis response

Tweet classification for crisis response for low-resource languages has the additional issue of limited labelled data duplicates caused by the absence of good external language resources. Thus, we apply some changes to our proposed domain adaptation approach in Chapter 6 to be used to automatically label tweets from low-resource languages like Arabic. Chapter 7 demonstrates our fourth contribution to the crisis response field from Twitter data, especially in languages with limited resources. Unlike the original version, our adaptive method does not rely on human-labelled data for the labelling task. It also expands our approach's ability to use corpora from other crisis types in the target data to create keyword sets that suit the situation of Arabic tweets. We evaluate our work on data from seven 2018–2020 Arabic events from different crisis types (flood, explosion, virus and storm). Preliminary results show that our method boosts the performance of the Arabic crisis-related tweet classifier in real-time scenarios.

1.3 Thesis Structure

The remainder of this thesis is structured as follows. Chapter 2 presents the background literature on the topics discussed in this thesis. It provides a detailed analysis of prior tweet classification approaches in the context of crisis response and highlights gaps in this field. It also details existing research on distant supervision and domain adaptation. Chapter 3 outlines our work in investigating the best word embedding and deep learning algorithms for building a good tweet classifier for crisis response. Chapter 4 presents the experimental setup used in this thesis, including the datasets, the classification models and the evaluation metrics. Chapter 5 introduces an automatic labelling framework that employs distant supervision to generate training data from new crisis events, whereby we incorporate the generated labelled data into the limited available manually labelled data in training tweet classifiers for crisis response as a means to improve the generalisation level of the tweet classification models. Chapter 6 discusses the application of our framework in a domain adaptation method to label unseen (target) tweets from emerging crises. This adds important new features to the training data, helping to improve the model's

performance. Chapter 7 presents an adaptive distant supervision labelling framework in the setting of domain adaptation for low-resource languages. This approach automatically labels Arabic tweets to classify data from current Arabic events. This proves that the proposed method is flexible enough to be extended to other languages and disrupts the need for human-labelled data. Finally, we provide a general conclusion and possible future directions for our work in Chapter 8.

Appendix A presents examples of the initial and expanded keyword lists created and used by our framework. Appendix B provides further analysis using the two-way Analysis Of VAriance (ANOVA) test for the results from Chapter 6. Finally, Appendix C provides the results and discusses the impact of excluding the distant supervision step from our framework in Chapter 7.

Chapter 2

General Background and Literature Review

This chapter highlights three areas related to our work in this thesis: tweet classification for crisis response, distant supervision and domain adaptation. First, Section 2.1 focuses on tweet classification for crisis response, the field of our contributions. Next, Section 2.2 discusses approaches and prior studies on distant supervision, which is the concept behind our proposed framework (see Chapter 5). Finally, Section 2.3 focuses on transfer learning and domain adaptation, which is the approach we use to incorporate new training data.

2.1 Tweet Classification for Crisis Response

Event Detection (ED) is an essential but challenging information extraction task that includes classifying instances of specific event types in the text [33]. Specifically, Tweet Classification For Crisis Response (TCFCR) is an ED task that aims to identify whether a tweet is related to a specific crisis event/type [32]; the main objective of TCFCR is to filter crisis-related tweets to simplify the work for humanitarian organisations.

2.1.1 TCFCR Categories

ED models are classified into two approaches: matching-based and learning-based models.

2.1.1.1 Matching-based models

The purpose of matching-based models is to identify related tweets based on predefined keywords and hashtags. First, a set of keywords and hashtags belonging to a specific crisis is collected. For example, if the crisis type is "wildfire", then the keywords and hashtags may contain "wildfire" and "fire". Next, a dictionary of more relevant hashtags can be constructed after searching a tweet collection using core keywords related to the crisis [124]. All candidate hashtags are then refined by crowd reviewers to improve the quality of the hashtag collection. Subsequently, the refined hashtags and keywords related to the crisis are combined into a final list to search for crisis-related tweets. The matching-based approach is outlined in Figure 2.1.

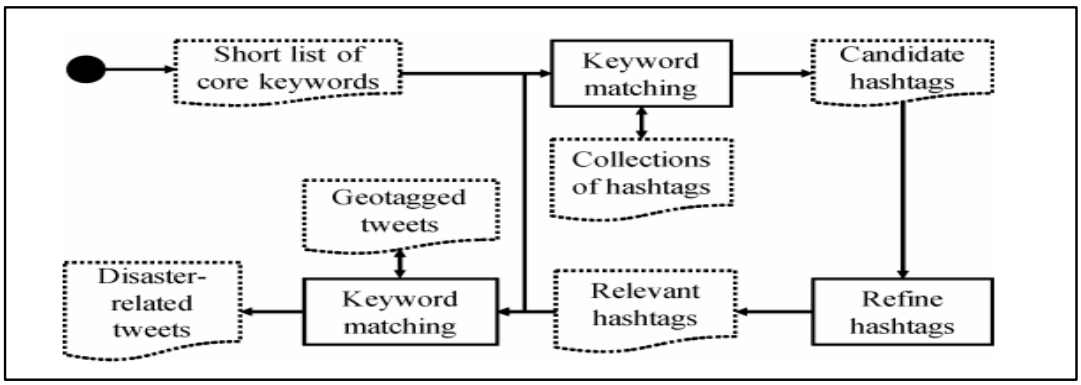

FIGURE 2.1: Matching-based approach. Source: [124].

However, original matching-based systems have various disadvantages. One issue is that they cannot retrieve related tweets that do not contain these keywords or hashtags, even if the tweets contain words with similar meanings. Another issue is that they mislabel irrelevant tweets that mention one of the hashtags or keywords where no noise reduction technique is used. Geolocation has also been used to retrieve related tweets; however, this feature does not exist for most of the posted tweets [116].

2.1.1.2 Learning-based models

This category addresses Natural Language Processing (NLP) problems – including text classification – by applying learning algorithms, including traditional and deep

learning models on texts [93]. In traditional Machine Learning (ML) models, standard classifiers and feature engineering methods are used, such as Support Vector Machine (SVM). Unlike traditional models, deep learning models use artificial learning models such as Convolutional Neural Networks (CNNs) with little or no feature engineering. In addition, bag-of-words models are applied in traditional models, while word embeddings are used in deep learning models. For crisis data, Nguyen et al. [92] have stated that deep learning models are perfectly suitable for such situations because they use distributed representation of words, learn the features automatically and can be applied in an online learning fashion.

2.1.2 TCFCR Approaches

Prior research studies have utilised different learning techniques to filter crisis data in an attempt to reduce the information overload problem. Sakaki et al. [111] developed a Japanese earthquake report system to alert citizens by sending SMS messages to registered mobile devices if the earthquake hit Japan. The authors considered Twitter users as sensors and addressed event detection problems as sensory observations. They monitored Twitter stream data and applied semantic analysis to tweets using an SVM classifier, thus classifying them into a positive event (earthquake) and a negative event (other events or non-events) based on three features: the existence of manually defined keywords in the tweet, the number of words (statistical) and their context. Furthermore, since some tweets are also associated with time and location, Kalman filtering and particle filtering were used as location estimation methods to detect an earthquake's location. The results showed that the combination of the three features improved the model's performance, and particle filtering outperformed Kalman filtering for location estimation. One drawback of this application is that the manually predefined keywords are limited (i.e., "earthquake" and "shaking"). Another is that the estimation of the earthquake's location sometimes showed unrealistic results.

Verma et al. [128] identified situational awareness tweets during crisis events using NLP techniques combined with ML algorithms, such as naive Bayes and maximum entropy. The crisis data were taken from the 2009 and 2010 Red River Floods,

the 2010 Haiti Earthquake, and the 2009 Oklahoma Grass Fire. According to the results, the classifiers generalised well across the 2009 and 2010 Red River Floods, but not across the other two disasters. For instance, the accuracy was low when using the classifier learned from the Oklahoma Grass Fire data to identify the Haiti Earthquake data, and vice versa. This is because the two events featured different characteristics regarding location, crisis type and people response.

Imran et al. [63] discovered valuable and self-contained crisis-related information terms in domain adaptation settings around two crisis events: the source data was the Joplin Tornado data, and the target data was the Hurricane Sandy data. First, the authors categorised different kinds of informative tweets using naive Bayes classifiers. They then employed Conditional Random Fields (CRF) to extract operational information, such as the number of victims or the name of the infrastructure. Finally, they built supervised classifiers from either source data or source data plus 10% labelled target data. The classifiers were tested on all target data and the remaining 90% of the target data. The authors compared the domain adaptation results with the outcomes of the supervised classifiers learned from 66% of the labelled target data and tested on 33% of the target data. Their findings revealed that incorporating target data increased the detection rate while not affecting recall.

Imran et al. [64] performed similar experiments by investigating the utility of tweets from prior crises and the utility of adding tweets from several languages. They built a random forest classifier from source data to be tested on target data using tweets from different crisis types, locations and languages. Their findings demonstrated that data from previous disasters of the same type as the current event, regardless of the language used to write the tweets, may benefit the disaster response.

Previous studies have addressed the problem of classifying tweets for crisis response using traditional ML algorithms. However, the supervised classifiers' performance in these works remains poor when tested across multiple types of events, especially for identifying tweets relevant to a particular disaster. As a result, deep learning models that have previously proven effective in text classification have been adopted for crisis tweet classifications.

Caragea, Silvescu, and Tapia [34] have studied TCFCR using a CNN to classify

disaster-related tweets into informative and uninformative tweets. They tokenised the texts into token sequences to be passed to CNN. CNN filters then perform as n-grams over continuous representations; these n-gram filters are then combined with subsequent network layers (dense layers). The results showed that CNN outperformed traditional ML models. This is because CNN can learn the features and distinguish between them automatically. Therefore, CNN does not require hand-engineered features, which saves human effort and time and eliminates the need for prior knowledge. Unlike a Multilayer Perceptron (MLP), the number of free parameters can be reduced by CNNs, and the vanished or exploded gradients can be prevented during the training process. Furthermore, all the weights in the convolutional layers are shared, which means that the same filter is used for all the fields within a layer, thereby improving performance and decreasing the memory space required.

Nguyen et al. [93] have argued that the informative class still contains much information to be handled by organisations. To simplify their work and save time and effort, they introduced a model that classified the informative class into multiple subclasses (e.g., infrastructure damages, affected people, donation and volunteering, sympathy and support, and other helpful information). They used two datasets, CrisisNLP and CrisisLex, and different pre-trained word embeddings: crisis embedding, domain-specific word embedding and Google word embedding. Their results reported that using different and multiple word embeddings slightly improved model performance. This was due to the variability of the corpora used when training the word embeddings. The authors also highlighted that out-of-event labelled examples could be used to train an event detection model when no in-event labelled examples were available. For example, labelled tweets from the Queensland Floods event could be used to train a model that classifies tweets from the Nepal Earthquake event. However, the results were unstable and highly dependent on the training data (source event); results differed when changing the source events with no existing criteria to choose the best source event in real time.

Liu et al. [84] proposed a transformer-based model that applied the concepts of Bidirectional Encoder Representations from Transformers (BERT) to crisis data (CrisisBERT) to enhance humanitarian aid. This approach showed promising results

across accuracy and F1 scores when tested on three different datasets for crisis detection and crisis recognition tasks. The authors also proposed contextual document-level embedding (Crisis2vec). In turn, other deep learning algorithms with different word embeddings were compared to the proposed model. Although they reported the F1 scores of CNN with GloVe and Long Short-Term Memory (LSTM) with word2vec, they did not report the results of using LSTM with GloVe embedding, which is similar to the classification model and word embedding used in this thesis. Unlike [84], Li et al. in [81] presented a comprehensive study of identifying disaster tweets using learning algorithms with different pre-trained word embeddings. Their experiments compared the use of word2vec, GloVe and fastText for sentence embeddings. Their findings demonstrated that GloVe recorded the best overall performance on the three different datasets. Paul, Sahoo and Balabantaray [100] highlighted the effectiveness of utilising CNN as a feature extraction layer in hybrid deep learning models, where local features can be detected in multidimensional texts. They combined CNN with Gated Recurrent Unit (GRU) in the first model and with SkipCNN in the second model. The researchers performed one-to-one in-domain experiments for collections related to four crisis events: the Nepal Earthquake, the California Earthquake, the Pam Cyclone and the Hagupit Typhoon. The results showed that their model (CNN-SkipCNN) outperformed GRU and CNN up to 16.55 absolute points for detecting crisis-related tweets.

Recently, tweet classification models have also been successfully applied for languages other than English. Alqaraleh and Işik in [14] have used CNNs to train a classifier to identify crisis-related Turkish tweets. Dharma and Winarko utilised a similar architecture with BERT embedding to classify tweets in the Indonesian language [45]. Alabbas et al. in [5] and Adel and Wang in [4] have used supervised traditional ML algorithms to classify flood-related Arabic tweets, while Alharbi and Lee in [9] have finetuned the Arabic BERT model using manually labelled Arabic crisis tweets from flood events.

However, the work presented in the prior research with deep learning algorithms has certain drawbacks. Deep learning approaches require a massive amount of labelled training data to build a robust model. This issue introduces a significant challenge to researchers when limited labelled data are available during training. The

datasets currently available for event detection using Twitter data are imbalanced and limited to specific crisis types. Some crises do not occur frequently enough to be collected, such as building collapses. This reduces the generalisation level of the classifiers and their ability to adapt to new domains. Notably, there is an urgent need to address these issues for event detection from Twitter data to build a robust and reliable model to serve humanitarian organisations, and this need intensifies when classifying crisis tweets in low-resource languages such as Arabic.

Domain adaptation approaches have been proposed for TCFCR to leverage the gap between source and target data, that is, by helping the classification models generalise to new different events from the training events. Alam, Joty and Imran in [6] have combined domain adversarial training and graph embeddings to propose a semi-supervised domain adaptation approach for crisis data; graph embeddings persuade structural similarity, while adversarial training reduces the distribution shift between source and target data. The authors used labelled and unlabelled data from two crisis events: the Nepal Earthquake and the Queensland Floods. Their results demonstrated that the domain adaptation approach outperformed the supervised learning method (CNN) in this study.

Li and Caragea in their paper [83] have jointly trained the sequence-to-sequence (seq2seq) model with Recurrent Neural Networks (RNNs) on disaster data for tweet classification on source data and reconstruction tasks for target data. The reconstruction task contained an autoencoder that reconstructs the target data, while the source shared the encoder; its reduced representation was used to learn a source classifier. The findings demonstrated that the reconstruction task could benefit domain adaptation settings. Further domain adaptation approaches related to our work in this thesis will be discussed in Section 2.3.

Human-labelled datasets from Twitter data have been publicly available to enhance the crisis response during a disaster for English [63, 41, 43, 7], French [73] and Arabic [8, 58, 137, 3]. Other datasets contain tweets from other languages besides English, such as Italian and Spanish [95, 96, 64]. However, manually labelling texts is expensive and requires time and effort. Therefore, several researchers have investigated similar approaches to this thesis in creating datasets using automatic

labelling to improve situational awareness using hashtags and emotions for crisis-related tweets.

Chowdhury et al. [39] have suggested that hashtags can be helpful in automatically annotating informative tweets. They built a unique dataset from Twitter data by filtering the crisis corpus using these hashtags. Hashtag prediction models were then trained using this dataset. LSTM achieved the best performance with 0.9222 F1 score. Desai, Caragea and Li [44] have used emotions to create an emotion dataset of 15,000 tweets from three hurricanes: Harvey, Irma and Maria. The authors suggested that the introduced dataset, HurricanEmo, could be used to analyse emotions in disaster tweets for classification tasks. Using this dataset, thier BERT model achieved 68% accuracy. Khare in his PhD thesis [70] has explored using semantics extracted from knowledge bases, such as DBpedia and Wikipedia, to identify crisis tweets for hurricane events. He applied Name Entity Linking (NEL) to determine the exact context of the extracted entities (semantics) within the tweet, helping gather contextual information about the tweet. These works are related to this thesis in terms of dataset creation. Unlike the existing research, however, this thesis introduces an automatically labelled dataset for crisis responses from Twitter data using a distant supervision-based approach. Our work differs from the dataset creation studies for TCFCR by using crisis-keyword lists expanded with external resources rather than original hashtags or emotions. Our dataset was also created to fulfil the need for more labelled training data in the extant literature.

Recently, Wahid et al. in [131] have presented a work similar to Chapters 6 and 7 of this thesis in terms of the method's general structure. This method consists of two main parts: annotating tweets using topic-oriented labels and building classifiers using these tweets. Latent Dirichlet Allocation (LDA) was used to extract meaningful topics – sorted using the Topic Term Frequency-Inverse Document Frequency (T-TF-IDF) ranking algorithm. The dominant topic was used to label the given tweets automatically. The classifiers used BERT embeddings and various deep learning architectures to classify crisis-related tweets based on their information types. The overall method was tested on two datasets: a combination of data from several crisis events and the Covid'19 dataset. In addition, the authors performed in-domain

experiments in which the training and the test data were related to the same crisis events. The promising results show that the topic-to-label framework could be used to classify tweets for crisis response. However, the authors did not examine the ability of their framework to classify crisis-related tweets effectively in domain adaptation settings (cross-domain or out-of-domain). Thus, no target data were used in their study, unlike in our work.

2.2 Distant Supervision

Distant Supervision (DS) is an approach to generate training data. It is an alternative method to reduce labor cost. In DS, we use an existing knowledge bases to label training data for a specific task [119]. Traditionally, DS data has mainly been used to train models in the absence of manual annotations or as additional training data to improve the generalisation performances of deep learning classifiers requiring extensive training data.

DS is first introduced in [89] via an external source as a combination of the method in [117], wherein (is-a) relations between entities were discovered using WordNet and the method in [42], in which weakly labelled data were used in bioinformatics. Since 2009, DS has been successfully applied to label training data via an external knowledge base in many NLP tasks, such as event extraction [36, 138], sentiment analysis [52], topic classification [90, 85] and relation extraction [104].

DS typically uses the following components to give corpora pseudo-labels (unreal labels): seeds, selective methods, external sources, noise reduction and negative example generating. Specifically, several resources are used to label positive training data based on certain language rules or restrictions extracted from the seeds. Noise reduction techniques are then applied to reduce noise in the generated data. Finally, negative examples are generated to be utilised with positive ones to train a given classifier. Note that, in the field of automatic labelling of text, the term "pseudo-labelling" is used with a different meaning than the term used in this thesis. According to Lee in his paper [77], pseudo-labelling uses human-labelled data to train a model to predict the labels for unlabelled data. However, we use the term

to describe the process of giving pseudo-labels to the unlabelled tweets from target events.

Some selective methods have been used in the text classification field to select essential words or sentences. Pointwise Mutual Information (PMI; [40]) is a well-known method to calculate the importance of a feature (keyword) in a category (class). In the context of event detection, Win and Aung have used the mean PMI in [134] to select the most related features in the disaster lexicon as

$$MeanPMI = PMI(t, informative) - PMI(t, non - informative)$$

where ***t*** is a given term (feature) and ***PMI*** for both informative and non-informative classes is calculated as:

$$PMI(t, orientation) = \log_2 \frac{\text{freq}(t, orientation).N}{\text{freq}(t).\,\text{freq}(orientation)}$$

where ***orientation*** is either informative or non-informative, freq(t) is the number of times term ***t*** appears in a tweet and ***N*** is the total number of terms in all tweets. If the ***meanPMI*** value is positive, then the term or the word is related to the disaster lexicon, and vice versa. Term Frequency-Inverse Document Frequency (TF-IDF; [65]) is another selection method that aims to measure the importance of a feature (word) to a given document (event type) in a given corpus (collection of tweets). TF-IDF is calculated as:

$$TF - IDF(t) = TF(t).IDF(t)$$

where

$$TF(t) = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

and

$$IDF(t) = \log \frac{\text{Number of documents}}{\text{Number of documents including term t}}$$

TF-IDF is widely used by search engines to retrieve the most relevant documents to the user query. It is also very efficient in removing stop words for text classification tasks. Key Rate (KR) is another selective method developed by Chen et al. in [36] to select the most important triggers and arguments of a specific event type for

an event extraction task. The KR value depends on two variables: role saliency (RS) and event relevance (ER), such that:

$$KR_{(}ij) = RS_{(}ij) * ER_{(}i)$$

and

$$RS_{(}ij) = \frac{Count(W_{(}i), ET_{(}j))}{Count(ET_{(}j))}$$

where $RS_{(}ij)$ reflects the appearance of an argument or a trigger $\boldsymbol{i}$ in representing a specific event type $\boldsymbol{j}$, $Count(W_{(}i), ET_{(}j))$ is the count of word $W_{(}i)$ in all the sentences related to event type $ET_{(}j)$, $Count(ET_{(}j))$ is the count of all the words in all the sentence representing event type $ET_{(}j)$, and

$$ER_{(}i) = \log \frac{Sum(ET)}{1 + Count(ETC_{(}i))}$$

where $ER_{(}i)$ measures the event relevance of the argument or trigger i. $Sum(ET)$ is the sum of all the event types and $Count(ETC_{(}i))$ is the count of all the event types, including argument or trigger i.

Several resources have been used in DS to label training data based on some language rules or restrictions. Freebase is an accessible semantic resource that uses Compound Value Types (CVTs), also called meditators, to combine several values into one value [30]. For example, Miami Heat is a CVT; location, member, founded and coach are instances; and Miami, National Basketball Association, 1988 and Erik Spoelstra are the respective values of these instances. Another widely used English-language external knowledge base is FrameNet, which consists of more than 1,000 semantic frames, with more than 100,000 Lexical Units (LUs), lemmas and Part-Of-Speech (POS) tags. Each frame in FrameNet is associated with a group of LUs that evoke that frame [25]. For example, in the sentence "The team took revenge with a resounding victory", revenge is an LU related to multiple frames such as agent, offender, injured party, injury and punishment. For the Arabic language, Almanny is one of the known dictionaries used in distant supervision for Arabic texts [87, 57] and recently for Arabic tweets [12, 15]. It is a comprehensive dictionary that provides meanings, synonyms (semantically similar words) and roots for Arabic words.

Noise is a recognised labelling problem when using distant supervision to label raw data. This problem can seriously affect the performance of deep learning models and hence has been well addressed in the literature. For the relation extraction task, Riedel, Yao, and McCallum [108] have introduced a multi-instance single-label model, assuming that each entity pair holds at least one relation expression. This work has been extended by Hoffmann et al. in [60] for a multi-instance multi-label model, where more than one label was allowed for each entity pair. In addition, noise has been reduced in other works via other approaches. Zheng et al. [140] filtered the noise in positive examples using a threshold for the frequency of the dependency paths among these examples. Li, Wu and Vijay-Shanker [78] and Su et al. [119] have applied three heuristic labelling methods initially proposed by Takamatsu et al. in [120]: top trigger words, closest pairs and high-confidence patterns. Chen et al. [36] have used two external knowledge bases instead of one to generate large-scale distant supervision data in the event detection literature. FrameNet has been used to eliminate noisy trigger words and expand the trigger list to include new triggers.

The simplest way to generate negative examples is to apply the assumption against distant supervision. For example, suppose the distant supervision assumes that every sentence contains at least one existing pair in the external dataset. In this case, the sentence expresses the relation and is labelled positive, and negative examples can be generated directly when the sentence has no such pairs. However, generating negative examples for distant supervision data requires great skill, and each case has its own rules and suitable ways to label negative examples from the raw data.

In this thesis, we create the seeds (initial keywords) from available labelled data (for English tweets) and clusters (for Arabic tweets). To select the essential words from the initial list, we use the key rate from [36]. However, we change some variables to suit the case of the binary classification task rather than the multiclassification task discussed in the prior work. We then apply distant supervision via external knowledge bases, using FrameNet and Almaany dictionary for English and Arabic tweets, respectively. For English, if one of the top crisis-type keywords exists as an LU of a frame in the database, then distant supervision assumes that all the LUs

related to the given frame express the given crisis type. For Arabic, distant supervision assumes that all semantically similar words to the top Arabic keywords express the crisis type for Arabic tweets. To filter the noise in our distant supervision data, we only consider tweets with two keywords from the final list instead of only one keyword. In addition, all the tweets containing only one keyword are ignored to reduce the noise caused by using keywords from FrameNet. To generate negative examples, we assume that the tweet with no keywords from the final list does not express the crisis type in any way; thus, we label them as negative tweets. A detailed explanation and justification is given in Chapters 5, 6 and 7.

2.2.1 DS Approaches for Text

From the literature, we can say that DS approaches differ in using one or more of the above components. In this thesis, we study the distant supervision approaches most related to our work, which are the approaches for automatically labelling textual data.

Since 2009, various approaches have been successfully applied to automatically annotating texts. Marchetti-Bowick and Chambers in their paper [86] have used pre-defined keywords to apply distant supervision on political tweets to determine the relevant subtopics, such as ideology. They assumed that if a keyword exists in the tweet, the relevant subtopic is assigned to the given tweet. Their topic classifier was trained with naive Bayes using the automatically created datasets. Besides the political keyword, the authors identified sentiment words to determine the sentiment of the post: a second classifier, the sentiment classifier, was separately trained with naive Bayes using the automatically labelled sentiment dataset. Ultimately, this two-stage model was used to first identify the subtopic and then the tweet's sentiment. The results showed that the performance of the topic identifier dramatically dropped from 90% to 10% when tested on general tweets containing political and non-political examples. For sentiment classification, it was evident that the combination of topic keyword and sentiment word led to a better aspect-based sentiment analysis approach. This work is related to ours in its use of keywords for applying distant supervision to tweets; however, we use different techniques and sources

from those used in this prior research. Our crisis keyword list primarily consists of essential and expanded keywords through external language resources, which improves the generalisation level of our classifiers. Moreover, while this paper used one word to identify the subtopic of the tweets, we use two keywords and discard all tweets containing one keyword.

Distant supervision techniques using text features like those used in this thesis have also been proposed in the literature. Go et al. [52] have discussed using tweets with emotion for distant supervision learning. They assumed that the emotions in texts express the feelings of the writers. For example, if the tweet contained a happy-face emoji, then the tweet was labelled positive. This assumption was used to label tweets for the sentiment analysis task. Results showed that using emotions as noisy labels is an effective DS as their best classifier reached 83% accuracy for tweets across all domains. A more recent study by Krommyda et al. [74] used Plutchik's eight basic emotions to annotate tweets. The authors identified the emotions expressed in the tweets using emojis, keywords and semantic relationships that appeared in the given tweet. As a result, they built a dataset of emotional tweets with eight categories and then trained an LSTM classifier using this dataset. Research by Mohammed, Ghelani and Lin in [90] has employed distant supervision for the topic classification task. Specifically, they transferred labels from tweets of topically focused accounts to tweets posted by general Twitter accounts. Magdy et al. [85] have used YouTube URLs for topic classification of Twitter data. They assumed that if the tweet contained a link to a YouTube video, then the topic related to its title was expressed in the tweet. Both researches yielded good tweet classifiers. In this thesis, we use essential keywords extracted from either labelled or unlabelled data, as we see that keywords are vital in classifying crisis tweets.

The most commonly used approach to automatic text labelling is to use available annotated data as a training set to automatically label data from exact domains. For example, Athira et al. [23] have applied self-training to an amount of human-labelled data to automatically label the remaining examples of the given medical dataset. The base classifier was built using 100 labelled medical tweets. They used traditional and deep learning methods, including SVM, K-Nearest Neighbour (KNN), CNN, LSTM and Bidirectional LSTM (BiLSTM) with BERT embedding.

The results showed that BiLSTM with BERT reported the best performance for classifying discussion topics in online health communities. In the context of this study, Win and Aung in their paper [134] have used a similar approach to classify crisis tweets to improve situational awareness. The researchers employed self-training learned on small available human-labelled data to label a new disaster collection, Myanmar _Earthquake_2016, derived from Twitter. They proposed a two-layer network: the first layer classified the tweets into informative and not informative, then the second layer classified the output tweets based on their information type. They applied word- and phrase-level features such as n-gram, POS tags and sentiment lexicon (disaster lexicon) with the LibLinear classifier. An annotation accuracy of 80% was achieved compared to the human-labelled dataset from the same event. De Carvalho et al. [35] have created an automatically labelled sentiment dataset for public security in a specific region in Brazil. They selected a general sentiment annotated collection of the Brazilian Portuguese language to be used as training data to classify tweets related to public security. To ensure the topic relevance of the tweets, they employed LDA on corpora posted from the same Brazilian region. Experiments were then run to train classifiers using different ML algorithms, and the results showed that SVM is the most accurate model. This work can be helpful in automatically labelling texts from languages with complex features. Importantly, Menini, Aprosio and Tonelli [88] have stated that context could be necessary to label text accurately. They created a dataset by re-annotating previously manually labelled texts into specific contextual categories for abuse detection of textual data. An end-to-end BERT model and a classifier were then applied to classify abusive texts in a specific context. The authors argued that a context-aware classification is more challenging but also more similar to real application scenarios. However, these existing works applied self-training to automatically annotate texts, which duplicates the label noise that exists in the training dataset. On the other hand, our work in this thesis explores different ranges of unseen features by expanding the original keyword list to include new linguistic units (new keywords with similar meanings) derived from FrameNet, which provides the opportunity to improve the generalisation level of the classification model.

Other distant supervision approaches are related to this thesis in their use of

external resources. Chen et al. in [36] have employed distant supervision to automatically generate a large-scale dataset using an external knowledge base for event extraction tasks. Their approach comprised four components: key argument detection, trigger word detection, trigger word filtering and expansion, and automatically labelling data generation. First, they selected the important key arguments for each event topic using FreeBase. These key arguments were then used to label texts from Wikipedia data preliminarily; these sentences were later used to extract trigger words. The linguistic knowledge base (FrameNet) was then used to filter the noisy trigger words and further expand the list. The data-generated component assumed that if the sentence from Wikipedia data contained the key argument detected in the first component and any trigger words from the filtered expanded list in the third component, then the relevant event topic was assigned as a label to the sentence. This approach was tested on the widely used ACE05 dataset. The authors stated that the automatically labelled dataset was comparable to human-labelled data regarding the quality of the data. The results showed that training neural models using a combination of distant and available manually labelled data improved the performance of the event extractor.

Zeng et al. [138] have argued that detecting triggers is not essential for determining the event type for event extraction tasks, unlike key arguments. They extracted the most related arguments that best described the event from existing structured knowledge (FreeBase). Event topics could be automatically inferred using the valuable information of event arguments in the structured tables provided by FreeBase. Thus, they assumed that if an argument existed in the text, then the event topic of this argument was assigned to the given text. They created a distant dataset by employing their assumption on Wikipedia data. Although this work used the same test data as [36] and reported similar findings and results, they simplified the work, reducing the time and effort needed to apply such approaches. While previous studies have used FrameNet and FreeBase for English data based on the relations between words in well-formed texts, we use FrameNet for English ill-formed texts based on the existence of essential keywords in the LUs of a related form. As a result, the techniques used after applying distant supervision to data are different from those in the

aforementioned studies. The same approach is followed to expand the initial keyword list to include new keywords from an Arabic resource (Almaany dictionary) for Arabic tweets.

In Arabic NLP research, most studies have concentrated on sentiment analysis, which involves labelling a sentence or a word with positive, negative or neutral labels. Most strategies for automatically annotating corpora use a sentiment lexicon [54, 55, 62]; automatic translation from English to Arabic is used to create the lexicon, based on which positive and negative tags are automatically annotated. Alzanin and Azmi [20] have reported an expectation-maximisation-based semi-supervised method for detecting rumours in Arabic tweets using solely "news" tweets for rumour detection, which can be considered a labelling technique.

2.3 Transfer Learning

According to Torrey and Shavlik in their book [125], Transfer Learning (TL) refers to transferring the knowledge from a learned task to a new task, the main goal of which is to leverage the knowledge between source and target domains to improve the learning process in the target domain, as shown in Figure 2.2.

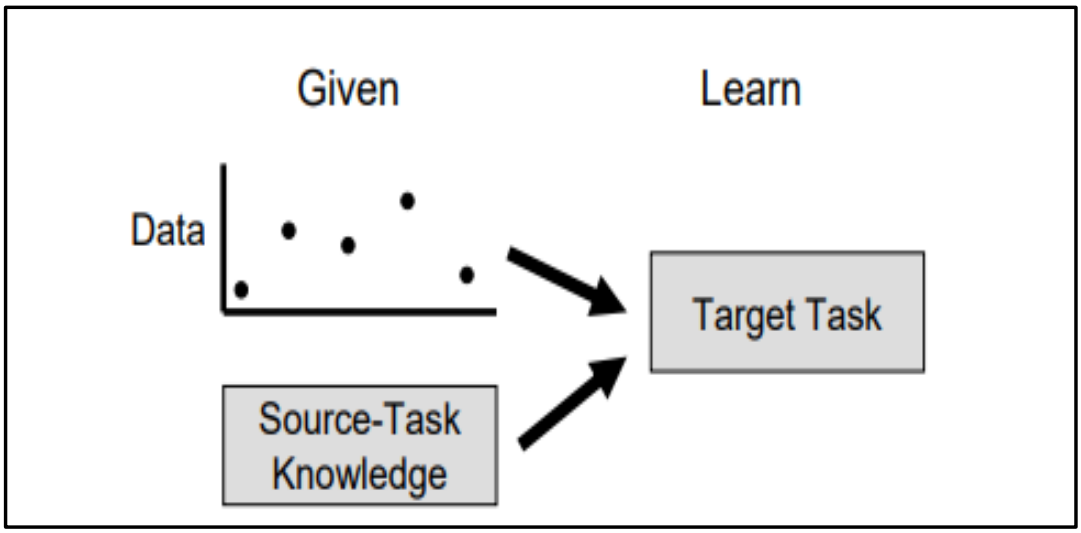

FIGURE 2.2: Transfer Learning. Source: [125].

Transfer learning is described as the process of improving a given predictive function fT(.) in a given domain (Dt) using the knowledge in another domain (Ds)

and another learning task (Ts), where Ds is the source domain, Dt is the target domain, Ts is the learning task on the source domain, Tt is the learning task on the target domain and $Ds \neq Dt$ or $Ts \neq Tt$ [98]. The two main paradigms for transfer learning in NLP are resource-based transfer and model-based transfer [136]. Resource-based transfer uses extra linguistic annotations for transfer learning. Model-based learning exploits a previously learned model on a given task to achieve another task by developing the similarity between them and adapting the feature representations or model architectures. Unlike resource-based transfer, model-based transfer does not utilise additional resources.

This thesis discusses cross-domain transfer through models and resources, also known as domain adaptation.

2.3.1 Domain Adaptation

Domain adaptation is a particular case of transfer learning. Given a domain D = x, P(X) and a task T = (y, P(Y | X)) where Ds is the source domain, Dt is the target domain, Ts is the source task, and Tt is the target task, domain adaptation is when $P(X_(s)) \neq P(X_(t))$ in which the marginal distributions of the source domain P($X_(s)$) and the target domain P($X_(t)$) are dissimilar, and $X_(s)$ and $X_(t)$ are from different topics [98]. Domain adaptation is a sub-discipline of machine learning that addresses scenarios in which a model trained on a source distribution is used in the context of a different (but related) target distribution. Domain adaptation, also known as cross-domain transfer, aims to learn a classifier that transfers knowledge from a source domain to a target domain. The domains may have similar or disparate label sets.

Standard NLP models can be trained on a large source domain and used on a limited target domain. However, some cases face domain shift, which causes a gap between the source and the target domains, especially if the similarity level decreases. The main purpose of domain adaptation techniques, therefore, is to reduce the gaps between the source and target domains caused by domain shift or distribution change when transfer learning is applied [133]. Domain adaptation approaches can be categorised based on their characteristics, such as the number of steps used

(one-step, multi-step) or the type of target data (supervised, semi-supervised, unsupervised). However, this thesis focuses on a specific one-step, semi-supervised domain adaptation method using pre-trained models.

2.3.1.1 Domain adaptation methods using a pre-trained model

Features of state-of-the-art models pre-trained on a large dataset can be used for a new task while either tuning or fixing the pre-trained parameters. Specifically, finetuning pre-trained models can be achieved through one of the following three techniques:

1. Copying the model architecture. The architecture used to train the source data and which proved successful in classifying similar texts is then transferred to train the target data.

2. Feature extracting. This fixes the weights of all the layers except the output layer, the last layer of the pre-trained model. With all other layers frozen, the last layer is replaced with a new layer to solve the given task. In this technique, the source and target domain are assumed to share high-level features.

3. Finetuning. This involves updating the weights of all or some of the layers of the pre-trained model. Unlike feature extraction, this technique assumes that the source domain is different from the target domain.

2.3.1.2 Approaches to DA using pre-trained models

Previous studies outlined in Section 2.1 have suffered from generalising to new crisis events; model performance dropped when tested in cross-domain settings. For this reason, several works have introduced online domain adaptation approaches for disaster response to reduce the domain shifts between tweets from past and emerging events. Zhang and Vucetic in their work [139] have assumed that a small amount of target-labelled data is available from the event onsite. However, according to the authors, the performance of a classifier trained on limited labelled data is low. To remedy this issue, they proposed a supervised domain adaptation model that used a large unlabelled corpus from the target domain to create word clusters, which were

then used for feature extraction to train a logistic regression classifier. They tested their approach on CrisisLexT6 and changed the amount of labelled and unlabelled data used for each experiment. Their experimental results showed that this proposed approach did not constantly improve the classifier's performance compared to classifiers that used the bag-of-words feature. Their findings also showed that using more labelled training data also improved performance.

Another supervised domain adaptation approach was introduced by Nguyen et al. [93], who used the CrisisNLP dataset to build a model with a single CNN layer after a look-up layer and before a pooling layer. A dropout layer was then added so that each node was removed with the probability of *1 - p* or kept with the probability of *p* only in the training time to avoid training all the nodes, thus reducing the overfitting problem. The authors trained the initial model first and then finetuned the weights of the last layers (freezing the initial layers) with small mini-batches of an emerging event online to suit the early crisis response situation. However, they assumed some manually labelled tweets during the event's occurring time, which we think is infeasible. This model has been tested on the Nepal Earthquake dataset as the out-of-domain tweets, where the training dataset contained tweets related to the same crisis type (Chile earthquake). The results showed that manually labelled data (labelled by paid crowdsourcing) is better than those labelled by volunteers. Furthermore, the model's performance dropped after approximately 2,000 tweets due to both binary and multi-classification problems. In addition, many NLP studies have finetuned pre-trained large models using manually labelled tweets in TCFCR [9]. However, supervised domain adaptation models assume that limited human-labelled data is available for the target event, which is infeasible in real time.

Li et al. [82] have proposed a semi-supervised domain adaptation approach that did not require limited labelled data from the target domain. Rather, it required labelled source data and unlabelled target data from three classification tasks with different datasets. The authors used a pre-trained model on one crisis dataset to classify tweets from an emerging (current) event – to be added to the training data in the retrained stage. Their iterative self-training method showed good results, particularly when classifying tweets related to a specific crisis. The authors then extended their work in [81] by comparing naive Bayes and self-training with hard

labels (NB-ST) to naive Bayes and expectation-maximisation with soft labels (NB-EM) in classifying tweets related to an emerging crisis. The domain adaptation classifiers were compared with their corresponding supervised classifiers learned only from labelled data from the source domain source. The findings indicated that using unlabelled target data resulted in better adaptation performance. Moreover, the authors compared the F1 scores of 11 event pairs for cross-domain adaptation settings and showed that the adaptation between similar event pairs is likely to deliver better performance. When comparing NB-ST and NB-EM, the results indicated that NB-ST is generally better than NB-EM when evaluated on the CrisisLexT6 dataset.

Proven to outperform other proposed semi-supervised domain adaptation approaches in the literature, the authors continued their work with iterative self-training. Lie et al. [80] combined self-training with deep learning models (CNN and BERT) and tested the classifiers on three different crisis datasets for Twitter data. Their results highlighted that self-training could help in improving the performance of CNN and BERT classifiers with large, not small, unlabelled target data. This thesis compares our work (distant supervision-based approach) to the iterative self-training approach proposed by Li et al. in [82] which is considered the current state-of-art domain adaptation method for TCFCR. We also combine distant supervision-based approach with BiLSTM, which leads to similar findings, indicating that self-training does not improve performance when combined with deep learning models on small unlabelled target data. Unlike this outcome regarding self-training, our proposed approach in this thesis improves the adaptation performance when combined with BiLSTM using small unlabelled data (50 examples for each class) and is therefore suitable for situations where time is critical, such as crisis response.

Although previous works have shown good results in using domain adaptation for crisis response, room for improvement remains to reach robust performance of supervised target classifiers when labelled target data are not available [79]. According to Wang and Deng in [133], domain adaptation can be achieved by building a target model using manually labelled source data with pseudo-labelled target data. However, to the best of our knowledge, there have been no works on domain adaptation approaches that use a distant supervision-based framework to classify crisis-related tweets from an emerging event. Thus, our work investigates using a distant

supervision-based framework to give unlabelled emerging tweets pseudo-labels to be then incorporated into labelled source data from several similar past events as a means to build a robust disaster-related classifier and compare it to the widely used automatic labelling technique (iterative self-training).

Wang, Nulty and Lillis [132] have argued that domain adaptation could be achieved without using unlabelled or labelled data from target events. They finetuned T5 seq2seq models – introduced earlier by Raffel et al. in [106] – using tweets from source crisis events and task descriptors and event descriptors for both target and source events. Specifically, the training data used to finetune the model consisted of three parts: the tweet text, the event descriptor, and the task descriptor. The crisis descriptor included the location and the name (type) of the crisis. In turn, the authors constructed the input example in the form of a question-answering sequence like "Contents: ***{tweet_text}***, Question: Is this tweet related to ***{location_name} {crisis_name}***?". During the testing phase, preliminary information was given about the task and the event of the target disaster. Experiments were run with in-domain and cross-domain settings from single- or multiple-source events. The results showed that using multiple similar source events to the target disaster improved the adaptation performance. The authors noted that selecting source events was critical when testing models in domain adaptation settings; in other word, using dissimilar events to the target event may harm the classifier's performance. Our work differs from this study in using unlabelled target data during the training phase, which is proven to be a valuable target resource at run-time.

2.4 Summary

In this chapter, we reviewed the current literature in the field of text classification related to distant supervision and domain adaptation. We mainly focused on texts from Twitter data, starting with tweet classification models that aim to help organisations respond to people during a crisis. We found several key points in our review of tweet classification for crisis response. First, current tweet classification models suffer from the lack of labelled training data, which prevents them from reaching a generalised model, as tweets related to various crisis events have different features

and social media responses. However, it is infeasible to manually annotate tweets for every crisis event, especially in real time.

Given their ability to create a large-scale training dataset with no effort, time or cost, distant supervision can help improve the generalisation level of crisis-related classifiers by automatically generating new labelled training data from an unlabelled corpus. However, to the best of our knowledge, the application of distant supervision has not been investigated in labelling unseen tweets from new/emerging crisis events, and no research exists to report the impact of using distant supervision in a domain adaptation approach.

Existing keyword-based systems use predefined keyword lists to label current disaster events and fail to identify tweets that contain other essential crisis keywords. On the other hand, distant supervision has the capacity to expand the predefined keyword list to include linguistically similar words. A domain adaptation approach that utilises distant supervision to give labels to unlabelled current tweets can hence detect crisis data using keywords that do not exist in the predefined list.

Several supervised domain adaptation techniques have been introduced in the literature. However, they require a certain amount of manually labelled data from the target incidents. This causes a delay in the work of the humanitarian organisation in real-time scenarios while paid workers label the incoming tweets; time and shortness of workers are the essential issues in adapting the application of Twitter data in their daily operations. Conversely, semi-supervised domain adaptation techniques do not require any labelled data from target events that suit real-time scenarios.

The widely used labelling method in the literature is self-labelling. Self-labelling is utilised to incorporate unlabelled target tweets under the umbrella of a semi-supervised domain adaptation approach. However, error amplification is a considerable drawback of this technique, especially when past and emerging crisis events differ. This can be minimised by incorporating tweets that contain different words with similar meanings, which can be achieved by employing distant supervision in a labelling framework instead of self-labelling.

Arabic is considered a low-resource language with many noticeable challenges in the crisis literature. Relatively few techniques have been introduced to address

the problem of limited labelled data for Arabic crises. However, at the start of this study, domain adaptation approaches had not been applied or experimented with using Arabic crisis tweets. We posit that domain adaptation techniques employing distant supervision can improve performance regardless of the limited labelled crisis data by incorporating tweets with diverse words from the emerging events into the available training data. Thus, this thesis attempts to fill the gaps in the extant literature.

Chapter 3

Deep Learning and Word Embeddings for TCFCR

This chapter represents our investigation to find the best deep learning architecture and word embedding for TCFCR as a start point in our research. This work has been published (in 2018) and cited by many research papers in the field. Later on, our findings in this chapter have been verified by another research paper with a comprehensive study.

3.1 Introduction

Multi-class tweet classification for crisis response is a text classification task that identifies if a tweet is related to a specific type of predefined informative class. Deep neural networks have proven their ability to automatically learn deep and complicated mappings from input to output using the distributed representation of words without requiring any feature engineering. It is also noticeable that deep learning approaches have outperformed traditional ones in many NLP tasks, including tweet classification [34]. In addition, robust word embedding can be a key factor in improving the neural network performance in any NLP task [92]. Recently, general-purpose and domain-specific word embeddings have been proposed, such as GloVe and Crisis embedding [93]. However, few experiments have been conducted to examine the effectiveness of different deep learning architectures and different word embeddings in improving tweet classification models. Our work is similar to [92] and [93]; however, we use different neural network architectures and

general-purpose word embeddings. We also train our four models separately in an offline fashion without integrating any network components. To the best of our knowledge, at the time of the publication of our paper in 2019 [16], this is the first study of using GloVe embedding with BiLSTM or CNN in tweet classification for crisis response.

The input of our networks are tweets that may contain any information related to any natural crisis. This work targets natural crises such as earthquakes, floods, storms, typhoons, and etc. First, we clean the tweets by removing unnecessary parts such as emojis and HTTP addresses. Then, the tweets are tokenised into words, and a pre-trained word embedding (GloVe or Crisis) is used to capture similarities between words and semantics of word sentences. After that, a deep learning architecture (CNN or BiLSTM) is applied to encode and leverage the information from the input text sequence, tweets. Finally, a fully connected layer with a softmax layer is used to compute the class distribution for each tweet.

3.2 Experiments

3.2.1 Models

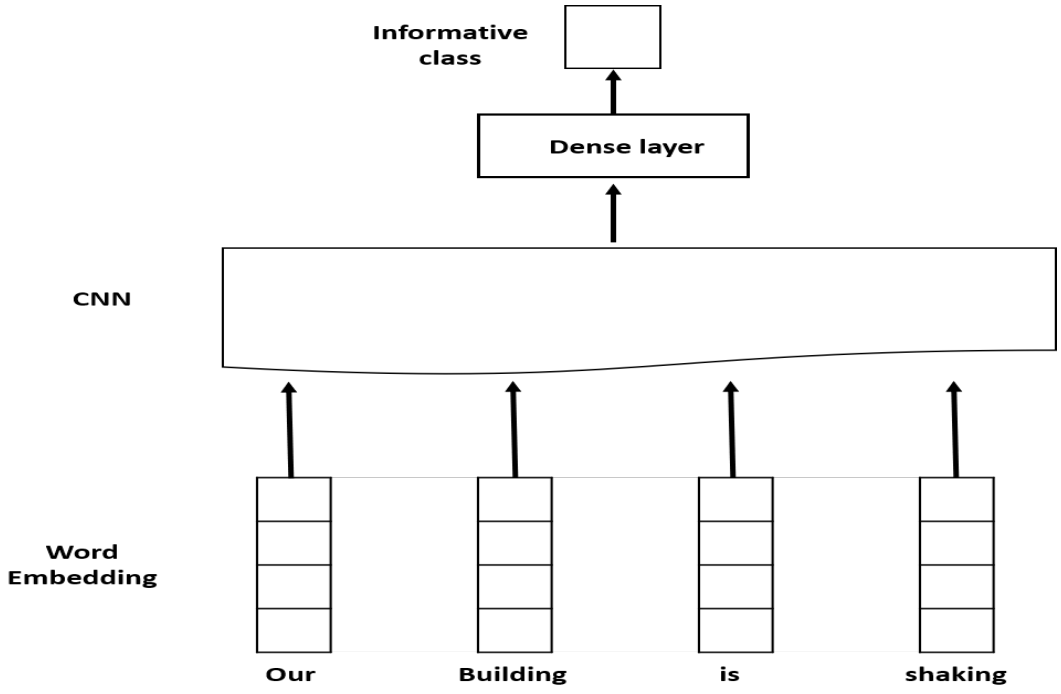

FIGURE 3.1: CNN architecture with word embedding.

We conduct four experiments using different word embeddings (Crisis embedding and Glove) and deep learning architectures (CNN and BiLSTM). We have re-implemented the CNN and Crisis embedding model from [93] to compare it with the other three models to investigate the effectiveness of integrating different word embeddings with different deep learning architectures. Figure 3.1 describes the first and the second classifiers used CNN with GloVe and Crisis embedding separately. We use BiLSTM in the third experiment with GloVe embedding and Crisis embedding in the fourth experiment, as shown in Figure 3.2.

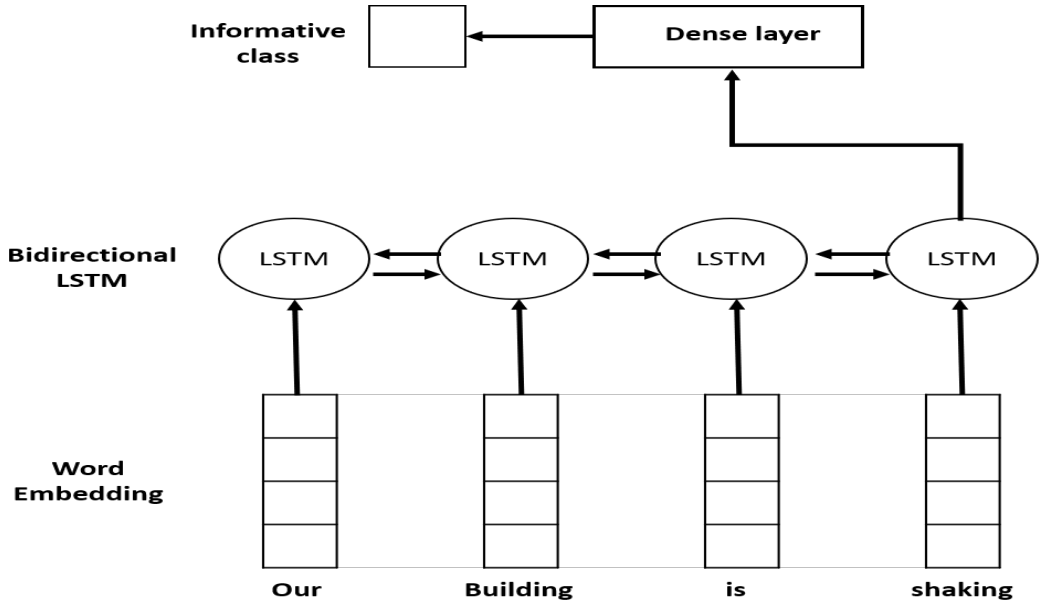

FIGURE 3.2: BiLSTM architecture with word embedding.

3.2.2 Datasets

We use the CrisisNLP dataset [64] to evaluate the four classifiers mentioned in the previous section. CrisisNLP is a collection of small datasets where each dataset contains annotated tweets related to a crisis event. The tweets are labeled based on their corresponding informative class (e.g., affected individuals, donations and volunteering, infrastructure and utilities, sympathy and support, other helpful information, and irrelevant). The number of tweets for each set is shown in Table 3.1.

TABLE 3.1: Training, validation and testing data used in the experiments.

Class number	**Class title**	**Number of tweets**		
		Train set	**Dev set**	**Test set**
1	Injured of dead people	1611	487	233
2	Missing, trapped or found people	741	221	106
3	Infrastructure and utilities damages	676	177	94
4	Sympathy and emotional support	1526	436	232
5	Donation needs or offers or volunteering services	2352	712	350
6	Other useful information	5690	1623	766
7	Irrelevant	6254	1756	886
Total number of tweets		18850	5412	2667

3.3 Results and discussions

CNN with Crisis embedding model achieved a F1 score of 0.6138, slightly higher than the model containing BiLSTM with the same embedding. On the other hand, BiLSTM with GloVe embedding reported the best result among all the four models with a 0.6204 F1 score, and CNN with GloVe embedding recorded the worst performance with a 0.5987 F1 score. The results of all the experiments are shown below in Table 3.2.

TABLE 3.2: Results in F1 score for different deep leaning architectures and word embeddings

Experiment	**Model components**		
	Deep Learning architecture	**Word Embedding**	**F1-score**
1	CNN	Crisis embedding	0.6138
2		GloVe embedding	0.5987
3	BiLSTM	Crisis embedding	0.6088
4		GloVe embedding	0.6204

According to the results, BiLSTM with GloVe model obtains the best performance for text classification for crisis response. That demonstrates the effectiveness of general pre-trained word embedding such as GloVe and sequence models such as BiLSTM in improving the classifier's ability to distinguish between crisis-related tweets. However, domain-specific embedding outperforms general word embedding when

integrated with CNN. This shows the importance of choosing a word embedding depending on the selected deep learning architecture for tweet classification.

We believe that the main reason behind these results is that Crisis embedding is initially built using the Skip-gram model of the word2Vec tool, which is a powerful method for detecting the semantic meaning of words with a small semantic space. On the other hand, the GloVe embedding needs more information than the Crisis embedding to detect the semantic meaning of words successfully. This is consistent with the fact that BiLSTM captures more information than CNN. BiLSTM captures the sequence of tweets in both directions, while CNN captures the local patterns of tweets and may lose some information, such as the order of the words in tweets.

Another possible reason is that Crisis embedding may contain twitter-specific text irregularities such as emojis, mentions, hashtags, and other domain-specific words. These were not taken into consideration when training GloVe embedding. Because we perform pre-processing for our dataset and remove such words, the performances of both GloVe and Crisis embeddings are expected to be close. However, misspelling words such as 'flods' for 'floods' can only exist in Crisis embedding, which gives it a minimal advantage over Glove embedding.

Finally, we discovered that GloVe as a general word embedding can be used instead of Crisis embedding as a domain-specific word embedding to improve the performance of the BiLSTM-based model to classify tweets for crisis response.

3.4 Related Work

Recently, a limited number of experiments have been reported on successfully applying deep learning architectures and word embeddings to tweet classification for crisis response. It started when the authors in [93] argued that the informative class in the previous studies still has much information to be handled by organisations. To simplify the organisations' work and save their time and effort, they introduced a model that classified the informative class into multiple subclasses (e.g., infrastructure damages, affected people, donation and volunteering, sympathy and support, and other helpful information). This work is very similar to our first model, where the authors build their model with a single CNN layer after a look-up layer and

before a pooling layer. After that, a dropout layer is added to reduce the model's overfitting. However, they trained the initial model first. They then retrained it with small mini-batches in an online fashion to suit the early crisis response situation where we use their pre-trained word embedding (Crisis embedding) without retraining the model. The same model (CNN and Crisis embedding) has also been used by the authors in [92]. However, they integrated the domain-specific word embedding with Google word embedding, and the results reported a slight improvement in the model's performance. Another deep learning model has been introduced in [31]—the semantically-enhanced dual-CNN consists of a semantic layer that captures the contextual information and a traditional CNN layer. The results show that the dual-CNN model has a comparable performance with a single CNN.

3.5 Conclusion

This work investigates the effect of using domain-specific and general word embeddings with two deep learning architectures: BiLSTM and CNN. Results reported that using different word embeddings slightly improves the model performance due to the variability of the corpora used when building the word embeddings. Further experiments will be done to examine the effectiveness of N-Gram CNN, another architecture introduced in [71], in classifying tweets for crisis response. In addition, we will consider recent works in integrating general word embedding such as GloVe for rich semantic representations of general words and domain-specific embedding for domain-specific words such as ill-words within tweets in our case.

Chapter 4

Experimental Setup

This chapter explains the experimental setup used in this thesis: the datasets and data collections, including human-labelled, automatically labelled and unlabelled data, the pre-processing of the data and the specifications of the model used to classify the test data. Some of the datasets, network architectures and evaluation metrics described in this chapter have been used by previous works cited in the previous chapter.

4.1 Datasets and Data Collections

4.1.1 Human-labelled Datasets

4.1.1.1 English tweets

We use specific collections from the CrisisNLP [64], CrisisLexT6 [95] and CrisisLexT26 [96] datasets to evaluate our framework, as shown in Table 4.1. CrisisNLP contains human-labelled English tweets from crisis events from 2013 to 2015. CrisisLexT26 and CrisisLexT6 contain human-labelled tweets from crisis events from 2012 to 2013. These widely used datasets are labelled by paid workers based on either their relatedness to a given crisis event (CrisisLexT6) or their corresponding informative class (CrisisNLP and CrisisLexT26; e.g., affected individuals, donations and volunteering, infrastructure and utilities, sympathy and support, other useful information or not related). However, for the CrisisNLP and CrisisLexT26 data, we relabel the available tweets into two classes: related and not related to a given crisis event. First, we combine all the tweets containing similar information, such as “Personal updates”

and "Affected individuals". We then relabel all the tweets from the four classes except "not related" to a related class; "not applicable" and other unclear labels are discarded. We also eliminate the non-English tweets, as our main goal is to build a reliable model for English tweets only, although it may then be transferred to other languages. It is worth noting that the Malaysia Crash (MMC) data, unlike the other human-labelled data, are labelled by volunteers rather than paid workers.

TABLE 4.1: Summary of the human-labelled English data from CrisisNLP, CrisisLexT6 and CrisisLexT26. The abbreviations in the table represent the type of the data, the place of the crisis and the crisis type. For example, MGC represents manually labelled data for the Glasgow Crash event.

Collection	# related tweets	# not related tweets	Total # tweets
Bohol Earthquake (MBE)	969	30	999
Queensland Floods (MQF)	919	280	1199
Colorado Floods (MCoF)	924	74	998
Manila Floods (MMF)	920	79	999
Alberta Floods (MAF)	982	17	999
Yolanda (Tornado) Typhoon (MYT)	939	108	1047
Sandy (Hurricane) Typhoon (MST)	1581	429	2010
Oklahoma (Tornado) Typhoon (MOkT)	1769	241	2010
Nepal Earthquake (MNE)	2839	177	3016
Chile Earthquake (MChE)	1648	364	2013
California Earthquake (MCE)	169	13	182
Pakistan Earthquake (MPE)	1676	336	2012
India Floods (MIF)	1500	502	2002
Pakistan Floods (MPF)	1985	27	2012
Hagupit Typhoon (MHT)	1779	233	2012
Pam (Cyclone) Typhoon (MPT)	1515	497	2012
Odile (Hurricane) Typhoon (MOT)	178	4	182
Lac-Mégantic Crash (MLC)	537	18	555
Glasgow Crash (MGC)	918	181	1099
New York Crash (MNC)	998	1	999
Australia Fires (MAFi)	949	249	1198
Brazil Fires (MBFi)	333	9	342
Colorado Fires (MCFi)	953	246	1199
Malaysia Crash (MMC)	70	65	135

For the experiments in Chapter 5, we use all the datasets mentioned in Table 4.1. In Chapter 6, we use specific human-labelled collections from Table 4.1, including different crisis events related to three crisis types: earthquake, flood and typhoon. It is noted that we do not use collections related to crash or fire events. One minor change is that we use the other collections for the Hurricane Sandy event (MST) and the Oklahoma Tornado event (MOkT), since more data are available.

Prior studies state that supervised models trained on training/source data can classify test/target data when training/source and test/target data share a specific feature: the crisis type [47, 81]. Based on this finding and to obtain the best possible

performance, our training/source data in these experiments include multiple events from the same crisis type to the test/target event. To test a model trained only on test/target data, we split the data for each test/target event into training (70%) and testing sets (30%). The testing set is then used to evaluate all the models on the given test/target events.

4.1.1.2 Arabic tweets

Kawarith is the first Arabic dataset which contains more than 12,000 Arabic labelled tweets for crisis response [9]. It is also a multi-label dataset (i.e., a tweet can be labelled with more than one class). However, only 4.4% of the tweets in Kawarith have more than one label. This dataset provides seven collections of crisis events that occurred in Arabic cities. The tweets are labelled by native Arabic speakers based on their corresponding informative class: "Affected individuals and help", "Infrastructure and utilities damage", "Caution, preparations and other crisis updates", "Emotional support, prayers and supplications", "Opinions and criticism" and "Irrelevant". However, for the experiments in Chapter 7, we relabel the tweets into two categories: "related/informative" and "not related / not informative" to a given crisis event or humanitarian organisation. We relabel all the tweets from the categories "Affected individuals and help", "Infrastructure and utilities damage", and "Caution, preparations, and other crisis updates" to the related/informative class. Tweets in the "Emotional support, prayers and supplications", "Opinions and criticism", and "Irrelevant" categories are relabelled to the not related / not informative class. If the tweet has been labelled with more than one class, we pick the first one.

Table 4.2 demonstrates the number of related and informative tweets and the number of irrelevant or not informative tweets for each crisis event in Kawarith. We split the data for each Arabic target event into training (70%) and testing (30%) sets to evaluate a model trained solely on test/target data. In turn, the testing set is used to assess all the models for the given target events.

TABLE 4.2: Summary of the human-labelled Arabic data used in our experiments from Kawarith. The abbreviations in the table represent the type of data, the place of the crisis and the crisis type. For example, MJF represents manually labelled data for the Jordan Floods event.

Collection	# related tweets	# not related tweets	Total # tweets
COVID-19 Virus (MCV)	1363	136	1499
Jordan Floods (MJF)	525	965	1490
Kuwait Floods (MKF)	1266	1711	2977
Hafr Albatin Floods (MHF)	513	599	1112
Cairo Bombing (Explosion) (MCEx)	261	253	514
Dragon Storms (MDS)	305	476	781
Beirut Explosion (MBEx)	346	492	838

4.1.2 Data Collections

4.1.2.1 English tweets

For the experiments in Chapter 5, we use the Twitter API to collect unlabelled tweets from 2018 crisis events for five different crisis types: Texas Floods (UTF), Indonesia Earthquake (UIE), Sunda Strait Tsunami (Typhoon) (UST), California Fire (UCFi) and Amritsar Crash (USC). These events were selected based on four factors: availability, crisis location, crisis type, and number of tweets. For example, the Texas Floods event was selected because its data were freely available during our work with a massive number of tweets. In addition, Texas does not exist in the current location list of flood events in the human-labelled datasets. Therefore, we preferred to choose different locations to enhance the generalisation level of the model when added to the training data. Our proposed method is not restricted to these specific events and can be applied to any unlabelled tweets from any crisis event.

The 2018 Texas Floods, caused by extreme rainfall, hit Central Texas on October 16, 2018 and forced the government to declare a state of emergency after the floods claimed human lives. Texas Floods data are crawled for a five-day period, from October 16 to October 20, 2018, using the hashtag "#floods" and geolocation information of Texas, resulting in more than 48,000 unlabelled data (tweets).

We use the same methodology to collect 3,099 unlabelled tweets from the 2018 Indonesia Earthquake event, this time using the hashtag "#earthquake" and the geolocation information of the damaged area. The data are crawled for only one day, October 16, 2018, from 16:00 to 23:59.

The Sunda Strait Tsunami (Typhoon) struck Indonesia in 2018, killing at least 426 people and injuring 14,060. We collect 4,955 unlabelled tweets using the hashtag "#tsunami" and the geolocation of Sunda Strait. The data are crawled for the entire day of December 23, 2018.

The 2018 Amritsar Crash, in which 59 people were killed and more than 100 were injured, occurred in Amritsar, India, when two trains crashed into a crowd standing on the railways while celebrating the Dussehra Festival during the late evening of October 19, 2018. The Amritsar Crash data are crawled for three days, from October 19 to October 21, 2018, using the hashtag "#amritsartrainaccident" and the corresponding geolocation. As a result, we collect 3,033 unlabelled tweets from the event.

The 2018 California Fires are known as the deadest wildfires in the state's history. The fires covered a total of 1,893,913 acres and caused substantial infrastructure damage. The California Fires data are crawled for three days from November 12 to November 14, 2018. We use the hashtag "#wildfires" and the geolocation of California to collect 2,965 unlabelled tweets.

On the other hand, CrisisNLP provides tweet IDs for several previous crisis events. We retrieve the tweets' text content using their IDs. We create corpora for four crisis events: Pakistan Earthquake (UPE), Pakistan Floods (UPF), Hagupit Typhoon (UHT), and Malaysia Airline Accident (Crash) (UMC). It is worth noting that we cannot retrieve all the tweets by their IDs, as some of the tweets or the Twitter account of the posted person have been deleted; in other cases, and due to recent changes to the Twitter platform, the tweets have been protected or disabled. Table 4.3 shows the number of unlabelled tweets gathered for the nine target corpora after the cleaning process. These collections are related to four crisis types: earthquake, flood, typhoon and crash.

To train the model for the experiments in Chapter 5, we generate automatically labelled data by applying our framework, in Section 5.1, to the unlabelled data collections in UPF, UPE, UHT and UMC from Table 4.3 and the collected data. As a result, 5 new crisis: AIE, ATF, AST, ACFi and AAC and 4 prior crisis events: APE, APF, AHT and AMC are generated (A represents Automatically). We randomly select examples from the automatically labelled collections to approximately match the

TABLE 4.3: Number of unlabelled tweets gathered by their publicly available IDs.

Crisis event	**# unlabelled tweets**
California Earthquake (UCE)	5430
Chile Earthquake (UChE)	10685
Hagupit Typhoon (UHT)	12807
Nepal Earthquake (UNE)	7432
Pakistan Earthquake (UPE)	8954
Pam Cyclone (Typhoon) (UPT)	5195
Pakistan Floods (UPF)	6352
Queensland Floods (UQF)	6088
Malaysia Airline Accident (Crash) (UMC)	5566

number of tweets of the other human-labelled training datasets from the same crisis type. For example, we randomly select 1,600 positive examples and 200 negative examples from the automatically labelled data from the Indonesia Earthquake event (AIE) to be added to the training data along with the manually-labelled data from the available earthquake events: Bohol Earthquake (MBE), Nepal Earthquake (MNE), Chile Earthquake (MChE), California Earthquake (MCE) and Pakistan Earthquake (MPE). The number of positive tweets in most events from this collection ranges between 1,650 and 2,800 (and between 180 and 400 for negative examples). On the other hand, for APE, APF, AHT and AMC, we randomly select examples to exactly match the number of positive and negative tweets from the human-labelled peers. For instance, 1,676 positive and 336 negative examples are randomly selected from APF data to match the number of tweets mentioned in Table 4.1 for the manually-labelled data from the same event, MPF.

On the other hand, in Chapter 6, we use our distant supervision framework (see Chapter 5) to give pseudo-labels to the unlabelled tweets from eight crisis events: UPE, UPF, UHT, UCE, UPT, UQF, UNE and UChE mentioned in Table 4.3 except UMC (unlabelled tweets for Crash crisis type). As a result, eight automatically labelled sets are created: APE, APF, AHT, ACE, APT, AQF, ANE and AChE. To replicate the real-time scenario, we sort the tweets based on their IDs, which reflect their posting order. To train the model, we take examples from the top of these collections to match approximately the number of tweets of the other human-labelled training datasets from the same crisis type.

4.1.2.2 Arabic tweets

Kawarith contains over one million unlabeled tweets from different dialects and locations for crisis events. These tweets were collected during 22 incidents between 2018 and 2020 from different disaster types. For the experiments in Chapter 7, we retrieve the tweet text from the public tweet ID in Kawarith. These tweets relate to previous but very recent Arabic crisis events. Specifically, we choose seven events (in Table 4.2) that serve as the gold standard in our experiments and create the following seven corpora: UCV, UJF, UKF, UHF, UBEx, UCEx and UDS. The number of collected tweets is restricted because some tweet texts are irretrievable. In addition, some of the posted tweets or Twitter accounts have since been removed, while others have been protected or disabled due to recent changes to the Twitter platform. However, we successfully gathered the number of tweets required for our experiments.

In Chapter 7, the proposed Arabic distant supervision-based framework is used to label the unlabelled Arabic corpora outlined in this section. The same methodology of selecting tweets in Chapter 6 is followed in Chapter 7. However, the number of automatically labelled tweets required in one of our experiments is 2,000 for each class . We could not reach this number for some of the Arabic crisis events (Cairo Explosion, Hafer-albatin Floods and Jordan Floods), which led to exclude these events in the given experiment (further analysis; Section 7.4.1).

4.2 Data pre-processing

Tweets are full of noise due to incomplete sentences or words, irregular expressions, ill-formed sentences or words, and out-of-dictionary words. Therefore, to improve the performance of the tweet classifiers, we pre-process the data before training the models. We follow [92] in cleaning English input tweets. We convert all the tweets to lower case and replace hyperlinks with "HTTP address", numbers with "D", usernames with "userID", and hashtags with "hashtag". We also remove all emojis, Twitter-specific words such as "RT", special characters, elongation and punctuation. All tweets are split into tokens to be passed to the classification model.

For Arabic tweets, we follow [9] in cleaning Arabic input tweets. We substitute hyperlinks with the Arabic word "رابط", which means HTTP address or URL. Similarly, we replace user mentions with "مستخدم", hashtags with "هاشتاق", and numbers with "رقم". Four types of letter normalisations are performed: (1) "إ ، آ، أ", the different forms of *alef* are normalised to "ا"; (2) "ئ، ى", forms of *elaf maqsora*, to "ي"; (3) "ؤ", a form of *waw*, to "و"; and (4) *ta marboutah* "ة ، ه" to "ه". This is done because users typically misspell *alef* and do not know the difference between *ta marbouta* and *ha* when these letters appear at the end of any Arabic word. As with English, we eliminate stop words, special characters, punctuation, Twitter-specific words such as "RT", elongation and emojis. We also remove non-Arabic characters, diacritics and short vowels.

4.3 Training Classification Model

This section provides details of the architecture utilised for training any of the data given in our experiments in this thesis.

4.3.1 Word Embedding

Word embedding is a set of feature learning techniques or language models in which texts (phrases or words) are mapped to real number vectors. The main goal of word embedding is to learn efficient and expressive text representations, such that similar words or phrases have similar representations that capture their semantic meaning [91]. The application of word embedding has drawn significant interest in NLP in recent years. Robust word embedding can be a critical factor in improving neural network performance in any NLP task [92, 16]. General-purpose word embeddings have recently been proposed, such as GloVe [101] and fastText [29] embeddings.

GloVe is a strategy that combines count-based (e.g., Principal Component Analysis, PCA) with direct prediction (e.g., word2vec) techniques. Word2vec is a predication-based embedding which is also a combination of two techniques: skip-gram model and Continuous bag of words (CBOW). Unlike word2vec, which only uses local information from words with local context windows, the GloVe method uses a combination of word co-occurrence information and global statistics to determine semantic

links between words in the corpus. GloVe employs the global matrix factorisation algorithm [101], which creates a matrix that encodes the presence or absence of words in a text. GloVe embedding is a publicly available embedding trained on 6 billion words from web texts and Wikipedia, including social media texts such as tweets. It has improved many NLP tasks for Twitter data, such as event detection [46] and sarcasm identification [48]. In Chapter 6, we use the 100-dimensional GloVe as a pre-trained word embedding for English tweets since it is the best reported deep learning architecture for tweet classification for English crisis response [16].

Facebook presented fastText in 2017 as an extension of word2Vec. Unlike word2Vec and GloVe, fastText can provide representations of words that do not exist in the training data: when a word does not occur during model training, its vector embedding may be determined by breaking it down into n-grams [29]. For example, if the word is "earthquake" and n = 3, then fastText produces the following representations: < ea, ear, art, rth, thq, hqu, qua, uak, ake, ke>. As a result, fastText considers misspelling and provides meaningful representations for rare words, whereas other embeddings ignore them. In addition, fastText Arabic embedding has been pre-trained using Arabic Wikipedia articles. It outperforms other embeddings in text classification [45], such as Twitter classification in healthcare applications [56], sentiment analysis [22, 66] and hate speech detection [21]. For this reason, we use fastText Arabic embedding for Arabic crisis tweets in this thesis.

The initial embeddings of GloVe and fastText were finetuned during the gradient modification of the deep learning model using back-propagating gradients in every experiment. Finetuning word embedding represents transferring the knowledge from the initial corpus (Wikipedia) where the embedding is built to our domain dataset (Twitter data).

4.3.2 Classification Algorithms

4.3.2.1 Convolutional neural networks

CNN is a deep learning architecture that consists of an input layer, multiple neural hidden layers and an output layer. The main layers of basic CNNs (as shown in Figure 4.1) are the convolution layer, the pooling layer and the fully connected layer.

Typically, in NLP tasks, token sequences are used as input to the convolution layer. Then, in the convolution layer, the CNN filters perform as n-grams over continuous representations. The size of the n-gram filters, the feature map, are then reduced in the pooling layer to minimise the computational costs. The most popular pooling operation is max pooling, where the largest element is picked from the feature map. Finally, the reduced n-gram filters are combined with subsequent network layers, a fully connected layer [71].

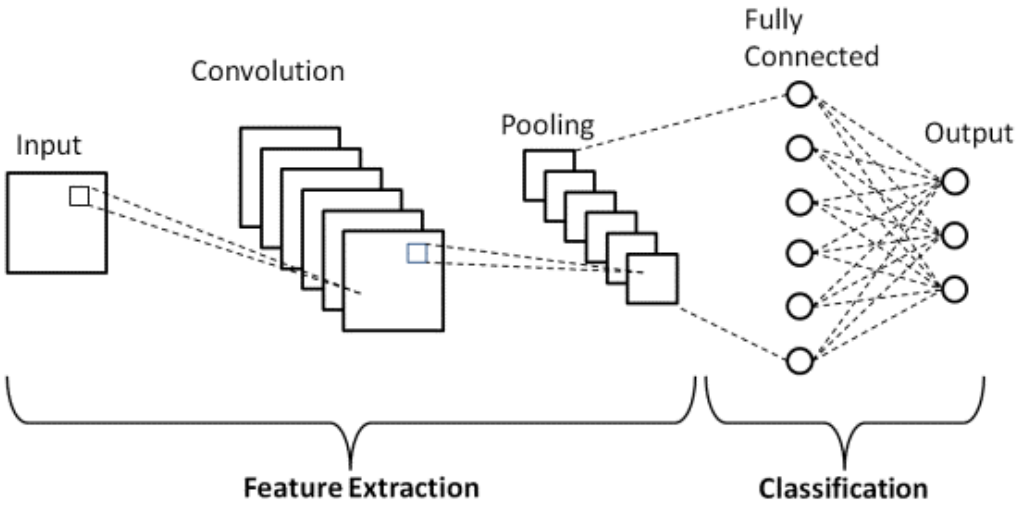

FIGURE 4.1: Basic CNN architecture. Source: [71]

Two further parameters are important in the architecture of CNNs: activation functions and dropout. Activation functions are non-linear functions; the most popular activation in CNN layers is the Rectified Linear Unit (ReLU). Dropout is where each node is removed with the probability of *1 - p* or kept with the probability of *p* only within the training time (to avoid training all the nodes).

CNN can learn the features and distinguish between them automatically; therefore, CNN does not require hand-engineered features, which saves human effort and time and eliminates the need for prior knowledge. Unlike a MultiLayer Perceptron (MLP), the number of free parameters can be reduced by CNNs, and the problem of the vanished or exploded gradients can be prevented during the training process. Moreover, all the weights in the convolutional layers are shared, which means that the same filter is used for all the fields within a layer to improve performance and decrease memory space required. Given these capabilities, CNNs can be successfully

used for feature extraction in several text classification problems [28].

4.3.2.2 Bidirectional Long Short-Term Memory (BiLSTM)

RNN is a type of artificial neural network adapted to work for data that involves sequences like texts because it stores the states or information of previous inputs to generate the next output of the sequence. However, RNN suffers from gradient vanishing/exploding problems where it loses the ability to propagate useful gradient information from the output end of the model back to the layers near the input end of the model. LSTMs are RNNs that solve the gradient vanishing/exploding problems of RNNs [59]. Unlike RNNs, LSTMs pass only the important information to the next layer [26]. LSTMs are designed to capture long-distance dependencies within texts. They hold the contextual semantics of each word using the surrounding information and store long dependencies between words. As shown in Figure 4.2, each LSTM unit consists of three gates – controlling which portions of information to remember, forget and pass to the next step.

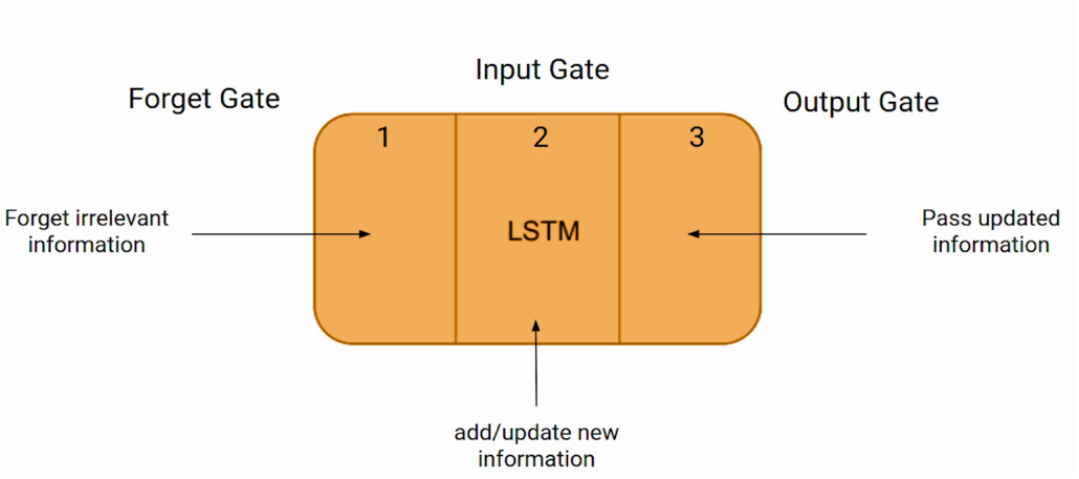

FIGURE 4.2: Basic LSTM architecture. Source: [26]

However, LSTMs only focus on one direction of the input: the past. On the other hand, BiLSTMs focus on the past and the future directions of the input, as shown in Figure 4.3. This method allows the network to capture more information than before; at every token position, hidden representations from each direction are concatenated. BiLSTM has hence been reported as the best deep learning architecture

for tweet classification for crisis response [16, 79].

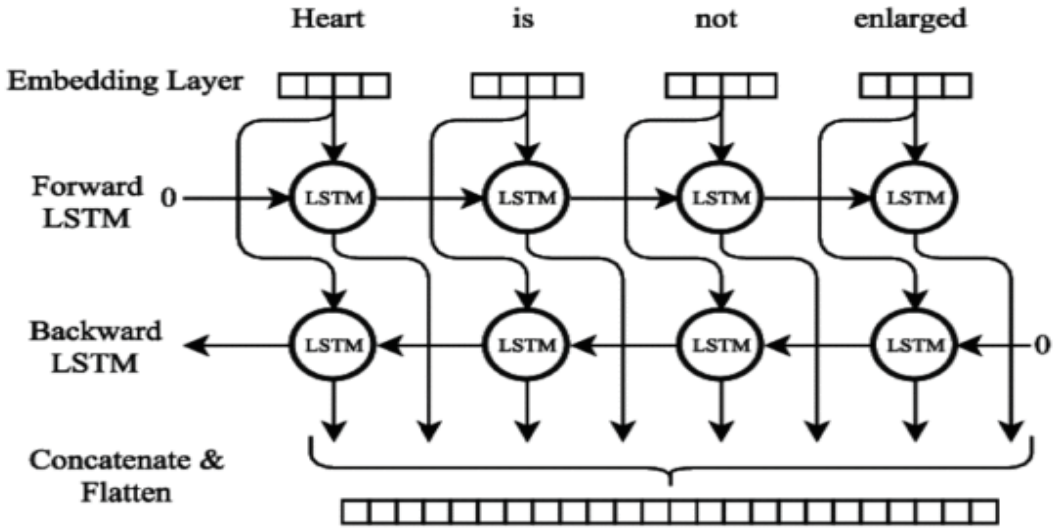

FIGURE 4.3: Basic BiLSTM architecture. Source: [121]

4.3.2.3 Implementation

Although the convolutional layer and the max-pooling layer enable CNNs to extract high-level local information effectively, they are unable to learn sequences of correlations. BiLSTM, on the other hand, improves the contexts by capturing long-distance dependencies within tweets in two directions. However, BiLSTM cannot capture local features in parallel. Therefore, an integrated structure of CNN and BiLSTM, ConvBiLSTM, is used in our experiments. We started using this architecture in 2018 for the experiments in Chapter 5 based on our findings while searching for the best deep learning model for crisis tweets (see appendix 3). We continue using it for the experiments in Chapters 6 and 7 because recent studies have shown the effectiveness of using ConvBiLSTM for Twitter data [114, 121].

It is helpful to think of ConvBiLSTM architecture (in Figure 4.4) as defining two sub-models: the CNN model for feature extraction and the BiLSTM model for interpreting the features across time steps in both directions. We define a sequential

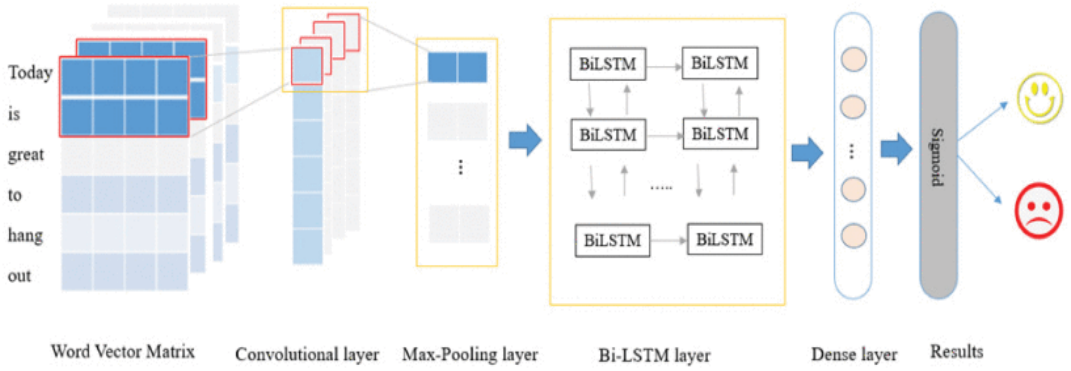

FIGURE 4.4: ConvBiLSTM architecture. Source: [121]

model and add various layers to it. The first is the embedding layer, which represents using either GloVe word embedding for English tweets or Arabic fastText embedding for Arabic tweets. The embedding layer converts tweets into numerical values and feature embedding. Feature embedding is then fed into the CNN layer with 64 filters and max pooling of size 4. The output of the CNN layer (reduced dimensions of features) is received by the BiLSTM layer with 100 neurons, followed by dropout layers with a rate of 0.5 for regulating the network. The final dense layer is the output layer with two cells representing categories along with a sigmoid activation function to produce classification results. To obtain the best parameter for our model, we utilise Adam as an optimiser and binary cross-entropy loss. We also set the class-weight parameter to "auto" – to solve the problem of imbalanced training datasets – and the maximum length to 100. In the end, our model with 25 epochs and a batch size of 32 yields better results.

4.3.3 Performance Evaluation Measures

4.3.3.1 F1 score

The F1 or F score is the weighted average of precision and recall. It is calculated using formula (4.1) or formula (4.2), and the final value lies between 0 and 1, where 1 indicates a perfect model.

$$F1\ score = 2 * \frac{1}{\frac{TP+FP}{TP} + \frac{TP+FN}{TP}} \tag{4.1}$$

$$F1\ score = 2 * \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \tag{4.2}$$

where true positive (TP) represents the correctly predicted positive values; false positive (FP) represents the wrongly predicted positive values; and false negative (FN) represents the wrongly predicted negative values; and:

$$Precision = \frac{TP}{TP + FP} \tag{4.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.4}$$

According to formula (4.2), the F1 score will be high only if both recall and precision are high. However, the F1 score has its limitations when high precision and low recall are desirable, and vice versa. In addition, Lador has stated that precision is the best evaluation method in the case of an imbalanced dataset when the negative class is in the majority, or when correctly detecting the negative class is less important than correctly detecting the positive class [75]. Otherwise, the F1 score with both recall and precision is better than using precision or recall in isolation. Furthermore, [72] has noted that the F1 score is instrumental when comparing classifiers, especially with imbalanced datasets. We use the F1 score to evaluate our work in this thesis because of the imbalanced datasets and the fact that correctly detecting a positive class (related class) is not more important than detecting a negative class (irrelevant class). Moreover, due to the stochastic nature of the learning algorithm, we repeat every experiment 30 times and take the mean as the final score.

4.3.3.2 Elbow curve

Elbow curve is a method used to determine the optimal number of clusters (k) before applying K-means to the data. This method utilises Sum Square Error (SSE) to establish a visualisation where the point position on the elbow arm indicates the optimal number of clusters [27]. Figure 4.5 visualises the elbow method when the number of clusters varies from 1 to 10. The point position on the elbow arm is located at k = 3, which represents the optimal number of clusters. [114, 121].

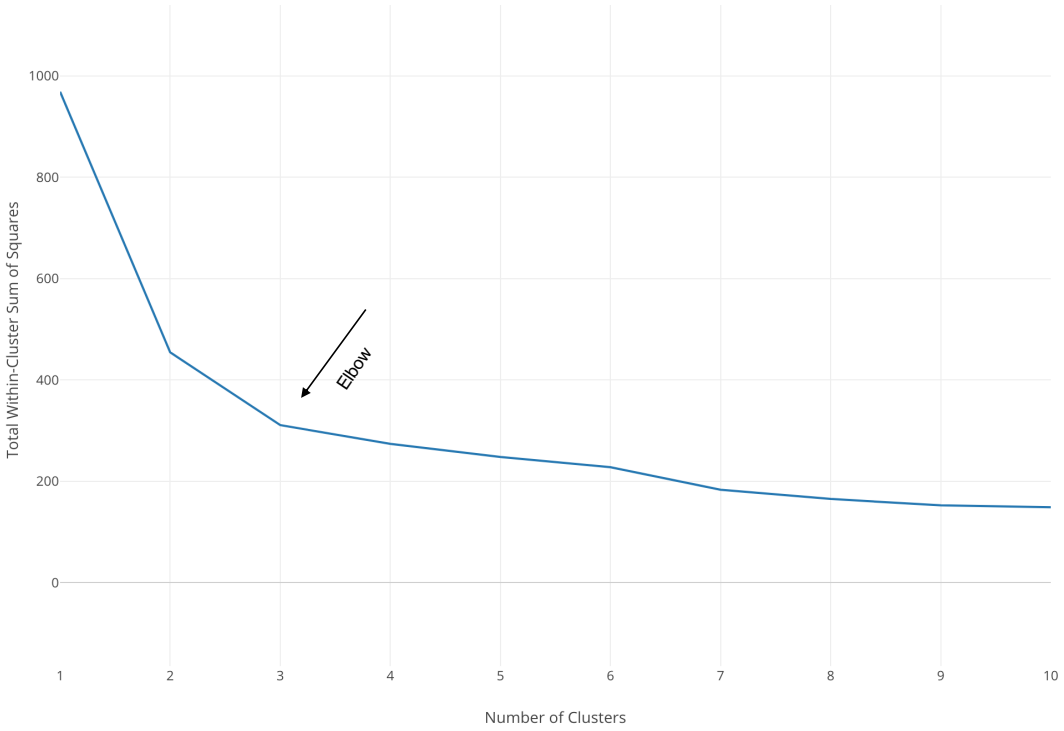

FIGURE 4.5: Elbow method visualisation. Source: [94]

.

4.3.3.3 Silhouette analysis

The silhouette plot is a measure that displays how close each point in one cluster is to points in the neighbouring clusters and thus provides a way to visually assess parameters such as the number of clusters [67]. In other words, this measure, known as the silhouette score, dictates the quality and strength of the cluster. The score ranges from -1 to 1. The value 1 indicates that the clusters are clearly distinguished from each other, value 0 indicates overlapping between the clusters, and a negative value indicates that the data points could be assigned to the wrong cluster. Figure 4.6 illustrates the silhouette diagram for a given K-mean clustering ($k = 4$); the dashed line represents the silhouette score. From Figure 4.6, we can say that all the clusters have a silhouette score above 0.6, which suggests that the setting when k = 4 is a good choice of number of clusters. Moreover, the cluster width can help in deciding the optimal number of clusters if the dataset is balanced.

Statistical significance test

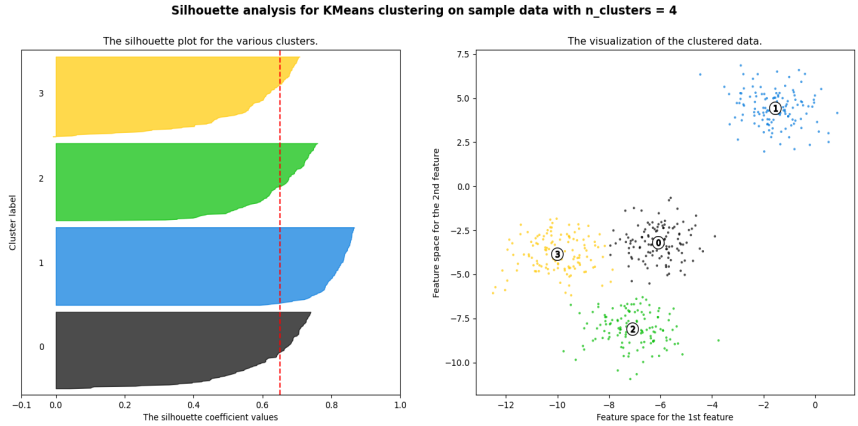

FIGURE 4.6: Silhouette diagram. Source: scikit-learn.org

.

The two-way ANOVA is a statistical test for determining the impact of two nominal predictor variables on a continuous outcome variable [49]. The effect of the independent factors on the predicted outcome and their connection to this outcome can be investigated using a two-way ANOVA. In our thesis, we use this test to determine whether the F1 score results of the model types, the number of tweets and a combination of both are statistically significant. Two hypotheses are presented:

- The null hypothesis (H0) states that no statistically significant relationship exists between the given factors.
- The alternative hypothesis (HA) states that a relationship exists between the factors.

The alpha value is set to 0.05 in this test. The p-value indicates whether we reject the null hypothesis; if $p < 0.05$, we reject the null hypothesis.

Chapter 5

Automatic Labelling Using Distant Supervision

Current tweet classification models aimed at enhancing crisis response are based on supervised deep learning. They rely on the quality and quantity of human-labelled training data. However, the available training data is small in size and imbalanced in coverage of crisis types [34], which prevents the models from generalisation [81], as tweets related to various crisis types have different features and social media responses [97]. In addition, it is infeasible to manually annotate tweets for every crisis event, especially in real time; any manual labelling is also expensive to produce. As a result, semi-supervised approaches that automatically generate new labelled training data from an unlabelled corpus are desirable.

Distant supervision can be applied to automatically generate large-scale labelled data for TCFCR. Our work in this chapter is inspired by [138] and [36] (see Section 2.2). These authors have successfully used distant supervision to generate large-scale training data for event extraction whereby triggers and arguments express the event type. This task is similar to ours: event detection for crisis response. Here, we assume that keywords can express the crisis type; thus, keywords can determine the relatedness of a given tweet to a given crisis event. However, some challenges persist when applying distant supervision to crisis data. Unlike the event extraction task, the initial keyword list does not exist for our task. Additionally, our data (tweets) do not constitute well-formed text. Tweets are full of noise and suffer from the absence of relations between words.

In Chapter 5, we introduce a novel framework to improve the ability of the classification models to generalise to unseen crisis events. It creates an initial keyword list for each crisis type using the available human-labelled data related to the given type. It then employs distant supervision [89] via the external linguistic knowledge base (FrameNet) [25]. Our framework explores different ranges of unseen features by expanding the original keyword list to include new linguistic units (i.e., new keywords with similar meaning) derived from FrameNet. If one of the crisis-type keywords exists as a lexical unit of a frame in the database, then distant supervision assumes that all the lexical units related to the given frame express the given crisis type. Unlike self-training in [134], our framework does not replicate the label noise that exists in the current dataset. In addition, crisis data that cannot be detected using existing keyword alert systems, as in [111], will be detected by our framework because of the new crisis keywords derived from FrameNet.

This chapter presents our attempt to minimise the problem of the low generalisation level of the crisis-related classifier caused by the lack of annotated tweets. To reduce the generalisation error, we test a new framework to label new crisis event data, which will then be added to the available data to train the classifier to filter the massive volume of tweets posted by users during crises. Our main goal is to investigate, for the first time, the application of distant supervision in producing good-quality labelled training data for crisis response.

Experimental results on different crisis events from five crisis types show that our work can produce good-quality labelled data from past and recent events. Specifically, substituting automatically labelled training data for part of the manually labelled training data has minimal impact on model performance, indicating that automatically labelled data can be used when no hand-labelled data are available. To evaluate our work, we create a new collection of crisis-related labelled tweets from the following new disaster events: 2018 Texas Floods, 2018 Indonesia Earthquake, 2018 Sunda Strait Tsunami, 2018 California Fires and 2018 Amritsar Crash.

5.1 Method

The method in this chapter is shown in Figure 5.1. It contains the following components:

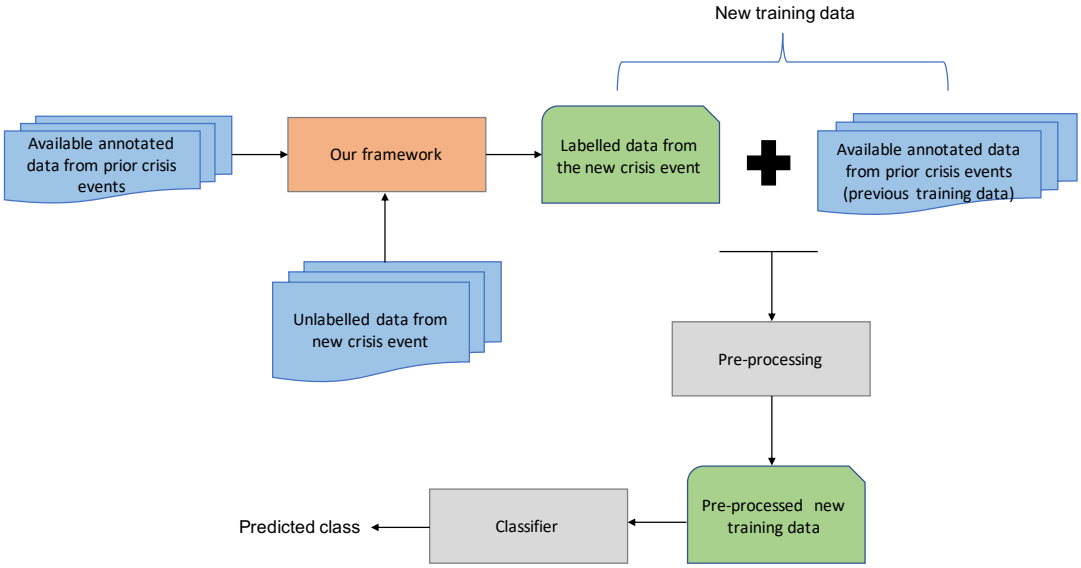

FIGURE 5.1: The procedures followed in Chapter 5, including our proposed labelling method (in orange).

Distant supervision-based labelling framework (ours): The available human-annotated data from previous events are used by our framework to automatically label the corpus from a new event from the same crisis type as the labelled data. More details will be given in the following subsection.

Pre-processing tweets: We apply the pre-processing techniques mentioned in Chapter 4 to the updated training data, including the human-labelled data from prior events along with the automatically labelled data from the new disaster event.

Tweet classifier: To evaluate the effectiveness of adding automatically labelled training data, we classify the tweets of unseen test events using the model mentioned in Chapter 4.

5.1.1 Distant Supervision-based Framework

We first create the initial keyword list based on the crisis type using the available labelled data. We then expand this list to include similar words from the external knowledge base FrameNet. We then use the expanded list to label the crisis data

collections. The proposed labelling framework (shown in orange in Figure 5.1) is described in detail by the steps shown in Figure 5.2.

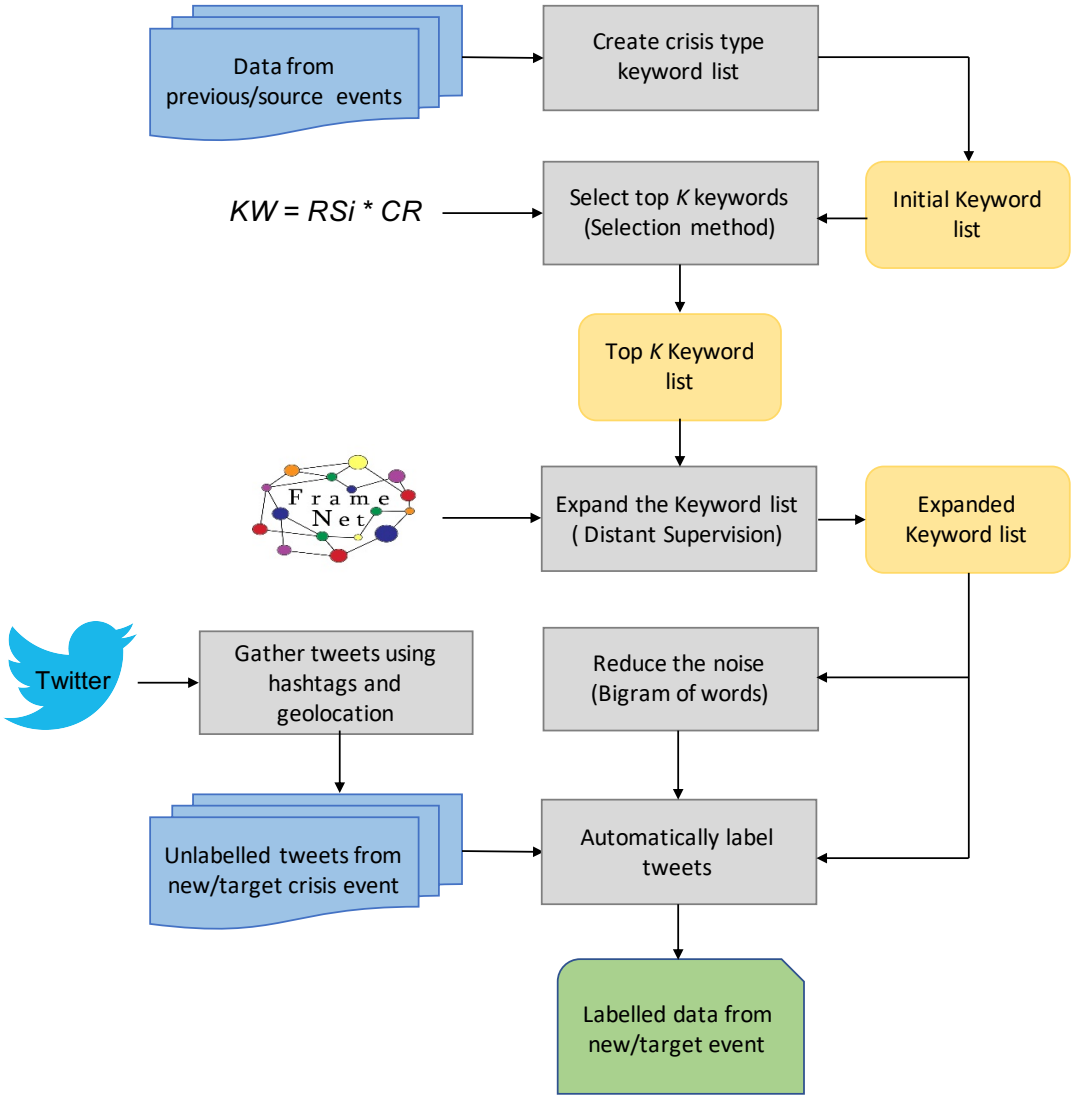

FIGURE 5.2: Proposed distant supervision-based framework.

Step one: Creating the initial keyword list. In this step, the list is created based on the available annotated tweets from different collections related to the same crisis type. For example, the entire available manually labelled data from the related class for all earthquake events are used to establish the initial keyword list for the crisis type Earthquake. The initial Earthquake keyword list in this step includes an unlimited number of words with no restrictions. To avoid word redundancy and reduce the amount of linguistically similar words, we use the Snowball Stemmer tool from NLTK 3.4 to stem each word to its root. To conduct fair experiments, at this point,

we eliminate the test event and the prior event where we automatically labelled the data in the first group experiment, and the test event only in the second and third experiment groups. Further details are given in step three.

Step two: Selecting top K keywords. After extracting the initial list of crisis-type keywords, the top K keywords are then chosen based on an intrinsic filtering method, in which a statistical measurement is used to pick the keywords with the highest scores. We calculate the keyword (KW) value for each keyword in the initial keyword list. In a tweet, a word that describes a given crisis type can be a verb, a noun or an adverb. For instance, "magnitude" (noun), "shake" (verb) and "deadly" (adverb) can be considered keywords of the crisis type Earthquake. Intuitively, a word describing a crisis type appears more than other words in the related tweets. In addition, if the same word appears in both related and unrelated tweets, it has a low probability to be a keyword of this crisis type. Thus, KW is calculated as follows:

$$RS_(i) = \frac{Count(W_(i), CT)}{Count(CT)} \tag{5.1}$$

$$CR_(i) = \log \frac{3}{(Count(CTC_(i)))} \tag{5.2}$$

$$KW_(i) = RS_(i) * CR_(i), \tag{5.3}$$

where $RS_(i)$ (role saliency) represents the saliency of i-th keyword to identify a specific word of a given crisis type, $Count(W_(i), CT)$ is the number of a word $W_(i)$ that occurs in all the tweets related to the crisis type CT,and *Count(CT)* is the count of times that all words occur in all the tweets related to the crisis type. The KW equation is inspired by the work in [36] (see Section 2.2), who used a similar key rate (KR) value to detect key triggers and arguments in event extraction tasks. However, unlike [36], $CR_(i)$ (crisis relevance) in our work represents the ability of the i-th keyword to distinguish between the tweets related to the crisis type and irrelevant tweets, and $Count(CTC_(i))$ equals 1 if the i-th keyword occurs only in the related tweets and 2 if the i-th keyword occurs in both related and irrelevant tweets.

Finally, and after removing stop words such as "and", hashtags such as "#earthquake", places such as "Nepal" and useless Twitter-specific words such as "RT" and "via", we compute for $KW_(i)$ all the words in the initial keyword list from step one

and sort them according to their KW values. This allows us to select the top K keywords for a given crisis type. For example, for the crisis type Earthquake, the top K keyword list contains "earthquake", "hit" and "magnitude", which have the highest KW values compared to the other words in the initial Earthquake list. The KW value for a given word increases when the RS or CR value of the same word increases; RS rises only if the frequency of the word in the related tweets increases. In contrast, CR increases in one case where the word occurs only in the related tweets. Table 5.1 shows how KW values play an important role in indicating the strongest keywords of the Earthquake crisis type; clearly, crisis-related and earthquake-related words have higher KW values than the unrelated ones.

TABLE 5.1: KW values of selected words from the initial Earthquake keyword list.

Keywords	**KW values**	**Ranking**
Help	0.00495	5
Quake	0.00702	3
Hit	0.00449	6
Kill	0.00216	25
Aftershock	0.00199	28
Give	0.00114	77
New	0.00129	62

Raw word frequency can be seen as a poor measurement for calculating the importance of word for a specific category due to the skews whereby certain words (including stop words like "the" or "of") can be very frequent but not informative. However, we already eliminate this disadvantage by removing all such words and stemming all words to their roots. Other standard methods such as PMI or TF-IDF have not been used here for solid reasons. Calculating PMI for positive examples and PMI for negative examples to give the final PMI score is not a fair metric in our case because of the imbalanced data problem, given the limited available manually labelled data where the number of positive examples is higher than the number of negative examples in all events (as shown in Table 4.1). On the other hand, the imbalanced dataset problem does not affect our formula as *Count(CT)* accounts for the total number of words in the related tweets only, while the total number of words in the unrelated tweets is ignored.

TF-IDF is another selection method that aims to measure the importance of a feature (word) to a given document (event type) in a given corpus (collection of tweets). However, this selection method is not suitable in our case because IDF has more impact on the final result than TF; in our case, they should be equally important since tweets are short and full of noise. If we used TF-IDF on our data, rare words such as misspelled words would have higher TF-IDF than essential keywords. Additionally, an important keyword may appear in both related and not related tweets. For instance, in earthquake crisis-type data, “earthquake” may appear very frequently in related earthquake event tweets but only once or twice in unrelated earthquake event tweets. On the other hand, our method does not discard the impact of word frequency if the word appears in both related and unrelated tweets.

Step three: Applying distant supervision. The list containing K keywords is then expanded to include similar linguistic units from FrameNet. FrameNet is an external linguistic knowledge base for English that consists of more than 1,000 semantic frames that have more than 100,000 LUs, lemmas and POS tags, which in our work are used as crisis keywords. Each frame in FrameNet is associated with a group of LUs that evoke that frame. Here, we map each keyword in the keyword list to linguistic units in FrameNet associated with the related frames only. We retrieve all the LUs of a given frame if the crisis keyword is one of these LUs and the frame is related to the crisis type. For example, “aid.v” is a linguistic unit related to the frame “Assistance” in FrameNet, which is inherited from “Intentionally_act” and can be mapped to "help" – a crisis keyword gathered in the first step and selected in the second step as one of the top K keywords based on its high KW value. In other words, if one of the top crisis-type keywords exists as a lexical unit of a frame in the database, then distant supervision assumes that all the lexical units related to the given frame express that crisis type. As a result, the number of keywords greatly increases in the final list. For instance, the number of keywords rises from 10 to 443 in the keyword list for the Fire crisis type. This list contains two types of keywords: strong keywords (top K keywords) and weak keywords extracted from FrameNet. If a word exists in the top K keywords and is an LU associated with another top K keyword at the same time, then we consider it a strong keyword. Weak keywords may bring noise to the data, which we try to reduce in step five. Figure 5.3 shows

an example of a frame and its associated LUs, as well as how we map them to keywords.

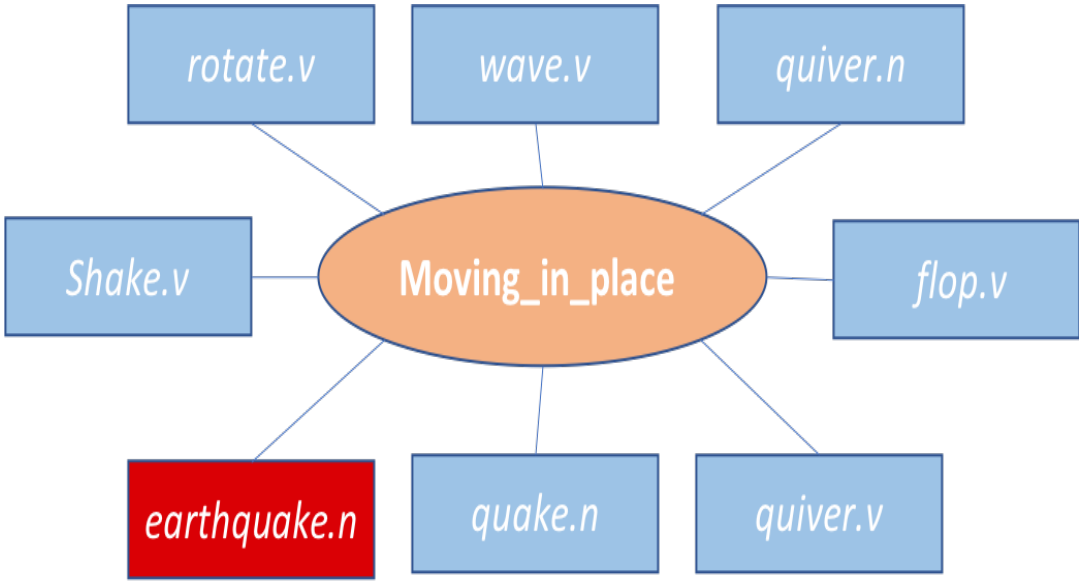

FIGURE 5.3: FrameNet example of Moving_in_place frame (in orange), its associated lexical units (in blue) and the keyword from the top K keywords for the Earthquake crisis type where the LUs are mapped (earthquake.n, in red).

It is worth noting that we choose the related frames from FrameNet manually, as it is hard to decide this without human involvement. Certain cases arise where we cannot map a top K keyword to LUs in FrameNet. Some of the top K keywords may not exist as LUs in FrameNet, such as "magnitude" in the top K keyword list for the Earthquake crisis type. Another case is when the keyword exists in FrameNet but none of the frames are related to the crisis type. For example, "toll" (i.e., number of deaths) in the top K keyword list for the Earthquake crisis type evokes three frames – Sounds, Make_noise and Cause_to_make_noise – none of which are related to crises. In these cases, the keyword is not mapped to any LUs, and the number of expanded keywords remains the same. Table 5.2 lists some of the keywords from different crisis types and their mapped LUs from FrameNet when the keyword and the related frame(s) exist.

We choose FrameNet to be used in our framework because it is more suitable for our work than other external knowledge bases such as FreeBase. First, FrameNet features LUs that evoke the frame element that can be used as keywords that evoke

TABLE 5.2: Keywords for different crisis types (keyword list) and their mapped LUs from FrameNet.

Keyword	**Crisis Type**	**Related Frame**	**Associated Lexical Units (LUs)**
Victim	Typhoon	Catastrophe	accident.n, apocalypse.n, befall.v, betide.v, calamitous.a, calamity.n, casualty.n, cataclysm.n, catastrophe.n, catastrophic.a, crisis.n, debacle.n, disaster.n, disastrous.a, incident.n, mischance.n, misfortune.n, mishap.n, suffer.v, tragedy.n,victim.n.
Burn	Fire	Fire_burning	ablaze.a, alight.a, backfire.n, blaze.n, blaze.v, bonfire.n, brush fire.n,burn.v, campfire.n, conflagration.n, fire.n, flames.n, hot spot.n, inferno.n, on fire.a, rage.v, spot fire.n.
Train	Crash	Vehicle	aircraft.n, airplane.n, ambulance.n, automobile.n, bicycle.n, bike.n, bird.n, boat.n, buggy.n, bus.n, cab.n, canoe.n, car.n, carriage.n, cart.n, chopper.n, coach.n, convertible.n, ferry.n, helicopter.n, helo.n, kayak.n, limousine.n, liner.n, lorry.n, minivan.n, pick-up.n, plane.n, schooner.n, scooter.n, sedan.n, ship.n, submarine.n, tank car.n, tank.n, taxi.n, tobogganing.n,train.n, tram.n, tricycle.n, truck.n, van.n, vehicle.n, vessel.n, warplane.n, yacht.n
Affect	Earthquake	Objective_influence	affect.v, effect.n, impact.n, impact.v, influence.n, influence.v, power.n
Dead	Flood	Dead_or_alive	alive.a, dead.a,dead.n, deceased.a, dirt nap.n, late.a, life.n, lifeless.a, live.v, living.a, living.n, nonliving.a, undead.a, undead.n

a crisis event. Second, we can retrieve all the LUs of a given frame if the initial keyword is one of these LUs and the frame is related to the crisis type. Third, unlike FreeBase, FrameNet can be used with the lack of relations in the sentence, as is the case in most of Twitter data.

In our work, K is set to 10 (K = 10). After step three, we have , originally, a top 10 keyword list for each crisis type. We choose 10 because using the top 15 keywords leads to general expressions, which breaks the borders between crisis types. On the other hand, using the top 5 keywords limits the impact of our framework. This result is based on earlier investigations, where we conduct experiments with K = 15 and K = 5.

Since we eliminate the test data from the annotated tweets used to create the initial keyword list (step one), and since we have 16 test events, 16 final keyword lists are generated according to the crisis type of the test event. For instance, when we test our model on Nepal Earthquake data, this data must be excluded from the annotated Earthquake tweets utilised to generate the automatically labelled data (by our framework) used to train the model. For this reason, we should have a specific keyword list based on the crisis type of the test data and the test data itself. This process is repeated for another experiment (first group experiment), where we create another 23 final keyword lists because we need to eliminate data from another crisis event besides the test event data: Pakistan Earthquake (MPE), Pakistan Floods (MPF), Hagupit Typhoon (MHT) and Malaysia Airline Crash (MMC). For example, if the added training data are from the Pakistan Earthquake event, then MPE and the test event data should be eliminated in step one, and a specific final keyword list is created. As a result, a total of 39 different final keyword lists are generated. More examples of initial and expanded keyword lists for specific test events from English and Arabic tweets are shown in Appendix A.

Step four: Gathering unlabelled tweets from a new crisis event. Unlabelled tweets from a new crisis event are obtained from Twitter using the Twitter API. Hashtags are used as the initial indicator of the crisis-related tweets along with geolocation information on the crisis site. For example, we use the hashtags “#californiafires”,

"#prayforcalifornia", or any other widespread hashtags related to the 2018 California Fires event and the geolocation of California state. Unlabelled tweets from multiple hashtags can also be merged. Although hashtags can be a beneficial method to classify related and unrelated tweets, a considerable number of unrelated tweets remains in which the same hashtag is used while tweeting about irrelevant subjects, such as political tweets or advertising for a particular product or service. Moreover, this step can be seen as hashtag-based supervision where tweets may contain some of these topical hashtags. However, hashtags are ignored when creating the top K keyword lists for all the crisis types, thereby eliminating any possible active role of the topical hashtags in the list. Also, in the pre-processing section, we replace all the words starting with the symbol "#" with the word "hashtag", which reduces the possibility of any negative impact caused by the appearance of these hashtags in the training data.

Step five: Noise reduction. We filter the unlabelled corpus gathered from step four after deleting duplicated and non-English tweets by applying a specific lexical feature (bigram of keywords). After cleaning the unlabelled tweets, only the examples with two keywords from the final keyword list remain. This step reduces the effect of a powerful hashtag when the hashtag without the # symbol is one of the keywords. For example, if we use "#earthquake" as one of our hashtags in the previous step, and "earthquake" is one of the keywords in the final keyword list, then tweets like "@archpics: grammichele is located in sicily, in southern italy. the town was constructed in 1693 with a distinctive hexagonal street #earthquake" will not be selected for the 2018 Indonesia Earthquake event. On the other hand, the tweet "quake hits indonesia, and deaths are reported at least three people died when a 6.0-magnitude #earthquake" will be selected because of the appearance of at least two keywords from the final Earthquake keyword list: "quake", "earthquake" and "magnitude" in this case. This process also eliminates several tweets that contain only one weak keyword expanded from FrameNet, which decreases most of the noise caused by step three. For instance, the tweet "@antonioguterres: Saturday is the International Day for Disaster Reduction. Here in Indonesia, I have just seen the devastating impact" will not be chosen for the 2018 Indonesia Earthquake event since "disaster" is a weak keyword derived from FrameNet using "victim" as one of the top K keywords

for the Earthquake crisis type and an LU in FrameNet at the same time (as shown in Table 5.2).

Step six: Labelling the corpus as related and not related examples. A collection of data from the new crisis event is automatically generated by labelling tweets from step five as relevant (positive) examples and tweets with no keywords from the expanded keyword list as not related (negative) examples. Examples of automatically labelled data created by our framework are shown in Table 5.3.

TABLE 5.3: Examples from English automatically labelled data created by our framework.

Crisis Event	**Tweet**	**Automated Label**	**Reasons**
2018 Texas Floods	"@srich1953: #laketravis above 700 feet.more rain for #astin today and Friday #mansfield dam will have 8 floodgates open #texas #txfloods"	Related	Two keywords exist in the given tweets: one strong keyword ("floods") and a weak keyword ("rain"). "Rain" is associated with the strong keyword "floods".
2018 Amritsar Crash	"@hatindersinghr: this is the new video where organiser can be seen sayin to wrap the programme to the earliest as many people have died"	Related	Two keywords exist in the given tweet: a strong keyword ("people") and a weak keyword ("die"). "Die" is associated with the strong keyword "dead".

Continuation of Table 5.3

Crisis Event	**Tweet**	**Automated Label**	**Reasons**
2018 California Fires	"@weatherchannel: say it ain't #snow? Late week storm could bring up to 8 inches of snow to parts of the mid-atlantic and north-east."	Not related	Absence of keywords from the final Fire crisis-type list.
2018 Indonesia Earthquake	"I feel something weird in my stomach. like the way animals sense before an earthquake."	Discarded	Only one strong keyword occurs in the tweet ("earthquake").
2018 Texas Floods	"@7rixi: automatic tweets or bots flooding twitter for @ted-cruz get #teambeto trending #texas#betofortexas #texastownhall https:xe2/x80/xa"	Discarded	Only one strong keyword occurs in the tweet ("flood").
2018 Indonesia Earthquake	"@onslowhouse: fracking? i feel so sorry for everyone involved in this campaign to stop fracking ine. why must we wait"	Not related	Absence of keywords from the final Earthquake crisis-type list.
2018 California Fires	"@vcapethealth: as the fires still burn, please continue to pass the word on that vca is offering free boarding for displaced animals"	Related	Two keywords exist in the given tweet: one strong keyword ("fire") and one weak keyword "burn". "Burn" is associated with the strong keyword "fire".
2018 Amritsar Crash	"@rubeenajan1: #pictureoftheday another side of the #kashmir valley. #kashmiri children play with #indianarmy soldier"	Not related	Absence of keywords from the final Crash crisis- type list.

Continuation of Table 5.3

Crisis Event	**Tweet**	**Automated Label**	**Reasons**
2018 Sunda Strait Tsunami (Typhoon)	"@livestormchaser: breaking weather updated: at least 222 have been killed, 843+ wounded and many still missing after #tsunami strikes"	Related	Two keywords exist in the given tweet: one strong keyword ("tsunami") and one weak keyword ("weather"). "Weather" is associated with the strong keyword "storm".
2018 Texas Floods	" @libertyallday: #texans are smarter than this! no way they allow #ca elites to make decisions for #texas #votefromcruz"	Not related	Absence of keywords from the final Floods crisis- type list.
2018 Amritsar Crash	"No-one is Right Unless someone is wrong #AmritsarTrainAccident"	Discarded	Only one strong keyword occurs in the tweet ("train").
2018 Indonesia Earthquake	"@unocha: some footage taken today from indonesia showing the aftermath of the devastating earthquake hers how you can help https://xe2/x80/xa"	Related	Two keywords exist in the given tweet. Both are strong keywords ("earthquake" and "help").

5.2 Experiments

The crisis response literature presents two approaches to running experiments on crisis-related data. The first is where the training data and the testing data are related to events from different crisis types, while the second is where the training and the testing data are from the same crisis type. According to [47], tweets related to

events from the same crisis type share common features, which leads to better results. Thus, we follow the second approach in our experiments. However, and due to the very limited available manually labelled data for each crisis type, we combine tweets from similar crisis types. Data for train crashes, helicopter crashes and airplane crashes are combined as the Crash crisis-type data. Data for tornado, typhoon, hurricane and tsunami events are combined as the Typhoon crisis-type data. Wildfire and bushfire data are combined as the Fire crisis-type data.

More specifically, this chapter aims to answer the following research questions:

- Can we automatically generate labelled training data for tweet classification for crisis response that has a competitive quality level compared to manually labelled training data?
- When added to the available labelled data, does our automatically generated labelled data improve the performance of the crisis-related tweet classifier?

To answer these research questions, we run three groups of out-of-domain experiments (i.e., where all the data from the test event are not used as training data). This means that the model does not see any tweets from the test event during the training phase. The three groups of experiments are as follows:

1. Quality of produced data: The first group of experiments aims to examine the effectiveness of our framework by monitoring the quality of the labelled data generated by our framework compared to the hand-labelled data.
2. Adding data from new crisis events: The second group of experiments aims to determine the effect of incorporating new labelled data generated by our framework on the performance of a tweet classification model for crisis response.
3. The impact of using FrameNet. The final group of experiments aims to study the impact of step three in our framework (using an external linguistic knowledge base) on the quality of the generated labelled data and, in turn, on the performance of the tweet classification model.

5.2.1 Quality of Produced Data

This experiment aims to answer the first research question in this thesis:
Can we automatically generate labelled training data for tweet classification for crisis response that has a competitive quality level compared to manually labelled training data?

Here, we investigate the quality of the labelled data created by our framework compared to the manually labelled data from the same event. To do so, we conduct two sub-experiments for each crisis type: (1) using manually labelled data (see Table 4.1) and (2) using automatically labelled data from a given disaster event. For example, in the Earthquake crisis type, we train the model with all the manually labelled data, including Pakistan Earthquake (MPE). In the second sub-experiment, MPE is replaced with the automatically labelled data related to Pakistan Earthquake (APE) to train the classifier.

5.2.2 Adding data from new crisis events

This group of experiments aims to show the effectiveness of incorporating automatically labelled data from new crisis events on the performance of the tweet classification model for crisis response. The goal is to answer the second research question in this chapter:
When added to the available labelled data, does our automatically generated labelled data improve the performance of the crisis-related tweet classifier?

In these experiments, we compare two labelling methods that give labels to the unlabelled tweets from the new crisis events: self-labelling (SelfL), where similar manually labelled collections from the same crisis type are used to pre-train a model to be utilised to classify the new unlabelled data (as in [135]), and the distant supervision-based framework (DS), where our framework is used instead of the pre-trained model. We use UIE, UTF, UST, UAC and UCFi data (see Section 4.1.2) to generate the new added labelled data for Earthquake, Floods, Typhoon, Crash and Fire crisis types, respectively. As a result of DS labeling method, we use AIE, ATF, AST, AAC and ACFi data as the added training data.

Regarding the training data, we directly mix the new labelled data with the available human-labelled data to train the tweet classifier, as shown in Figure 5.1 in the

Method section. We also report the classifier performance when trained using the original manually labelled data without incorporating the new labelled data (E for Earthquake, F for Floods, T for Typhoon, C for Crash and Fi for Fires) and use it as our baseline.

5.2.3 Impact of Using External Knowledge Base (FrameNet)

In this group of experiments, we exclude step three from our framework (DS-FN): applying distant supervision via FrameNet. The main goal is to investigate the impact of using the external knowledge base on model performance. This can be seen as a simple keyword-based framework, where only the top K keywords are used to automatically label the unlabelled data from new or prior crisis events. We use UIE, UTF, UST, UAC and UCFi (see Section 4.1.2) to generate the new added labelled data for Earthquake, Floods, Typhoon, Crash and Fire crisis types, respectively.

5.3 Results and Discussion

5.3.1 Quality of Produced Data

Table 5.4 shows the F1 scores for the first experiment group across four crisis types. E represents all the available manually labelled Earthquake crisis datasets excluding MPE; F represents all the available manually labelled Floods crisis datasets excluding MPF; T represents all the available manually labelled Typhoon crisis datasets excluding MHT; and C represents all the available manually labelled Crash crisis datasets excluding MMC.

As shown in Table 5.4, using APE instead of MPE with training datasets from other earthquake events to classify the tweets in MChE, MBE, MCE and MNE datasets slightly diminishes the performance in F1 score by 1.2%, 2.8%, 0.5% and 1.2%, respectively. Similar results are presented for the Floods crisis-type data, where the maximum drop is 4.2% for the MAF dataset. F1 scores displayed for four Typhoon event datasets show a minor decline in model performance when using AHT rather than MHT in training the data along with the hand-labelled typhoon events data. However, this is not the case for MST. Here, the performance drops from 0.8023 to

TABLE 5.4: F1 score results for first experiment group (four crisis types).

(a) Earthquake crisis type				
Train/Test	**MChE**	**MBE**	**MCE**	**MNE**
E+MPE	0.8044	0.9335	0.8941	0.9168
E+APE	0.7882	0.9052	0.8890	0.9043
(b) Floods crisis type				
Train/Test	**MIF**	**MCF**	**MQF**	**MAF**	**MMF**
F+MPF	0.7389	0.9224	0.7855	0.9619	0.8987
F+APF	0.7100	0.9170	0.7496	0.9190	0.8881
(c) Typhoon crisis type				
Train/Test	**MOkT**	**MYT**	**MST**	**MOT**
T+MHT	0.8418	0.8781	0.8023	0.9618
T+AHT	0.8067	0.8372	0.7242	0.9126
(d) Crash crisis type				
Train/Test		**MGC**	**MLC**	**MNC**
C+ MMC		0.7774	0.9516	0.9985
C+ AMC		0.7634	0.9516	0.9985

0.7242 in F1 score. One possible reason is that the final keyword list for this case includes a higher level of noise than other cases (test events). We investigate the automatically labelled data created for the Hagupit event (AHT) using this particular final keyword list, when MST is excluded. We find that, unlike the other Typhoon events, one of the top K keywords is "go". This is because people tweet, in the data of the remaining events, about the directions of the emerging typhoons. However, the number of weak keywords driven from FrameNet through "go" is 22, including non-crisis words such as "zigzag". On the other hand, some of the important words regarding data from crisis events are not in the final list, such as "victim" and "safe". Table 5.5 presents examples of the mislabelled tweets for AHT in this case.

We also observe that the performance almost remains the same for the Crash events MGC, MLC and MNC. This result suggests that our generated data, AMC, has a good quality level and can be replaced with data labelled by volunteers, as in MMC.

To conclude, we answer the following research question:

Can we automatically generate labelled training data for tweet classification for crisis response that has a competitive quality level compared to manually labelled training data?

In general, it can be said that substituting automatically labelled data produced

TABLE 5.5: Examples of mislabelled tweets from AHT when MST is the test event.

Tweet	**Wrong Label**	**Reasons**
“@ancalerts: pagasa: #rubyph expected to be in oriental mindoro monday morning“	Not related	Absence of keywords from the final Typhoon keyword list.
“““ruby““(international name: hagupit) is expected to make landfall in tacloban city on saturday evening. #rubyph“	Not related	Absence of keywords from the final Typhoon keyword list.
"@inqnational: 943 passengers stranded in bicol ports as of friday - army #rubyph | via @smbarrameda"	Not related	Absence of keywords from the final Typhoon keyword list.
“omg buti na lang na send ko na first draft ko for thesis the other day thru google drive. #signofrelief #rubyph “	Related	Two keywords exist in the given tweet, a strong keyword “relief “ and a weak keyword “drive“. “Drive“ is associated with the strong keyword “hit“.
“does climate change has something to do with the #typhoonhagupit #typhoonruby #rubyph“	Related	Two keywords exist in the given tweet, a strong keyword “typhoon“and a weak keyword “climate“. “Climate“is associated with the strong keyword “storm“.
“@ukcovergirl: uk ink cover girl the stunning @nancy_harry #inkedgirls #ukcovergirl #ukcg #stunneroftheday @sxypb @oh_eddy #rubyph “	Related	Two keywords exist in the given tweet, two weak keywords: “ cover“and “girl“. “Girl“ is associated with the strong keyword "“people“while “cover“ is associated with “hit“.

by our framework with manually labelled data from the same crisis event in training tweet classifiers for disaster response has a minimal impact (average of 2.62%) on the classifier's performance for the three crisis types. This is due to the noise (mislabelling problem) in the produced data. These results demonstrate that data annotated by our framework can be used when no hand-labelled data are available for new disaster events because they have similar quality levels. This finding can be considered a good outcome, since manually labelling new data from multiple events requires large amounts of time, money and effort compared to the automatically generated data.

5.3.2 Effect of Adding Data from New Crises

As seen in Table 5.6, DS reports the best labelling method for two earthquake crisis datasets (MChE and MPE), with a maximum improvement of 2.1% in F1 score. On the other hand, the performance does not improve for the MBE, MCE and MNE datasets. In Table 5.7, for flood crisis datasets, DS is the best labelling method when tested on the MAF, MIF, MCF and MQF datasets, whereas SelfL is superior in the remaining two datasets. However, the improvements in F1 score are very minor. For the typhoon crisis datasets, in Table 5.8, the classifier performance improves when using DS as the labelling method for three out of six datasets (MYT, MPT and MHT). Regarding crash events, DS is better than SelfL for MNC but not MGC and MLC as can be seen in Table 5.9. For MCFi and MAFi in the fires events, in Table 5.10, DS reports the best labelling method.

After analysing the data, we observe that more than seven keywords from the top K keyword list used to label the extra data appear in the MPE, MIF and MYT datasets, which helps in providing new keywords from FrameNet to the training data. Conversely, only one keyword occurs in the MChE and MNE datasets, with more than 30 new keywords derived from the external knowledge base. These new keywords assist in recognising related tweets from the new data that would not have been identified by the old keyword list. Given that we only label tweets with two keywords, different (new) relations may emerge using these new keywords. On the other hand, in the case of the limited number of new keywords derived from

FrameNet, adding data from new crisis events does not improve the model performance, regardless of the number of top K keywords appearing in the test data, especially if the training and the testing data are dissimilar (MCE and MOkF). If the training and testing data are similar and the number of matched keywords is low, SelfL is the predicted best labelling method, as seen in MLC and MBFi.

The main observation here is that although we produce good-quality labelled training tweets, the improvement in model performance is not significant when adding these tweets to the original training data using any of the three labelling methods (DS, SelfL or DS-FN). However, we compare our results in this group of experiments to those in the previous group (see Table 5.5). We notice that, unlike our finding, model performance improves for some events when data generated by our framework are added to the original training data. F1 score for MMF increases from 0.8532 (F in Table 5.7) to 0.881 (F+APF in section (b) of Table 5.5) when APF data are added to the original Flood data (F). The same situation appears for MAF. The F1 score also rises for MBE by approximately 2% for the Earthquake crisis type when APE data are added to E. Another example is from the Typhoon crisis type, where performance improves from 0.7080 (T in Table 5.8) to 0.7242 (T+AHT in section (c) in Table 5.5). In contrast, the performance falls for some events, such as MChE, MQF, MYT and MGC. It declines when adding prior events to the original training data regardless of the labelling method. For instance, the F1 score decreases for MOT (0.9610, T in Table 5.8) for both cases when adding MHT (0.9180; T+MHT in section (c) in Table 5.5) and adding AHT (0.9126; T+AHT in section (c) in Table 5.5). This indicates that adding any data from Hagupit Typhoon, even if they are manually labelled by paid workers, negatively affects the model performance when tested on MOT. Thus, our second main observation is that adding more data does not always lead to better results, and this differs from one crisis event to another based on the similarity level between the added data and the test data. In other words, the selection of the added event plays an important role in developing the model performance on specific test data.

In sum, we answer the following research question:

When added to the available labelled data, does our automatically generated labelled data improve the performance of the crisis-related tweet classifier?

From the discussion above, we can say that DS improves the performance of TCFCR models in 12 out of 23 datasets from different crisis types. Generally, our results show that DS improves model performance if new keywords derived from FrameNet exist in the test data, especially if the similarity level between the test and original training data is low and the similarity level between the new added data and test data is high.

TABLE 5.6: F1 score results for second experiment group for Earthquake crisis type.

Model/Test	MPE	MNE	MCE	MBE	MChE
E	0.7915	**0.9068**	**0.8921**	**0.8876**	0.8356
SelfL	0.7903	0.9045	0.8842	**0.8877**	0.8302
DS	**0.7940**	0.8875	0.8769	0.8863	**0.8566**
DS-FN	0.7913	0.9026	0.8780	0.8855	**0.8581**

Note. E represents all available manually labelled earthquake crisis datasets excluding MPE. Best results for each test data are in bold.

TABLE 5.7: F1 score results for second experiment group for Floods crisis type.

Model/Test	MPF	MMF	MQF	MCF	MIF	MAF
F	0.962	0.8532	0.839	0.917	0.764	0.916
SelfL	**0.968**	**0.8544**	0.836	0.921	0.762	0.917
DS	0.960	0.8487	**0.840**	**0.922**	**0.767**	**0.925**
DS-FN	0.966	0.8437	**0.840**	0.920	0.761	0.920

Note. F represents all available manually labelled flood crisis datasets excluding the test data. Best results for each test data are in bold.

TABLE 5.8: F1 score results for second experiment group for Typhoon crisis type.

Model/Test	MHT	MPT	MYT	MOT	MOkT	MST
T	0.881	0.827	0.901	**0.961**	**0.793**	**0.708**
SelfL	0.881	0.825	0.897	**0.962**	0.787	0.703
DS	**0.882**	**0.829**	**0.911**	0.957	0.759	0.699
DS-FN	**0.883**	**0.828**	**0.910**	0.960	0.777	0.702

Note. T represents all available manually labelled typhoon crisis datasets excluding the test data. Best results for each test data are in bold.

TABLE 5.9: F1 score results for second experiment group for Crash crisis type.

Model/Test	MGC	MLC	MNC
C	0.7725	0.9419	**0.9977**
SelfL	**0.7758**	**0.9430**	0.9974
DS	0.7716	0.9421	**0.9976**
DS-FN	0.7677	0.9352	0.9971

Note. C represents all available manually labelled crash crisis datasets excluding excluding the test data and MMC. Best results for each test data are in bold.

TABLE 5.10: F1 score results for second experiment group for Fire crisis type.

Model/Test	MCFi	MBFi	MAFi
Fi	0.7858	0.9603	**0.8011**
SelfL	0.7956	**0.9621**	0.7725
DS	**0.8020**	0.9610	0.7889
DS-FN	0.7994	0.9600	0.7406

Note. Fi represents all available manually labelled fire crisis datasets excluding the test data. Best results for each test data are in bold.

5.3.3 Impact of using external Knowledge base (FrameNet)

The results in Tables 5.6, 5.7, 5.8, 5.9 and 5.10 show that incorporating FrameNet is an important stage in our framework, as the performance of the tweet classification model increases for 14 out of 23 datasets. Our framework is able to detect tweets that include unseen keywords such as "burn", "aid" and "bushfire" in MAFi, where the performance decreases from 0.7889 (DS) to 0.7406 (DS-FN) in F1 score when we remove the applying distant supervision step from the framework as presented in Table 5.10.

5.4 Conclusion

In this chapter, we examined the application of distant supervision in generating automatically labelled tweets from new crisis events to overcome the problematic lack of training data in the crisis response literature [34, 135, 47, 131, 92]. Adding more crisis data from different disaster events should improve the performance of

classifiers when identifying tweets from unseen events [81]. This leads to a more reliable system to be used by humanitarian organisations to help people in need during crises. The results show the effectiveness of our distant supervision-based framework in producing labelled training tweets from new crisis events, especially when no manually labelled data are available for the given incident. Substituting generated annotated data rather than manually labelled data in training tweet classifiers for disaster response has only a minor impact on performance. Specifically, model performance drops slightly by an average of 2.62% on 16 datasets from different locations and crisis types. This indicates that the generated data has a competitive quality compared to the manually labelled data – with less effort, time and money. Our results also suggest that our proposed framework is the best labelling method when the test data and the original training data are dissimilar. This is because it can recognise the related tweets in the test data that include new keywords retrieved from FrameNet and do not exist in the original training data. Based on our outcomes, the selection of the new disaster event plays an important role in improving the ability of the crisis-related model to classify tweets from unseen events. The data from new and test events should have a high level of similarity; otherwise, the performance drops when adding the new training data to the original ones.

Our work in this chapter has a number of limitations. First, our framework relies on manually labelled data to create the initial keyword list (step one in the Method section), which makes it inapplicable to crisis types with only one or two available labelled datasets, such as the Building collapse or Volcano crisis types. Second, the new event should be carefully chosen after analysing the data. Third, the noise reduction (step four) allows tweets with two weak keywords to be labelled as positive examples without any restrictions. According to step four in our framework, there is no difference between strong and weak keywords in the labelling process. This brings undesirable noise to the generated data. In the future, we plan to use our framework to label tweets from emerging crisis events to be adopted to the model during the training phase through a domain adaptation approach.

Chapter 6

Domain Adaptation for English Twitter Data

The information posted by users on Twitter during crises can significantly improve crisis response with regard to reducing human and financial losses [51]. Deep learning models can identify related tweets to mitigate the information overload that prevents humanitarian organisations from using Twitter posts [69]. However, these models rely heavily on labelled data, which is unavailable for emerging crises since it is infeasible to manually annotate tweets from these events in real time [93, 131, 132, 82]. In turn, because each crisis has its own features such as location, time of occurrence and social media response [97], current models are known to suffer when generalising to unseen disaster events after pre-training on past ones [34, 81].

From Chapter 5, it can be seen that pre-analysis steps are needed to ensure the classifier's ability to identify unseen tweets from a crisis event, as adding more data depends on the similarity level between the new extra training data and the test data as well as the number of keywords derived from FrameNet. However, our distant supervision-based framework has the potential to always improve the model performance if the new training data and the test data come from the same disaster event (high-level similarity). In addition, for crisis response, the gap between source and target tweets can be minimised through common similar keywords provided by employing distant supervision techniques. According to the authors in [109], unlabelled target data can be labelled using pseudo-labelling techniques and used as training data by retraining a pre-trained source model from scratch, finetuning the pre-trained model or building a new target model. However, to the best of our

knowledge, no work has studied the application of using distant supervision as a pseudo-labelling technique within a domain adaptation approach for TCFCR.

This chapter presents an attempt to minimise the domain shift between the target and the source tweets using a domain adaptation approach inspired by our previous work. Here, we use our distant supervision-based framework to label the unlabelled target data (pseudo-labelling), whereby an initial keyword list is established using the available annotated source data from past similar events. The most related keywords are then selected using a statistical method. The selected keyword list is then expanded by employing distant supervision via an external knowledge base (FrameNet), and those tweets with a bigram of keywords are labelled as positive tweets, while tweets with none of the keywords are labelled as negative tweets. The generated labelled data is then mixed with the available source data to train a new target model.

Our method is especially useful when tweets describing an emerging crisis may not include keywords derived from past events, since we provide an expanded keyword list by applying distant supervision via FrameNet. Our method also avoids the error amplification problem caused by using a basic semi-supervised approach (iterative self-training [82]), especially when the emerging event is different from the past events. We evaluate the method on eight 2012–2015 crisis events from three crisis types (Earthquake, Floods and Typhoon). Our results show that our approach can be seen as a general robust method to classify unseen tweets from current events.

6.1 Method

Semi-supervised domain adaptation techniques have been adopted to incorporate unlabelled target data to labelled source data to reduce the gaps between the two domains. Our method (described in Algorithm 1) contains two stages: the pseudo-labelling stage and the adaptation stage. In the pseudolabelling stage, unlabelled tweets from the current (target) crisis event are gathered from the Twitter API using publicly available tweet IDs. The unlabelled tweets are then given pseudo-labels by applying our distant supervision-based framework. In the adaptation stage, the pseudo-labelled target tweets are used to build a target model with several crisis

Algorithm 1 Robust domain adaptation approach with pseudo-labelled target data.

1. Given: Labelled tweets of several crisis events from different time intervals and locations from the same crisis type to the given target event (MLS); unlabelled tweets from target domain (UT) retrieved using Twitter API and publicly available tweet IDs; and manually labelled test data from target domain (MLTT).
2. Pseudo-labelling stage: Use our framework to label UT based on all the available MLS and employing distant supervision via external knowledge base (giving them pseudo-labels).
3. Adaptation stage: Build a target model using MLS with the pseudo-labelled data from the target domain.
4. Evaluate the model on MLTT.

events from different time intervals and locations from the same crisis type to the given target event. In other words, through this domain adaptation approach, we try to minimise the gap between source and target data by using the source data to create an expanding keyword list that can be used to label a limited number of real-time target tweets. The tweets from source and target events are then used to build a model that classifies the incoming unlabelled data from the emerging (target) crisis event in real-time scenarios. Further details will be provided later in this section.

6.1.1 Pseudo-labelling Stage

We apply the same distant supervision-based framework mentioned in Section 5.1.1 in the previous chapter. However, we now use the framework to give automated labels to the unlabelled data from emerging crisis events rather than new past events. Specifically, the human-labelled data described in Table 4.1 are used to create an expanded keyword list to label the data collected by tweet IDs as presented in Table 4.3 and to generate the automatically labelled data.

Our framework consists of six steps as shown in Figure 5.2 in Section 5.1.1, with some minor adjustments. In step one, we exclude data from the emerging/testing crisis event when we use the available annotated labelled data to generate the initial keyword list for a crisis type. In step four, we also gather unlabelled tweets from their publicly accessible IDs using the Twitter API instead of using hashtags and geolocation information.

Pseudo-labelled data

Table 6.1 demonstrates examples from pseudo-labelled tweets created by our framework from some of the target events in our work. It should be noted that the given target crisis event is excluded from source events used to create the crisis-type keyword list.

TABLE 6.1: Examples from English pseudo-labelled data created by our distant-supervision-based framework.

Target event	**Tweet**	**Pseudo-label**	**Reasons**
Queensland Floods	"wind still blowing like a freight train here! rain's eased. thoughts with all those struggling in floods. #qldfloods"	Related	Two keywords from the final keyword list exist in the given tweet: one strong keyword "flood" and one weak keyword "rain". "Rain" is associated with the strong keyword "flood".
California Earthquake	"why the earthquake near san francisco is just the start of the shaking in california http://t.co/4ezy0ev6pv"	Related	Two keywords from the final keyword list exist in the given tweet: one strong keyword "earthquake" and one weak keyword "shake". "Shake" is associated with the strong keyword "earthquake".

Continuation of Table 6.1

Target event	**Tweet**	**Pseudo-label**	**Reasons**
Nepal Earthquake	"@_banarasi_: all phone calls from # india to #nepal made by any#bsnl phone shall be charged at local rates and not at isd rates. #nepal"	Not related	Absence of keywords from final Earthquake crisis-type list.
Queensland Floods	"@girlposts: It's amazing how quickly your mood can change, how deep your heart can sink and how much one person can affect you."	Discarded	Only one strong keyword from the final keyword list occurs from the final keyword list in the tweet "affect".
Pakistan Floods	"@drgpradhan @sageelani lol, really this bastard has gone mad. i found this link http://t.co/gfgfupiqw4"	Not related	Absence of keywords from final Floods crisis-type list.
Pam Typhoon	"#cyclonepam: massive storm bears down on vanuatu, with 260,000 people in its path http://t.co/ex0u1tyys7"	Related	Two strong keywords from the final keyword list exist in the given tweets: "storm" and "people".
California Earthquake	"engineer interested in data, health; wearables? our data science team does data products; stories: https://t.co/91wsymje4z chat with us!"	Not related	Absence of keywords from final Earthquake crisis-type list.

Continuation of Table 6.1

Target event	**Tweet**	**Pseudo-label**	**Reasons**
Chile Earthquake	"chile lifts tsunami warning after quake kills 6 - yahoo news http://t.co/nn72g4gtmd"	Related	Two strong keywords from the final keyword list exist in the given tweet: "quake" and "kill".
Queensland Floods	"hi-ho hi-ho... off to work i go :)"	Not related	Absence of keywords from final Floods crisis-type list.
Pakistan Floods	"pakistan: 08 september 2014: asia – floods and severe weather source: european commission humanitarian aid department http://t.co/fu0buputq"	Related	Two keywords from the final keyword list exist in the given tweet: one strong keyword "floods" and one weak keyword "aid". "Aid" is associated with the strong keyword "help".
Chile Earthquake	"i miss you"	Not related	Absence of keywords from final Earthquake crisis-type list.
California Earthquake	"RT @peterhartlaub: The Greatest Generation: 96-year-old colleague @daveperlman, who served in WWII, wrote this story before I woke up"	Discarded	Only one weak keyword from the final keyword list occurs in the tweet "serve". "Serve" is associated with the strong keyword "help".

Continuation of Table 6.1

Target event	**Tweet**	**Pseudo-label**	**Reasons**
Nepal Earthquake	"@shelterbox: a 7.9 magnitude earthquake has hit 50 miles east of pokhara, #nepal. shelterbox is monitoring, ready to respond"	Related	Two strong keywords from the final keyword list exist in the given tweet: both are strong keywords "earthquake" and "magnitude".
Hagupit Typhoon	"@teamrubicon: what are the requirements to deploy internationally with our team? http://t.co/pjquxmoil7 #hagupit #rubyph"	Not related	Absence of keywords from final Typhoon crisis-type list.

6.1.2 Adaptation Stage

We add the pseudo-labelled target data created in the first stage to the available labelled source data from the same crisis type as the target crisis to build a new target model to classify the unseen tweets from the emerging event. Pseudo-labelled target data generated by our distant supervision-based framework provides new keywords than those existed in the source data. Adding these data to the manually labelled tweets brings target-related features to the training data, including location and crisis nature.

The event lifetime is the reason for mixing source and target data in training the target classifier. The information posted by people during a disaster differs based on the tweet's posting time [115]. For example, tweets containing advice, warnings and alerts start to appear at the beginning of the event onset and decrease thereafter. On the other hand, tweets containing reports on damage and affected individuals reach their peak in the middle of the disaster. Sympathy and prayers vanish after

the disaster dissipates. In turn, because we replicate the real-time scenario and order the tweets by posting time, the number of information types in the gathered target tweets is limited. This restricts the model in identifying tweets with types that occur in the middle of the event onset or which do not exist in the gathered tweets. Nonetheless, we still have the source data from complete past events in which all the information types are available. By mixing the source and target data in training the target model, we thereby increase the ability of the target classifier to identify related target tweets, including any type of information during the target event lifetime.

6.2 Experiments

Our main goal is to investigate whether automatically labelled target data generated by a framework via distant supervision can be used to build a robust model along with similar source events to improve model performance in classifying unseen tweets from emerging events.

To determine the effectiveness of using pseudo-labelled target data generated by our framework, we compare several component labelling and adaptation methods. Specifically, we use two methods to automatically give labels to the unlabelled target data (stage 1):

- Distant supervision-based framework (DS) – using our distant supervision-based framework proposed in Chapter 5 and modified in this chapter; and
- Self-labelling (SelfL; in [82]) – using a pre-trained model on MLS. Here, we use the self-training iterative method introduced by the authors [82], with a confidence interval of 80%. The authors compared this method with their novel feature-based, instance-based and hybrid feature-instance domain adaptation methods, finding that self-training performs better in building a target model for crisis data (see Section 2.3.1.2). We compare our labelling method with self-labelling because it is the best reported method in the existing literature for producing self-labelled target data in domain adaptation approaches.

In the adaptation stage (stage 2), we use three methods to incorporate target labelled data in the previous stage:

- Target Model (TM) – building a model following the source architecture using human-labelled tweets from the source domain and pseudo-labelled tweets or self-labelled tweets from the target domain;

- Finetuning (FT) – modifying all the weights of the pre-trained model using the pseudo-labelled or self-labelled target data; and

- Feature Extraction (FX) – treating the pre-trained model as a feature extractor. Here, we do not retrain the model as in FT; instead, we only train a linear classifier using pseudo-labelled or self-labelled data on the top of the extracted features.

As a result, we compare eight classifiers (supervised [SL] learning structure and domain adaptation models) on eight settings, as shown in Table 6.2: (1) SL-LT, trained on MLTT and tested on MLTT (upper limit); (2) SL-LS, pre-trained on MLS and tested on MLTT (lower limit); (3) DS-TM; (4) SelfL-TM; (5) DS-FX; (6) SelfL-FX; (7) DS-FT; and (8) SelfL-FT. We consider the lower limit model to be our baseline, while the upper limit model is our ideal case. We believe that the domain adaptation results should lies between the results of these two supervised learning models.

In particular, we ask the following research questions:

- What is the performance of supervised classifiers that have only been trained on source labelled data to classify target data?

- When used to classify target data, how do the results of domain adaptation classifiers that use labelled source data and unlabelled target data compare to the results of supervised classifiers that solely use source data?

- How do the results of self-labelling compare to those of distant-supervised labelling when used in domain adaptation settings?

- How similar are the domain adaptation classifiers' results to those of supervised classifiers trained with target labelled data?

TABLE 6.2: Source and target set for each setting (S) in our experiments in this chapter.

Setting	Source Sets	Target Set
S1	Earthquake events: 2014-Chile, 2015-Nepal, 2013-Bohol, 2013-Pakistan.	2014-California Earthquake
S2	Earthquake events: 2014-California 2015-Nepal, 2013-Bohol 2013-Pakistan.	2014-Chile Earthquake
S3	Typhoon events: 2015-Pam, 2014-Odile, 2013-Yolanda, 2013-Oklahoma, 2012-Sandy.	2014-Hagupit Typhoon
S4	Earthquake events: 2014-Chile, 2014-California, 2013-Bohol, 2013-Pakistan.	2015-Nepal Earthquake
S5	Earthquake events: 2014-Chile, 2014-California, 2015-Nepal, 2013-Bohol.	2013-Pakistan Earthquake
S6	Typhoon events: 2014-Odile, 2013-Yolanda, 2014-Hagupit, 2013-Oklahoma, 2012-Sandy.	2015-Pam Cyclone
S7	Floods events: 2013-Queensland, 2013-Manila, 2013-Colorado, 2014-India, 2014-Alberta.	2014-Pakistan Floods
S8	Floods events: 2014-Pakistan, 2013-Manila, 2013-Colorado, 2014-India, 2014-Alberta.	2013-Queensland Floods

6.3 Results and Discussion

Based on the results shown in Table 6.3, we answer our research questions mentioned in Section 6.2 below.

What is the performance of supervised classifiers that have only been trained on source labelled data to classify target data?

As shown in the first row of Table 6.3, LS can be helpful when classifying target data. F1 scores for most settings are above 0.70 except for setting 8 (0.68). This outcome is consistent with earlier studies ([128, 82, 79]). However, we observe that the results are especially high when one or more source events and target events are similar in features other than crisis type (e.g., nearby locations or close occurring time). This is the case in settings 4 and 7. Nepal Earthquake, in setting 4, shares nearby locations with two events in the source data: the Pakistan and Bohol Earthquake incidents. Pakistan Floods, in setting 7, happened at a very close time to India Floods. After looking at the datasets, we find that a considerable number of tweets in the India Floods collection, one of the source events in setting 7, mention 2014

Pakistan Floods. In other words, users post information about Pakistan Floods in relation to India Floods in the India Floods data, which definitely causes the high F1 score (0.96) shown in Table 6.3. On the other hand, Queensland Floods in setting 8 does not share any of the common features other than the crisis type with Floods events in the source data, nor did any of the events happen shortly after Queensland Floods or in Australia. This explains the low F1 score compared to other settings (0.68).

When used to classify target data, how do the results of domain adaptation classifiers that use labelled source data and unlabelled target data compare to the results of supervised classifiers that solely use source data?

In Table 6.3, it is evident that at least one of the domain adaptation models outperforms the supervised classifier learned only from source data: DS-FT for settings 2 and 5, DS-FX for setting 4 and DS-TM for the five remaining settings. Based on these results, DS-TM can be seen as the best general approach among the other five domain adaptation classifiers regardless of the similarity between source and target domains, as it reports the best results in 5 out of 8 settings with only a very minor gap compared to the best score in the other settings (< 3%).

In contrast, we observe that domain adaptation techniques are not always better than supervised learning models learned from source data alone. For example, FT (with DS target data) inhibits Nepal Earthquake model performance by 0.9%, and FX (with the two labelling methods) harms Chile Earthquake model performance by 6.9%. This is based on the level of similarity between source and target data and the nature of the adaptation methods: the Nepal Earthquake data is very similar to the Earthquake source data while the Chile Earthquake data is different. Moreover, in FX, the high-level features of the source data are transferred to the target data, which requires a level of similarity between the two domains; in FT, more specific target features are incorporated through changing the weights of certain layers. This result is not consistent with the study in [82], where iterative domain adaptation techniques were used.

How do the results of self-labelling compare to those of distant-supervised labelling when used in domain adaptation settings?

Comparing rows 2 and 3 in Table 6.3, we can say that DS performs better than

SelfL when TM is used as an adaptation method in 6 out of 8 settings. For the two remaining cases, settings 4 and 5, SelfL-TM outperforms DS-TM with a gap of 1% in model performance. When FX is used to adapt pseudo-labelled or self-labelled target data, the results of DS and SelfL are almost identical, with a very minor improvement of less than 1% when using DS in seven out of eight test sets. For the last adaptation method in our experiments, FT, DS outperforms SelfL in 5 out of 8 settings with an average increase of 6% in F1 score. For the other three settings (1, 3 and 4), using SelfL as the labelling method instead of DS slightly improves performance (> 1.3%).

These outcomes can be explained by the nature of finetuning, feature extraction, distant supervision and self-labelling. DS produces pseudo-labelled target data with important keywords extracted from source data and new keywords derived from FrameNet. This can be very useful if the test set includes these derived keywords and the target features exist in the target tweets, such as location, infrastructure damage and people response. However, if the source and target data are alike, then SelfL can produce accurate self-labelled target data. For the adaptation methods, feature extraction is better than finetuning when the domains are similar, since the high-level features from the pre-trained model may be relevant to the target data. On the other hand, finetuning the model layers with target data captures the dataset features of the target event.

It is noteworthy that DS works better with finetuning when the source and target data differ. Unlike DS-FT, SelfL-FX is very suitable when the source and target data are similar. For example, in setting 2, the Chile Earthquake collection is very different from the source data; they do not share any common features. This reflects the sizeable gap (10%) between the results of SL and TL when classifying tweets from the Chile Earthquake event. As a result, finetuning the pre-trained model with pseudo-labelled data from Chile Earthquake is the best domain adaptation model, as shown in Table 6.3. The same scenario is repeated for setting 5 with the Pakistan Earthquake event. For the same reasons, DS-FT improves the performance by 7% from the supervised learning model only trained on source data when tested on tweets from the Queensland Floods event. However, the best reported domain adaptation

models in classifying data from the Queensland Floods event is DS-TM. This is because the dissimilarity level between source and target domain is very high. In this situation, building a target model from scratch using source and target data is better than using a pre-trained source model in all settings. The same scenario is found in setting 6 for the Pam Typhoon event with a high level of dissimilarity: the gap in F1 score between SL and TL is 19%. Here, as expected, DS-TM is the best reported domain adaptation model.

In contrast, for setting 4, DS-FT causes a drop in performance, whereas SelfL-FX is one of the best-performing models. This is because the Nepal Earthquake data is similar to the Earthquake source data, as mentioned above. Although data from the Pakistan Floods event is highly related to data from the India Floods event in the Floods source data, DS-TM still produces the best results in setting 7. On reviewing the data, we note that 7 out of 10 top keywords are present in tweets from the Pakistan Floods incident; the DS method accurately labels tweets from this event to the extent that building a target model along with the highly related source data performs better than other models.

As shown in setting 3 from Table 6.3, DS-TM is the best reported model. However, we note that source and target data only share 2 top keywords and 3 new keywords derived from FrameNet. This restricts the ability of the DS labelling method to produce good pseudo-labelled data from the Hagupit Typhoon event. However, because of the level of divergence between the source and the target events, SelfL produces noisy self-labelled data related to the Hagupit Typhoon incident.

Another interesting observation is that DS-FX and SelfL-TM produce similar results when testing on diverse target events, even though they use different labelling methods. This is possibly because they both use the same weights of the pre-trained source model to label or classify the target data. This is not the case in rows 4 and 6, where FT performs better than FX when one or more source and target events are different.

The most interesting observation is that the best reported domain adaptation classifier always use DS as its labelling method. This can be seen in all eight datasets (DS-FT for settings 2 and 5; DS-FX for setting 4; and DS-TM for the five remaining

settings). This indicates that DS is generally better than SelfL in automatically labelling tweets from emerging events regardless of the adaptation method.

How similar are the domain adaptation classifiers' results to those of supervised classifiers trained with target labelled data?

The last row in Table 6.3 represents the upper limit or ideal case in our experiments where the model learned only from manually labelled target data. Here, the best recorded semi-supervised domain adaptation model is very close to the results for the upper limit in settings 3 and 4, with a gap of less than 2% in F1 score. Nonetheless, this does not apply to all the settings: we observe that the gap increases to approximately 6% in setting 5 and reaches the maximum gap in settings 6 and 8 at approximately 11%. In contrast, the best domain adaptation model outperforms LT in two settings (1 and 7), with gaps of 5% and 0.3%, respectively. One possible reason is that the training data used to build the LT model for the California Earthquake event is very small (133 tweets only). Another reason is the number of shared top keywords in the Pakistan Floods event (7 out of 10). This gives the maximum benefits of using DS labelling method to produce good-quality pseudo-labelled data from the Pakistan Floods incident. When used as training data along with related source data, DS-TM outperforms LT in classifying unseen tweets from the Pakistan Floods event. In general, however, the results for domain adaptation models show much room for improvement.

TABLE 6.3: Results of our experiments in F1 score for eight models in eight settings.

Models/Settings	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
SL-LS	0.883	0.812	0.841	0.905	0.785	0.702	0.960	0.680
DS-TM	**0.935**	0.864	**0.879**	0.906	0.773	**0.779**	**0.975**	**0.794**
SelfL-TM	0.892	0.743	0.856	**0.907**	0.792	0.688	0.971	0.677
DS-FX	0.890	0.743	0.858	**0.907**	0.790	0.698	0.973	0.691
SelfL-FX	0.883	0.743	0.851	**0.907**	0.788	0.683	0.966	0.680
DS-FT	0.874	**0.871**	0.844	0.896	**0.802**	0.768	0.972	0.750
SelfL-FT	0.892	0.743	0.853	**0.907**	0.787	0.692	0.969	0.677
SL-LT	**0.886**	**0.912**	**0.902**	**0.915**	**0.856**	**0.894**	**0.972**	**0.899**

Note. The upper limit and the best reported results are highlighted in bold.

6.4 Further Analysis

At this point, we want to specify how many pseudo-labelled instances are needed from current events to build an accurate target classifier in domain adaptation settings. Our goal is to understand how the performance of DS-TM varies with different amounts of target data. In our opinion, this is an important factor for the practical usage of our method in reality. Thus, our last research question is: *What is the minimum number of pseudo-labelled target instances needed to build a good model to classify tweets from an emerging crisis event?*

6.4.1 Experiments

To study model performance based on the changes of the amount of incorporated target data, we run the experiments mentioned in Section 6.2 for the eight target sets with different numbers of instances (50, 100, 250, 500 and 1,000) for each class (related and not related tweets). The results in F1 score can be seen in Figure 6.1.

6.4.2 Results and Discussion

To answer our last research question in this chapter, we run the same experiments using balanced datasets with different numbers of examples from related and not related classes. We then examine how the DS-TM performance varies with different numbers of pseudo-labelled target instances compared to other domain adaptation models. Finally, we determine the minimum number of instances needed to train a robust target model.

What is the minimum number of pseudo-labelled target instances needed to build a good model to classify tweets from an emerging crisis event?

As can be seen from Figure 6.1, in subplots (a), (b), (f) and (h), DS-TM performs better than other domain adaptation models regardless of the number of instances from the target domain. This clarifies our assumption that DS-TM works better when source and target domains are dissimilar. The crisis events in these subplots are very different from the source domain. Conversely, for less dissimilar target sets, as in subplots (c) and (g), DS-TM is not always the best recorded model. In both cases, namely the California Earthquake and Hagupit Typhoon events, the F1 scores

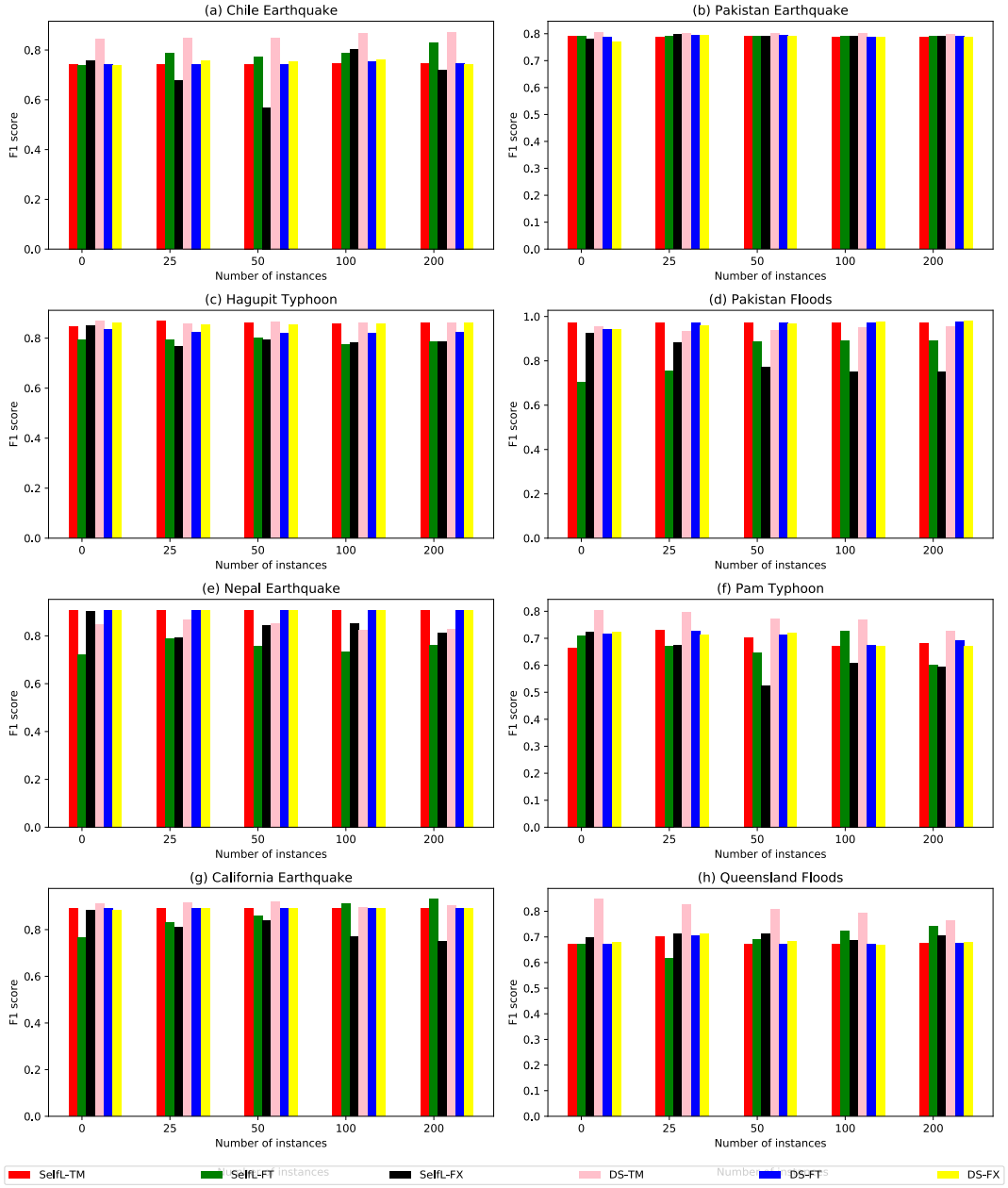

FIGURE 6.1: Results of English domain adaptation models in F1 score through different numbers of incorporated (pseudo-labelled and self-labelled) target data.

for DS-TM start at 0.91 and 0.834, respectively, and do not change greatly when incorporating more pseudo-labelled target data. For the last case, when source and target domains share a crisis feature other than crisis type, we observe that the results for DS-TM change in line with the number of instances if the incoming target tweets are added to the training data.

The most interesting observation here is that DS-TM always performs well at the beginning of the experiments (i.e., when the number of instances is 50 for each class).

The F1 score for DS-TM in all eight settings is above 0.80. This indicates that this domain adaptation model can be considered a robust model to classify unseen tweets from current events at disaster onset. Specifically, we can say that, for the Nepal Earthquake event, DS-TM is unstable and produces the worst results among all the other settings. However, this does not apply to the Pakistan Floods event, where the number of top keywords shared between the source and target data is 7 out of 10. Ultimately, the performance of DS-TM across different numbers of incorporated target tweets improves, reaching and sometimes bettering the performance of other domain adaptation models.

In summary, our results suggest that DS-TM is a robust model that can be used in practical situations. It only requires 50 pseudo-labelled tweets from both classes (related and not related) to successfully classify tweets from emerging events. Having said that, negative or not related examples can be gathered from tweets posted prior to the crisis occurrence time. This minimises the time required to collect these instances. We can say that our approach perfectly suits real-time crisis situations when quick decisions should be made by humanitarian organisations to help people in need.

Statistical analysis

Statistical analysis of the two-way ANOVA test is provided in Appendix B.

6.5 Conclusion

In this chapter, we introduced a simple but powerful semi-supervised domain adaptation approach for TCFCR by using a distant supervision-based framework to label the unlabelled target tweets. Our framework provides a new set of keywords rather than those extracted from available past events, thereby adding new features to the training data. This minimises the gap between the source and target domains caused by the domain shift of different crisis events. The experimental results show that our labelling method (DS) outperforms self-labelling (SelfL) in three different adaptation methods. Building a target model using the pseudo-labelled target domain data generated by DS and the available most-related source domain data improves model performance on seven out of eight datasets: from 0.1% to 11.4% absolute gain

in F1 score. Further analysis to determine the number of pseudo-labelled target data needed to build a robust target classifier shows that 50 tweets from each class (related and not related) are sufficient. This perfectly suits our task because it requires only minimal time at event onset, and it can be considered a general approach without need to predefine the similarity between source and target domains, unlike the other methods.

Aside from the limitations stated in Chapter 5, the proposed method in this chapter presents other drawbacks. First, the method requires an amount of unlabelled data from the emerging event. While pseudo-labelled data from the irrelevant class can be gathered before the emergence of the event, the pseudo-labelled target tweets from the related class must be collected after the crisis hits. There is no time specification to ensure this process; it could be seconds, minutes or hours, especially if we do not include duplicated or retweeted tweets. For some crisis events, this issue may cause a delay in the work of the humanitarian organisations. This is a general problem of semi-supervised or unsupervised domain adaptation methods; however, we believe that it can be partially solved by generating syntenic data from the limited number of available pseudo-labelled target tweets in certain cases.

In the next chapter, the proposed method will be applied to crisis tweets from low-resource languages such as Arabic. Some adjustments will be made to overcome the challenges in such languages, such as using clustering instead of labelled data to create the initial keyword list.

Chapter 7

Domain Adaptation for Arabic Twitter Data

1.8 billion Muslims, including 427 million native Arabic speakers, use Arabic as their liturgical language. Arabic thus represents the world's fifth most spoken language [76]. Furthermore, in terms of the number of Internet users, Arabic language users are the fastest-growing language group on the web. In February 2011, protestors in Egypt used Twitter as their main communication platform [127]. This emphasises the potential of Twitter in the Arabic world for spreading such crisis-related events and as an important and rich source of real-time and useful information. Humanitarian organisations could use this valuable human-generated information during a crisis in Arabic countries to help people in need. In prior works, deep learning algorithms have been used to identify crisis-related Arabic data to support disaster management and enhance situational awareness in the Middle East [8, 9, 3, 4] (see Section 2.1.2). However, they did not consider the domain-shift between source and target tweets posted during Arabic crisis events.

Our work in this chapter is inspired and motivated by the success of applying our proposed domain adaptation method to high-resource English-language tweets in the last chapter. Unlike English, Arabic is considered a low-resource language, with several notable issues highlighted in the crisis literature. Hence, some challenges are to be expected in applying our pseudo-labelling method to Arabic tweets. First is the lack of labelled Arabic tweets for crisis response. The first published dataset, Kawarith, has been recently released by Alharbi and Lee (in 2021) [9], which

contains labelled data from 7 Arabic crisis events (see Section 4.1.1.2; Table 4.2). Second, the lack of good supporting resources for Arabic, such as external knowledge bases or language dictionaries [8]. Finally, Arabic tweets are informal and regional in nature, and Arabic regions have unique dialects. Moreover, many written Arabic dialects differ in syntax, phonology and morphology [37]. Those users posting tweets usually write in their regional dialects. Thus, the keyword set generated from one crisis may be insufficient when used to label tweets from another if the two crises come from different regions with different dialects.

In this chapter, we propose an adaptive domain adaptation method for Arabic crisis response that overcomes all these challenges. To the best of our knowledge, this is the first attempt to use distant supervision under the umbrella of domain adaptation techniques to classify unseen crisis-related Arabic data from current events. The experimental results show that the Arabic version of our domain adaptation method can be seen as a robust approach to classifying unseen Arabic tweets from an emerging event. In addition, incident data related to different crisis types of target events can be used to create the initial keyword list. This finding reduces the necessity of having keywords from the same crisis type as the target disaster. Furthermore, it extends our framework's abilities to automatically label data from low-resource languages with limited capabilities.

7.1 Method

We use the same two-stage domain adaptation method mentioned in Section 6.1 in the previous chapter, driven by the method in Section 5.1. However, due to the expected challenges in applying this method to Arabic tweets, we make some significant adjustments to the details of the pseudo-labelling. We use clustering to produce annotated data rather than manually labelled data to create the initial keyword list. We also use an Arabic external resource (Almaany) instead of FrameNet to expand the original list. In addition, we use different crisis types to establish this list alongside using the same type of the target event. The changes are listed as follows:

The changes are listed as follows:

First change: In Chapters 5 and 6, we use manually labelled data from different

events related to the same crisis types in step one to apply our framework successfully and create a robust initial keyword list. However, this is infeasible for Arabic crisis tweets [9]. Thus, we use an unsupervised method – clustering – to classify several Arabic corpora from different events to overcome this problem. The chosen clusters are then used instead of the manually labelled data in Chapters 5 and 6 to create the initial keyword list. We follow the authors in [61] and [2] in utilising ISIRI Stemmer from NLTK 3.4 on Twitter data to stem each word to its root to avoid word redundancy and reduce the amount of linguistically similar words. This chapter calls the group of data used to create the initial keyword list the "keyword set". It is critical here to say that we create initial keyword lists for only two crisis types: Floods and Explosion. This is because most of the available unlabelled corpora (see Section 4.1.2.2) are related to Floods events (12 out of 20 events) or Explosion events (2 out of 20 events). Another reason is the lack of gold-standard testing data for the given crisis types. For example, two unlabelled corpora are available for Fire and Shooting events; however, the gold-standard testing data related to a disaster event from the Fire or Shooting crisis types are unavailable, which leads us to exclude Fire and Shooting incidents from our experiments.

Clustering model

Tweet clustering is an unsupervised method where posts are grouped according to the common characteristics that differentiate them from other groups. These groups are referred to as clusters. K-means is one of the most widely used clustering algorithms to classify tweets into groups based on their inherent distances from each other [113]. It is a centre-based method: the centre of each group is used to determine whether the data point is related to the given group. It is also a distance-based algorithm: the data point's group is assigned based on the calculated distances between the centre of the groups and the data points. First, the number of clusters (k) is determined. Random data points are then selected as cluster centroids, and all the points are assigned to the closest centroid. The centroids for the newly created clusters are then recomputed. Finally, we repeat the assignment of all data points and the recomputing of centroids until the process stops according to the given criteria. These criteria include reaching the maximum number of iterations, data points remaining

in the same clusters, and centroids that are unchanging [13]. The main objective of K-means is to minimise the sum of the distances between the data points and their corresponding centroids [113].

According to Alruily in his review paper [19], the K-means clustering model is the most suitable unsupervised method for Arabic tweet problems. It has been successfully applied to different Arabic Twitter data [1, 112, 110]. In addition, the concept of K-means is similar to the concept of word embedding in identifying the relationships between words in tweets. Word embedding has proven to be a decisive factor in tweet classification; we think that clustering has the same potential. This chapter applies K-means to each event corpus mentioned in this section. This is because if we merge all data regardless of the events, the created clusters will be divided by the events themselves. We also choose the optimal number of clusters, label the clusters and select the most related ones, as shown in the following subsections.

Cluster optimisation

Determining the number of clusters is required in advance when applying K-means to data. Although finding the optimal number of clusters is a time-consuming process, it is a very important step to ensure a good separation. To do so, we use elbow method and silhouette score measurements. We find the elbow method is uncertain for our data because the results shown in the figures are not clear. Figure 7.1 shows an example of the unclear results when using the elbow method to decide the optimal number of clusters for the Covid'19 corpus. We can see that the optimal number cannot be determined because, in our opinion, there is no point position on the elbow arm in the diagram. On the other hand, the silhouette score for the Covid'19 corpus is undoubtable on our data. Figure 7.2 shows silhouette diagrams for varying k (from 8 to 12) and including silhouette plots and visualisations of the clustered data. We observe from the diagram that almost all the silhouette coefficients from all the clusters are beyond the silhouette score dashed line and are close to 1. This means that all the settings are good choices. It also indicates that the clusters are probably well separated with no overlapped instances in the data. Additionally, the

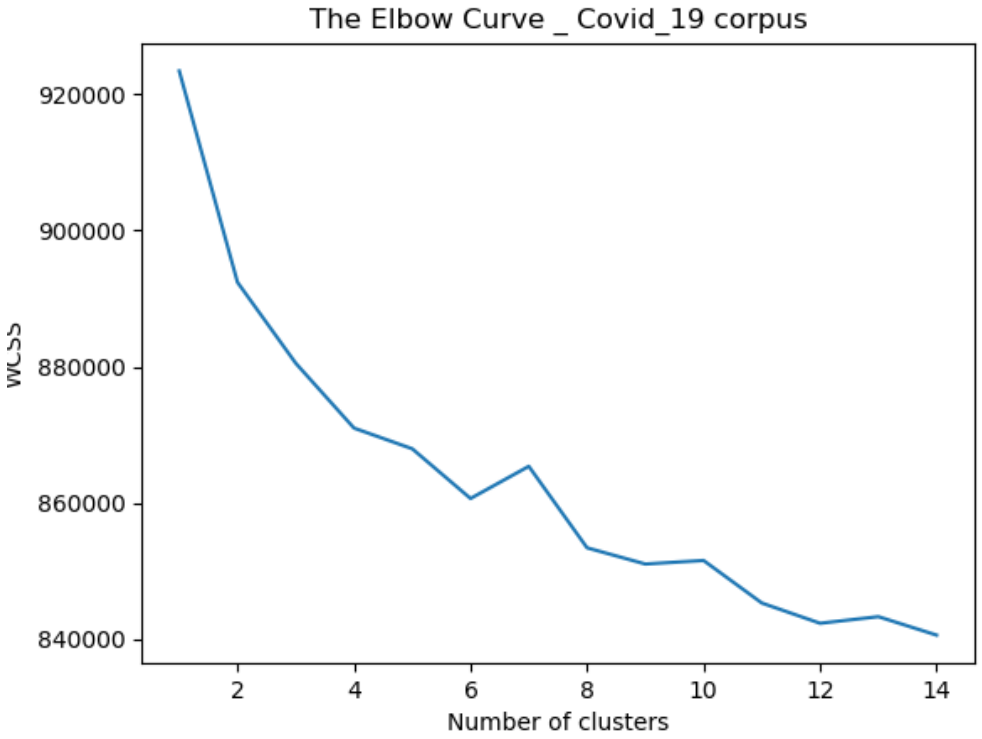

FIGURE 7.1: Elbow curve for Covid'19 corpus.

size and width of the clusters give no hint of the best setting because we know beforehand that the crisis dataset is unbalanced. Therefore, the best setting is the one with the highest silhouette score, which is setting 11 (0.488). In this example, we can say that the optimal number of clusters for the Covid'19 corpus is 11.

Extracting features of clusters

After determining the optimal number of clusters and applying K-means to the data for every crisis event from the unlabelled corpora, we need to assign profiles as labels for each cluster. The reason behind labelling the clusters is that assigning profiles that describe the tweets within the clusters is another way to decide whether the cluster is related to the crisis and informative. To do so, we follow the centroid approach: we pick the centre data point of each cluster to extract the cluster's features. This approach is suitable for our work because the variance within the clusters is slight, and the centre data point of the cluster is the closest one to represent the cluster. On the other hand, other approaches – such as supervised learning for cluster membership and the empirical approach – are not suitable for our data. The cluster

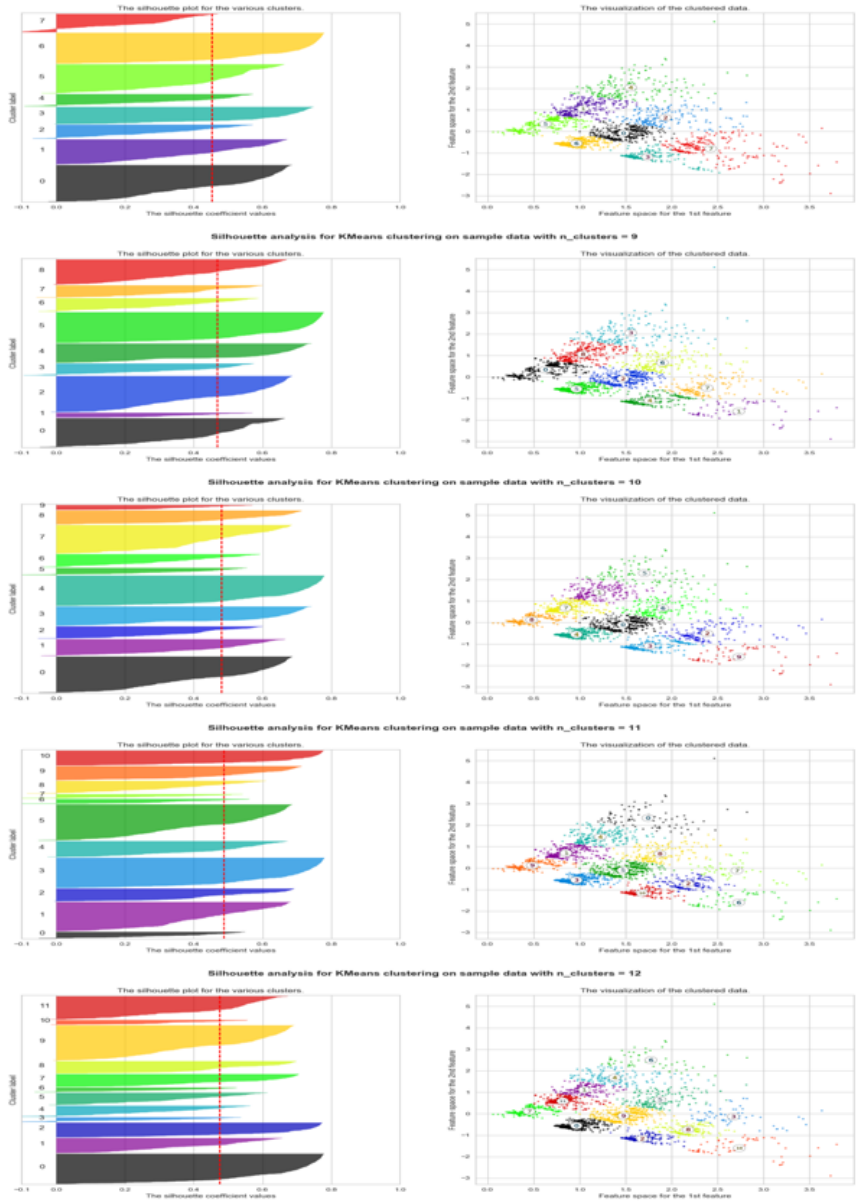

FIGURE 7.2: Silhouette diagrams for Covid'19 corpus.

membership approach uses certain data points to train the model to be applied to classify the others, while the empirical approach studies the characteristics of the clusters. Our data are similar in that the tweets are all posted during a crisis and different in providing information about the crisis. Therefore, such approaches can be misleading for our data.

To ensure the effectiveness of the centroid approach, we select the closest three data points instead of one. We then extract the features for each cluster and use them to assign profiles from these data points. Our data represents many crisis topics, including: advertisements; political opinions; irrelevant to the crisis; emotional

support; infrastructural and utility damage; dead, injured or affected people; and providing help and caution advice. For example, the closest three data points for cluster #3 in the Beirut Explosion corpus (3,445 tweets) are:

حصيلة قتلى انفجار ميناء # بيروت ترتفع إلى ١٣٥. https://t.co/Hbah7vFFKi. https://t.co/HLYVLF7zco

,

أشبه ما يكون بما حدث في هيروشيما وناجاساكي أعداد كبيرة من القتلى والجرحى # انفجار _بيروت # انفجار _مرفأ _بيروت

and

اوضح فيديو للانفجار الهائل الذي هز العاصمة اللبنانية # بيروت الانفجار نتج عنه دمار كبير وانباء عن سقوط عشرات الجرحى https://t.co/uwD2JWpyF4

These tweets contain the words قتلى, which mean (dead people), انفجار (explosion) and جرحى (injured people). It is obvious that the most represented tweets of the cluster talk about dead and injured people during the Beirut Explosion incident. As a result, this topic is assigned to cluster #3.

Choosing the most related clusters

After assigning topics to clusters, we divide the clusters into two classes: related and informative and irrelevant or not informative. In particular, infrastructure and utility damages, dead, injured or affected people, and providing help and caution advice are classified as related and informative. On the other hand, advertisements, political opinions, and emotional support are labelled as irrelevant or not informative. While doing this, we observe that all the crisis events have a cluster with a vast number of tweets advertising for specific products or services. Although clusters with tweets expressing political opinions are related to the crises, we decide to label them as not informative because the information in the posted tweets offers no

benefits to humanitarian organisations. This decision differs from one language to another since, unlike English, the number of Arabic tweets that reflect political opinion after an emerging disaster cannot be underestimated and thus can ultimately impact the keyword lists, which eventually influences the results of our proposed approaches. In addition, the performance of supervised models classifying English crisis-related tweets improves when the emotional support class is ignored in the training data [134]. Thus, we decide to label emotional support as not informative in addition to the fact that this category does not provide any useful information to humanitarian organisations.

Second change: To select the top K keyword list for Arabic crisis events, we remove stop words such as "هذا ، من ،في", hashtags such as "#انفجار", places such as "حفرالباطن" and useless Twitter-specific words such as "RT" and "via". We then compute $KW(_i)$ (mentioned in Chapter 5) for all the words in the initial keyword list from step one and sort them according to their KW values. We then select the top K keywords of a given crisis type. For instance, for the Floods crisis type, the top K keyword list contains "سيل","غرق" ,and "مطر", which have the highest KW values compared to the other words in the initial Floods list.

Table 7.1 shows how KW values play an important role in indicating the most vital keywords for the Floods crisis type. We can see that crisis-related and flood-related words have higher KW values than the unrelated ones.

TABLE 7.1: KW values of selected words from the initial Floods keyword list from keyword set #7.

Ranking	Keyword	KW Value
1	مطر	0.00019
2	غرق	0.00032
4	حذر	0.00045
32	كسر	0.00065
51	نهر	0.00130
67	سجل	0.00184
98	رادار	0.00371

Third change: To apply distant supervision to the Arabic initial keyword list, we expand the list to include similar semantic words from the Almaany Arabic dictionary rather than FrameNet. Arabic is one of the low-resource languages for which

powerful knowledge bases are unavailable. However, dictionaries have proven to be an excellent resource to expand Arabic keyword lists [24, 61].

The Almaany dictionary [11] has been widely used by Arabic researchers [87, 57], including recently to collect specific words to classify Arabic Twitter data [12, 15, 126]. Almaany is an online dictionary that provides corresponding meanings with similar semantic words for each term in many languages, including Arabic [126]. In our work, we use Arabic for the Arabic version of Almaany. We retrieve all the synonyms provided by Almaany for each crisis keyword if the corresponding meaning of the top keyword is related to the crisis type. For example, the top keyword سيل exists in the Almaany dictionary but with two corresponding meanings based on the shape and the signs of the word: سَيْل and سيّل.The meaning of سيْل is the water of the rain that rushes over the earth's surface, whereas سيّل refers to converting material from a solid state to a liquid state. According to their meanings, سَيْل is related to the Floods crisis type, but سيّل is not. Thus, all the synonyms associated with سَيْل, such , such as فيضان and طوفان can be mapped to سَيْل, which is a crisis keyword gathered from step one and selected in step two as one of the top K keywords based on its high KW value. In other words, if one of the top crisis type keywords exists in the Almaany dictionary and its meaning relates to a given crisis type (Floods or Explosion in this chapter), then distant supervision assumes that all the synonyms related to the given word express that crisis type. As a result, the number of keywords increases in the final list. For instance, the number of keywords rises from 10 to 78 in the keyword list for the Floods crisis type. This list contains two types of keywords: strong keywords (top K keywords) and weak keywords (extracted from Almaany). If a word exists in the top K keywords and is a synonym associated with another top K keyword at the same time, then we consider it a strong keyword. Weak keywords may bring noise to the data, which we try to reduce in step five.

We choose the related words based on their meanings from the Almaany dictionary manually, as it is hard to decide this without human involvement. Certain cases arise where we cannot map a top K keyword to any synonyms in Almaany. Some of the top K keywords may not exist as words in Almaany, such as صور in the top K keyword list for the Floods crisis type. In these cases, the keyword is not mapped to any synonyms, and the number of expanded keywords remains the same. Table

7.2 lists some of the keywords from the Floods and Explosion crisis types and their mapped synonyms from Almaany when the keyword and the related meaning both exist.

TABLE 7.2: Keywords from keyword lists for different crisis types and their synonyms from Almaany.

Keyword	Crisis Type	Related Meaning	Associated Synonyms
سيل	Floods	The water of the rain as it rushes over the surface of the earth.	تَدفُق، جَرَيان ، طُوفَان ، عُبَاب، فيضان ، مَسِيل ، انْصِبَاب ، اِنْصِبَاب ، اِنْسِكَاب انصباب، انسياب، انهيار
أَسعف	Explosion	Providing help and assistance to someone	آزَرَ ، أَجَارَ ، أَعانَ ، أَغَاثَ ، أَلْجَأَ ، أَنْجَدَ ، أَنْجَى ، أَنْقَذَ ، حَمَى ، خَلَص ، دَافَعَ عَنْ ، دَعَمَ ، رَفَدَ ، ساعَدَ ، ساعَفَ ، سَاعَدَ ، سَانَدَ ، صَانَ ، ضَافَرَ ، عانَ ، عَاضَدَ ، عَاوَنَ ، عَضَدَ ، نَجَدَ ، نَجَّى ، نَصَرَ ، وزَرَ ، وَقَى
انفجر	Explosion	Smashing something violently with a bang	تَدَفقَ ، أَرَاقَ ، إِنْبَثَقَ ، إِنْبَجَسَ ، خَرَجَ ، دَفَقَ ، سَالَ ، فاضَ ، نَبَعَ ، هَرَاق

Continuation of Table 7.2

Keyword	**Crisis Type**	**Related Meaning**	**Associated Synonyms**
كارثة	Explosion	Great misfortune and great ruin. Sometimes naturally caused by earthquakes, floods and hurricanes.	أَزْمَة ، خَطْب ، داهِيَة ، ضَرّاء ، فاقِرَة ، كَرْب ، مُصِيبَة ، مُلِمَّة ، نائِبَة ، نازِلَة ، نَازِلَة ، نَكْبَة ، هَلاَك ، وَاقِعَة ، أَزْمَة ، إِبْتِلاء ، بَغْضَاء ، بَلاَء ، بَلِيَّة ، بَلْوَى ، جائِحَة ، حَادِث ، حَادِثَة ، خَطْب ، داهِيَة ، شِدَّة ، صُروف الدَهر ، صُرُوف الدَهْر ، صُرُوف الدَهر ، ضائِقة ، ضِيق ، طامَّة ، طَارِقَة ، عُسْر ، غَائِلَة ، غَاشِيَة ، فَاجِعَة ، فَادِحة ، فَجِيعة ، فِلْق ، كَرِيهَة ، كُرْبَة ، لأْوَاء ، مُصِيبَة ، مُلِمَّة ، مِحْنَة ، نائِبَة ، نازِلَة ، نَقْمَة ، نَكْبَة ، وَاقِعة
إخلاء	Flood	Evacuation of the city; the displacement of its inhabitants.	إِعَاد ، إِجْلاء ، إِخْراج ، إِشْغار ، إِطْلاق ، إِفْراغ ، إِقْصاء ، تَرْحيل ، تَفْريغ ، جَلاء ، طَرْد ، نَفْي ، إِفْرَاغ

In this chapter, we follow the method in Chapter 6 to label the tweets gathered

by tweet IDs provided in Kawarith, as mentioned in Section 4.1.2.2 However, to expand the abilities of our framework when applied to low-resource languages that face the same problems, we use keyword sets from different crisis types to automatically label tweets from a crisis event. For example, we use the keyword set for the Explosion crisis type to automatically label tweets from the Covid'19, Dragon Storm and Cairo Explosion corpora.

Pseudo-labelled data

Table 7.3 lists examples from pseudo-labelled tweets created by our framework from some of the Arabic target events in our work. We note that some tweets are labelled by our framework with different classes when different crisis types are used to create the keyword lists and keyword sets. For example, in the Jordan Floods corpus,

RT @aa_arabic: # عاجل # سيول ـالأردن ٩ قتلئ و ٩ مصابين و إنقاذ آلاف و الحالة غير مستقرة (الحكومة متحدثة)

is labelled as not related when Floods are used as the keyword set because none of the strong or weak keywords are found in the tweet. On the other hand, the tweet is labelled as related when Explosion is used to create the keyword list due to the presence of a strong keyword "انقاذ" and a weak keyword "اصاب". The word "مصابين" is derived from "اصاب", which is associated with the strong keyword "كارثة". It should be noted that "انقاذ" and "كارثة" are in the keyword list for the Explosion crisis type but not for Floods, which is the crisis type of the Jordan Floods event. However, most tweets are labelled with the same class regardless of the crisis type used to create the keyword lists. For example,

RT @ww6223ww6: بالتوفيق بإذن الله لابناء العم في انتخابات الغرفة التجارية في فئة الصناعيين و فئة التجار #حفر ـالباطن https://".co/2payPj3hU

is a tweet from the Hafer-albatin Floods corpus that is labelled as irrelevant regardless of the crisis type of the keyword set (Explosion or Floods).

TABLE 7.3: Examples from Arabic pseudo-labelled data created by our distant supervision-based framework. Note that the given target crisis event is excluded from the source events used to create the crisis-type keywords.

Target Crisis Event	**Tweet**	**Keyword Set**	**Pseudo-label**	**Reasons**
Jordan Floods	ارتفاع عدد وفيات طلاب حادثة البحر الميت الى ٧١ طالب و لا زال جاري البحث عن ٤٤ طالبا و معلما جرفتهم السيول سيول_الاردن # https://t.“o/Cml2jy6Dxt	Same crisis type: Floods	Related and informative	Two strong keywords exist in the given tweet: ”بحر” and “سيل”.
Covid’19	ناس خايفين من كورونا و ناس تشل و تحط خطبات و ملكات بروید بروید ان شاء الله لاحقين ماهو ذا الشغل خطبات و ملكات بالله عندكم و الا بس عندنا ؟؟	Different crisis type: Floods/ Explosion	Not related / not informative	Absence of keywords from final Floods/Explosion crisis-type list.
Jordan Floods	لا تصدق حلو اللسان و لا جميل فوق الراس الجمايل يبرد دفاها و تطيح https://t.o/3zalDWn1eY	Same crisis type: Floods	Not related / not informative	Absence of keywords from final Floods crisis-type list.

Continuation of Table 7.3

Target Crisis Event	**Tweet**	**Keyword Set**	**Pseudo-label**	**Reasons**
Dragon Storm	RT @drzawba: وزارة الصحة تعلن رفع درجه الاستعداد للقصوى بالمستشفيات و توفير مستلزمات الطوارئ و تكثيف سيارات الإسعاف بعد إعلان الأرصاد الجوية.	Different crisis type: Explosion	Related and informative	Two strong keywords from Explosion keyword list in the given tweet: ” استعد” and ” طوارئ” .
Beirut Explosion	RT @AlArabiya: شاهد انفجار # بيروت يداهم السيارات و المارة عل مسافات بعيده و ينشر الذعر و الرعب # انفجار■بيروت # العربية https:/t.co/8UVEKJC	Different crisis type: Floods	Related and informative	Two weak keywords exist in the given tweet (” رعـب ” and “ ذعـر ”), which are associated with the strong keyword “خوف”.
Cairo Explosion	# معهد■الأورام اللهم اجرنا من نار جهنم	Same crisis type: Explosion	Not related / not informative	Absence of keywords from final Explosion crisis-type list.

Continuation of Table 7.3

Target Crisis Event	**Tweet**	**Keyword Set**	**Pseudo-label**	**Reasons**
Jordan Floods	RT @aa_arabic: # عاجل # سيول ـالأردن ٩ قتلئ و ٩ مصابين و إنقاذ آلاف و الحالة غير مستقرة (الحكومة متحدثة)	Different crisis type: Explosion	Related and informative	Two keywords exist in the given tweets: one strong keyword ("إنقاذ") and one weak keyword ("اصاب". "مصابين" is derived from "اصاب", which is associated with the strong keyword "كارثة".
Hafer-albatin Floods	RT @ww6223ww6: بالتوفيق بإذن الله لاابناء العم في انتخابات الغرفة التجارية في فئة الصناعيين و فئة التجار #حفر ـ الباطن https://".co/2payPj3hU	Same crisis type: Floods	Not related/ not informative	Absence of keywords from final Floods crisis-type list.
Dragon Storm	RT @3ashoouur: إن شاء الله العاصفه الجايه نكون محبوسين انا و إنتي في بيت واحد	Different crisis type: Floods	Discarded	Only one weak keyword occurs in the tweet: "عاصفة" is derived from "عـصف", which is associated with the strong keyword "اعصار".

Continuation of Table 7.3

Target Crisis Event	Tweet	Keyword Set	Pseudo-label	Reasons
Cairo Explosion	@RT @masrawy: # عاجل مصرع ٥ إصابة ٥١ آخرين الداخلية تكشف تفاصيل انفجار معهد الأورام	Same crisis type: Explosion	Related and informative	Two strong keywords exist in the given tweets: "إصابة" (derived from "اصاب") and "انفجار" (derived from "فجر").
Beirut Explosion	RT @azzawil: تعليق ترامب خطير عن احتمال ان يكون # انفجار■بيروت نتيجة لهجوم و ليس مجرد حدث عارض. هذه المرة كان يقرا نصا مكتوبا لم يرتجل.	Different crisis type: Floods	Discarded	Only one strong keyword occurs in the tweet: "خطير", derived from the strong keyword "خطر".

The adaptation stage mentioned in Section 6.1.2 remains the same in this chapter.

7.2 Experiments

Our main goal is to investigate whether automatically labelled target data generated by a framework via distant supervision can be used to build a robust model along with other source events to improve model performance in classifying unseen Arabic tweets from emerging events. To this end, we use the same experiments mentioned in Chapter 6 for domain adaptation for Arabic crisis response. In these experiments, we use the settings in Table 7.4: keyword sets mentioned in Section 4.1.1.2 and target sets from Table 4.2.

In this chapter, we ask the following specific research questions:

- What is the performance of supervised classifiers that have only been trained on source labelled data to classify target Arabic data?
- When used to classify target Arabic data, how do the results of Arabic domain adaptation classifiers that use labelled source data and unlabelled target data compare to the results of supervised classifiers that solely use source data?
- How do the results of self-labelling compare to those of distant-supervised labelling when used in domain adaptation settings for Arabic tweets?
- How similar are the Arabic domain adaptation classifiers' results to those of supervised classifiers trained with target Arabic labelled data?

We also extend these experiments to include using keyword sets from different crisis types of the target event. For example, in Chapter 6, we use keyword sets from Floods crisis events only to automatically label target data from an emerging Floods event. Here, we extend this to include using keyword sets from Floods crisis events to automatically label target data from an emerging Explosion event, Virus event, or Storm event, as shown in Table 7.4. Thus, we ask the following questions:

- How do the results of Arabic domain adaptation classifiers that utilise distant supervision to automatically label target data when using keyword sets from similar crisis events compare to those using keyword sets from different crisis events in classifying Arabic tweets from emerging events?
- How do the results of self-labelling compare to those of distant-supervised labelling with keyword sets from another crisis type of the target event when used in Arabic domain adaptation settings?

In addition, we add a standalone model that is trained only on the generated labelled data from the emerging events. The training data does not include manually labelled data from any crisis events, including the emerging event. We add this model to gain an insight into the quality of our generated labelled data. To this end, we ask the following question:

- How similar are the supervised classifiers' results trained on target Arabic generated labelled data to those of supervised classifiers trained solely on Arabic source data?

TABLE 7.4: Keywords and target sets for each setting (S) in our experiments.

Setting	Keyword Set	Target Set
S1	2020-Beirut Explosion	2020-Cairo Bombing
S2	2020-Cairo Bombing	2020-Beirut Explosion
S3	Hafer-albatin and Kuwait Floods	2018-Jordan Floods
S4	Hafer-albatin and Jordan Floods	2018-Kuwait Floods
S5	Kuwait and Jordan Floods	2018-Hafer-albatin Floods
S6	Floods	Covid'19
S7	Explosion	Covid'19
S8	Floods	Dragon Storm
S9	Explosion	Dragon Storm
S10	Floods	2020-Cairo Bombing
S11	Floods	2020-Beurit Explosion
S12	Explosion	2018-Jordan Floods
S13	Explosion	2018-Kuwait Floods
S14	Explosion	2018-Hafer-albatin Floods

7.3 Results and Discussion

Based on the results shown in Tables 7.5 and 7.6, we answer our research questions mentioned in Section 7.2 below.

What is the performance of supervised classifiers that have only been trained on source labelled data to classify target Arabic data?

LS can be useful when classifying target Arabic data, as shown in the first row of Tables 7.5 and 7.6. F1 scores for most settings are above 0.70, except for settings 8 and 9 (0.658), which represent the same target data (Dragon Storm). This outcome suggests that crisis data from other crisis types of the target event can be used to train a model for identifying Arabic tweets for crisis response. This result is consistent with prior studies [92, 81]. However, and unlike the English models in the previous chapter, we observe that the highest result is still below an F1 score of 0.80. This is because the models are trained using the same data from various Arabic crisis events of different disaster types.

Nonetheless, Jordan Floods, in settings 3 and 12, occurred after the emergence of Kuwait Floods. After analysing the Jordan and Kuwait Floods data, we note that an extensive number of tweets in the Kuwait Floods collection – an event in the source and keyword sets – comment on Jordan Floods. Users share information about Jordan Floods in relation to Kuwait Floods in the Kuwait Floods data. This definitely causes the high F1 score (0.79), as presented in Table 7.5. This score is not consistent with the highest score in English tweets in the previous chapter (0.96). This is because people in Kuwait and Jordan use different dialects while posting tweets about crises. On the other hand, Dragon Storm in settings 8 and 9 does not share any of the common features, especially the crisis types, with the source events or the keyword sets. None of the events happened shortly after Dragon Storm or at the locations of the event (Syria and Palestine). Moreover, because Arabic tweets are region-based, dialects used in the Dragon Storm data have not been used in the source data. These observations explain the low F1 score compared to other settings (0.658). This is not the case for the Covid'19 event, since dialects used to post tweets about Covid-19 have been used in the data of the source event, including Saudi and Kuwaiti. This observation clarifies the gap in F1 scores between Dragon Storm and Covid'19 (0.658 < 0.744).

When used to classify target Arabic data, how do the results of Arabic domain adaptation classifiers that use labelled source data and unlabelled target data compare to the results of supervised classifiers that solely use source data?

It is evident in Table 7.5 that at least one of the domain adaptation models outperforms LS. The highest scores are recorded by DS-TM for all the settings except settings 5 and 6, where SelfL-FX and SelfL-TM perform the best, respectively. In contrast, it is clear that domain adaptation techniques are not always better than supervised learning models learned from source data alone. For example, SelfL-FX (with self-labelled target data) causes the Beirut Explosion model's performance to decrease by 18%, while SelfL-FT (with self-labelled target data) causes the Hafer-albatin Floods model's performance to fall by 4%. This is based on the level of similarity between source and target data and the nature of the adaptation methods. In FX, the high-level features of the source data are transferred to the target data, which requires a level of similarity between the two domains; in FT, more specific target

features are incorporated through changing the weights of some layers. Having said that, the Beirut Explosion data differs from the source data even with the existence of another explosion event (Cairo Explosion). The Cairo and Beirut Explosion data are written in different dialects and have dissimilar characteristics: Cairo Explosion was a terrorist act, whereas Beirut Explosion was caused by mismanagement on the part of the Lebanese government. On the other hand, the two Floods events in the source data used to train the model make the Hafer-albatin Floods data very similar. This result is consistent with our work in Chapter 6 for English tweets.

To summarise, DS-TM can be seen as the best general approach among the other five domain adaptation classifiers – regardless of the similarity between source and target domains – as it reports the best results in three out of five settings and a very minor gap compared to the best score in the other two (< 1%).

How do the results of self-labelling compare to those of distant-supervised labelling when used in domain adaptation settings for Arabic tweets?

We can say from rows 2 and 3 in Table 7.5 that DS performs better as a labelling method than SelfL when TM is used as an adaptation method in 4 out of 5 settings. For the remaining setting, setting 5, SelfL-TM is better than DS-TM with a gap of 1% in model performance. However, it is clear from the results that DS-TM always improves the performance by an average of 5.5%. In contrast, SelfL-TM causes a decline in performance for 4 out of 5 target events (average of 12.2%). The model performance when feature extraction (FX) is used to adopt pseudo-labelled target outperforms that with self-labelled data in 3 settings (2, 3 and 5). The same scenario is replicated for the last adaptation method, finetuning (FT).

These outcomes suggest that the impact of the labelling method is greater than the impact of the adaptation methods when pre-trained models are used. This can be explained by the nature of DS and SelfL. DS produces pseudo-labelled target data with important (initial) keywords extracted from the keyword set with the same type and new keywords derived from Almaany. This can be very useful if the test set includes these initial or derived keywords. However, if the source and target data are alike in terms of having similar event features (e.g., location, infrastructure damage and people response) besides language features such as dialects, then SelfL can produce accurate self-labelled target data. As shown in settings 1 and 2 from

Table 7.5, DS-TM is the best reported model. Both incidents, Cairo Explosion and Beirut Explosion, are different from each other as described before. In addition, the data for both include 5 out of the 10 top keywords and more than 50% of the expanded Explosion keyword list. This is the ideal situation for DS as the labelling method in domain adaptation settings.

Although the data from the Jordan Floods event are highly related to the data from the Kuwait Floods event in the source data and the Floods keyword set, DS-TM produces the best results in setting 3. On review, we observe that 5 out of the 10 top keywords are present in tweets from the Jordan Floods incident. Additionally, 62.5% (50 out of 80) of the expanded keyword list occur in the target data. This increases the ability of the DS labelling method to accurately label tweets from this event to the extent that building a target model along with the source data performs better than other models. In setting 5 (Hafer-albatin Floods), SelfL-TM outperforms other domain adaptation methods. The reason behind this result is that Hafer-albatin is very similar to the other two Floods events, especially Kuwait Floods. Hafer-albatin and Kuwait are proximal locations and share a language feature (dialect). Another reason is that the incident data contain 5 out of 10 initial Floods keywords, yet the percentage of the expanded keywords from Almaany is low (38%).

For Kuwait Floods, it is clear that SelfL-TM should report better results than DS-TM because of the similarity level with Hafer-albatin Floods and the small number of common top keywords (3 out of 10). Surprisingly, however, DS-TM performs better than SelfL-TM for setting 4. This can be explained by the nature of the Arabic language. Unlike English, any root word in Arabic has more than 10 shapes regardless of the language signs; the expanded keyword list that contains only root words is extended automatically by all these shapes. This helps our framework to retrieve more related tweets. Setting 3 is an example of this: most of the expanded keywords occurring in the target data are shapes from root words such as "تحذير، حذر، يحذرون، يحذر" from "حذر". This represents a significant advantage in using our framework to automatically label Arabic crisis tweets from emerging events.

We also note that, in setting 2, both labelling methods cause a substantial drop in model performance when FT or FX is used as the adaptation method, unlike in the other settings. This is because of the high level of divergence between the source

and target domains – to the extent that using a pre-trained model in the domain adaptation method always inhibits model performance.

In general, similarity between source/keyword and target sets has a major impact on choosing the best domain adaptation model in detecting related Arabic crisis tweets. If the data are similar, then SelfL is most likely to be better than DS in labelling target data, as seen in settings 4 and 5. In addition, the number of common keywords between initial/expanded keyword list and target data is also considered to be a strong factor in selecting the best model. Conversely, if the number of common words is equal to or greater than 5 in the initial keyword list or more than 50% of the expanded keywords are shared, then DS is recommended as the labelling method instead of SelfL (as in settings 1 and 2). However, pre-determining the level of similarity or the common keywords is unfeasible in real-world crises.

How similar are the Arabic domain adaptation classifiers' results to those of supervised classifiers trained with target Arabic labelled data?

The last row in Table 7.5 represents the upper limit (ideal case) in our experiments, where the model learned from manually labelled target data. The best recorded semi-supervised domain adaptation models for all settings are very far from the results for the upper limit. This is not consistent with the results from the last chapter. One possible explanation is the source data used in our experiments for the two languages: while the source data are built from disasters with the same crisis type as the target event for the English experiments, the source data are collected from events from various crisis types for the Arabic ones. We also note that the minimum gap is approximately 5% in setting 2 (Beirut Explosion data), while the maximum gap appears in settings 1 (Cairo Explosion) and 3 (Jordan Floods), with F1 scores of approximately 11%. In general, therefore, the results of these Arabic domain adaptation models show much room for improvement. *How do the results of Arabic domain adaptation classifiers that utilise distant supervision to automatically label target data when using keyword sets from similar crisis events compare to those using keyword sets from different crisis events in classifying Arabic tweets from emerging events?*

As stated in row 2 from Tables 7.5 and 7.6, and as expected, the DS-TM results slightly decrease when using crisis data from different crisis types as the target data to create the keyword set. More specifically, we find that the number of the shared

TABLE 7.5: Experimental results in F1 score for 9 models tested on 5 crisis events from the same crisis type as the keywords set (with expanding the initial keyword list).

Model/Setting	S1	S2	S3	S4	S5
SL-LS	**0.753 (0.026)**	**0.768 (0.028)**	**0.798 (0.016)**	**0.746 (0.024)**	**0.717 (0.030)**
SSL-DS-TM	**0.833 (0.017)**	**0.831 (0.008)**	**0.822 (0.019)**	0.803 (0.016)	0.747 (0.025)
SSL-SelfL-TM	0.608 (0.098)	0.589 (0.062)	0.687 (0.074)	0.653 (0.087)	**0.757 (0.023)**
Standalone	0.806 (0.008)	0.776 (0.004)	0.712 (0.008)	0.728 (0.015)	0.680 (0.019)
SSL-DS-FX	0.683 (0.000)	0.618 (0.002)	0.804 (0.001)	0.708 (0.000)	0.754 (0.001)
SSL-SelfL-FX	0.784 (0.005)	0.584 (0.005)	0.647 (0.002)	**0.819 (0.000)**	0.679 (0.001)
SSL-DS-FT	0.628 (0.052)	0.635 (0.026)	0.803 (0.009)	0.725 (0.026)	0.754 (0.009)
SSL-SelfL-FT	0.795(0.015))	0.592(0.024)	0.625 (0.011)	0.802 (0.036)	0.670 (0.014)
SL-LT	**0.945 (0.006)**	**0.881 (0.013)**	**0.924 (0.010)**	**0.929 (0.007)**	**0.839 (0.009)**

Note. The upper limit and the best reported results are highlighted in bold.

top or expanded keywords occurring in the target data decreases. Evidently, when the number of shared keywords decreases, the performance of DS labelling method also declines. However, this is not the case in settings 1 and 10.

Our results are better in classifying the Cairo Explosion data when the Floods keyword set is used in place of the Explosion keyword set. This is because the number of the top Floods keywords exist in tweets related to Cairo Explosion event is higher than that of the top Explosion keywords (6 > 5). The high divergence level between the Cairo and Beirut Explosion data helps in producing such an outcome. For the Kuwait Floods event, the performance of DS-TM drops from 0.803 to 0.767 in F1 score. It is worth noting that the top keyword list changes from the previous list and does not include "حذر", which gives DS-TM an advantage in the previous section.

For the Covid'19 and Dragon Storm events, Table 7.6 shows that the results of DS-TM change when using different crisis types to build the keyword sets for Floods and Explosion. It seems that the framework with the Floods keyword set generates better pseudo-labelled data from Covid'19 and Dragon Storm than with the Explosion keyword set. This is definitely caused by the number of shared top or expanded keywords. The Dragon Storm data includes 6 top keywords and 55% of the expanded keywords from the Floods keyword set. On the other hand, only 2 top keywords and 16% of the expanded keywords are shared with the Explosion keyword set. The performance of our standalone model supports this finding: for

example, its F1 score for tweets related to Covid'19 in setting 6 is higher than in setting 7. This is because setting 6 uses the Floods keyword set, while setting 7 uses the Explosion keyword set.

Based on these observations, we can posit that Arabic tweets from an event of any crisis type can be used to generate keyword sets for any emerging disaster. However, the performance of DS-TM can be improved by using crisis data from the same or similar crisis type to establish the initial keyword list for the given emerging Arabic event.

How do the results of self-labelling compare to those of distant-supervised labelling with keyword sets from another crisis type of the target event when used in Arabic domain adaptation settings?

Using tweets from different crisis types to pre-train a model to classify target events presents several problems. The main issue is that keywords from related tweets in the source data can be remarkable keywords in the irrelevant target data. An example of this case is setting 9 in Table 7.6, where the Explosion crisis type included in the source data features terrorism-associated words due to the nature of bombings and explosions, while unrelated tweets from the Dragon Storm target event contain these words due to the crisis locations (Palestine and Syria), where people often post about terrorist acts. Using our framework to automatically label the Arabic target corpus – before merging with the manually labelled source tweets to build a reliable target model (DS-TM) – dramatically reduces this problem. DS-TM does not use models pre-trained on source data, and the DS labelling method labels the tweet as related and informative only if it contains two keywords from the expanded keyword set; it is rare to find two terrorist words in one tweet posted during the Dragon Storm crisis. Thus, the DS labelling method outperforms SelfL in the three adaptation methods.

Another issue is that the number of shared top or expanded keywords can be reduced when tweets from crisis events belonging to different crisis types to the target data are used to generate the keyword sets. This is the case in settings 11, 12, 13 and 14. Although this issue restricts the capacity of our DS labelling method to produce good pseudo-labelled data from the emerging disasters, the best reported domain

adaptation model for setting 11 is DS-TM. This is because of the divergence level between the source and target events, which leads SelfL to produce noisy self-labelled data related to the Beirut Explosion incident. In contrast, DS-TM does not outperform SelfL-FX for the Kuwait Floods event – even in setting 13 with the increased number of common top keywords. Nevertheless, this number is still too small (4 > 3) to change the performance of DS-TM. We also observe that DS-TM remains the best reported domain adaptation model for the Jordan Floods disaster in setting 12. Here, the length and content of the keyword set change when using incidents from another crisis type. Although the number decreases, the list becomes richer by including words with multiple shapes present in the Jordan Floods data: "انقاذ" and "كارثة". This is because of the powerful nature of the Arabic language in having multiple shapes on one root as discussed above. For setting 14 (Hafer-albatin), the number of common keywords decreases from 5 to 2, with no words with multiple shapes like "صوت". Thus, SelfL-TM produces the best results among the 6 domain adaptation models.

In general, DS-TM is the most robust tweet classifier among all the mentioned domain adaptation models. In all cases, it improves model performance after incorporating the pseudo-labelled data, unlike the alternatives.

How similar are the supervised classifiers' results trained on target Arabic generated labelled data to those of supervised classifiers trained solely on Arabic source data?

From columns 1 and 3 in Table 7.6 and rows 1 and 4 in Table 7.5, we see that in 4 out of 7 settings, the standalone model outperforms LS. For example, the F1 score changes from 0.753 to 0.806 and from 0.744 to 0.804 for the Cairo Explosion and Covid'19 incidents, respectively. Based on these results, we find that the standalone model is better than LS in classifying messages from upcoming current events if the keyword set is generated using similar events or events from similar crisis types and the similarity level between the target and the source data is low. This is not the case in Floods settings, where the F1 score shifts from 0.798 to 0.712 (Jordan Floods), from 0.746 to 0.728 (Kuwait Floods) and from 0.717 to 0.680 (Hafer-albatin Floods). This is because these events are similar to each other, which makes them similar to the source data where two of the Floods events are included in the training data.

Clearly, then, training the tweet classification model on automatically labelled

target data generated by our framework is better than training the tweet classification model on human labelled data with a low similarity level. This states the importance of the crisis and the language features in improving the performance of the classifiers.

Effect of using external knowledge base (Almaany)

As expected, the impact of excluding distant supervision from our framework for Arabic tweets is similar to the impact seen in Section 5.3.3 for English tweets. See Appendix C for more details (results of DS-A – DS without expanding the initial keyword list via Almaany).

However, this impact increases when excluding this step for Arabic tweets when using tweets from another crisis type to the target event to establish the keyword lists, as seen in the last 3 rows in Table 7.6. For 13 out of 14 settings, the F1 scores decrease for all the 3 models where we apply our framework without expanding the original keyword lists. For example, the performance drops from 0.846 to 0.715 (-13%) for Dragon Storm data when using the top 10 keyword list rather than the expanded one. Setting 11 is the only setting to increase the model performance after removing step three from our framework. F1 score raises from 0.771 to 0.788 (+1.7) for DS-TM. For this setting, less than 35% of the expanded Floods keyword list exist in Beirut Explosion data with 3 out of 10 top keywords ("خوف", "اخلاء" and "صور") .The word "صور does not have synonyms in Almaany.

[]

TABLE 7.6: Experimental results in F1 score for nine models tested on seven crisis events from different crisis types to the keyword sets in two settings (with and without expanding the initial keyword list).

Models/Settings	S6	S7	S8	S9	S10	S11	S12	S13	S14
SL-LS	0.744 (0.047)	0.744 (0.047)	0.658 (0.029)	0.658 (0.029)	0.753 (0.026)	0.768 (0.028)	0.798 (0.016)	0.746 (0.024)	0.717 (0.030)
SSL-DS-TM	**0.846 (0.029)**	**0.831 (0.025)**	**0.741 (0.013)**	**0.734 (0.019)**	**0.843 (0.032)**	**0.771 (0.022)**	**0.810 (0.019)**	0.767 (0.002)	0.737 (0.021)
SSL-SelfL-TM	0.741 (0.130)	0.741 (0.130)	0.560 (0.070)	0.560 (0.070)	0.608 (0.098)	0.589 (0.062)	0.687 (0.074)	0.653 (0.087)	**0.757 (0.023)**
Standalone	0.804 (0.025)	0.488 (0.031)	0.694 (0.010)	0.570 (0.020)	0.428 (0.309)	0.559 (0.054)	0.502 (0.071)	0.472 (0.026)	0.453 (0.028)
SSL-DS-FX	0.850 (0.000)	0.730 (0.000)	0.742 (0.000)	0.651 (0.006)	0.694 (0.006)	0.682 (0.002)	0.640 (0.000)	0.719 (0.000)	0.505 (0.000)
SSL-SelfL-FX	0.757 (0.002)	0.757 (0.002)	0.647 (0.002)	0.647 (0.002)	0.784 (0.005)	0.584 (0.005)	0.647 (0.002)	**0.819 (0.000)**	0.679 (0.001)
SSL-DS-FT	0.842 (0.008)	0.729 (0.063)	0.725 (0.011)	0.612 (0.039)	0.689 (0.019)	0.687 (0.009)	0.644 (0.018)	0.753 (0.014)	0.532 (0.028)
SSL-SelfL-FT	0.757 (0.024)	0.757 (0.024)	0.640 (0.016)	0.640 (0.016)	0.795 (0.015)	0.592 (0.024)	0.625 (0.011)	0.802 (0.036)	0.670 (0.014)
SL-LT	**0.954 (0.003)**	**0.954 (0.003)**	**0.852 (0.010)**	**0.852 (0.010)**	**0.945 (0.006)**	**0.881 (0.013)**	**0.924 (0.010)**	**0.929 (0.007)**	**0.839 (0.009)**
Without Expanding (i.e., without applying distant supervision via Almaany dictionary)									
SSL-DS-TM	0.715 (0.0380)	0.668 (0.041)	0.693 (0.014)	0.708 (0.025)	0.828 (0.024)	0.788 (0.014)	0.798 (0.012)	0.738 (0.018)	0.593 (0.042)
SSL-DS-FX	0.611 (0.004)	0.458 (0.000)	0.518 (0.005)	0.591 (0.000)	0.436 (0.000)	0.515 (0.002)	0.544 (0.000)	0.562 (0.000)	0.407 (0.000)
SSL-DS-FT	0.687 (0.085)	0.456 (0.041)	0.516 (0.0170)	0.600 (0.0160)	0.589 (0.130)	0.514 (0.0150)	0.497 (0.024)	0.566 (0.008)	0.408 (0.0020)

Note. E represents all available manually labelled earthquake crisis datasets excluding MPE. Best results for each test data are in bold.

7.4 Further Analysis

In domain adaptation settings, we want to know how many pseudo-labelled messages are required to create an accurate target classifier. Our objective is to identify how the performance of DS-TM changes as the amount of target data increases. In our opinion, this is a critical aspect in the practical use of our strategy in real-world settings. As a result, our last research question is: *How many pseudo-labelled Arabic target instances are required to develop a good model for classifying Arabic tweets from a current crisis event?*

7.4.1 Experiments

We repeat the experiments described in Section 6.4 for the 4 target sets with varied numbers of instances (50, 100, 250, 500, and 1,000) for each class – related/informative and irrelevant / not informative tweets – to see how performance varies as the amount of included target data changes.

7.4.2 Results and Discussion

We conduct these experiments using balanced datasets with varied numbers of samples from related/informative and irrelevant / not informative categories to address our last research question in this chapter. To do so, we compare the performance of DS-TM to that of other domain adaptation models as the number of pseudo-labelled target instances increases. Figure 7.3 presents the results in F1 scores. Finally, we estimate how many examples are required to train a reliable Arabic target model. *What is the minimum number of pseudo-labelled Arabic target instances needed to build a good model to classify Arabic tweets from an emerging crisis event?*

DS-TM outperforms other domain adaptation models in subplot (d) of Figure 7.3, regardless of the number of used tweets from the Beirut Explosion data. This supports our hypothesis that DS-TM performs better when the source and target sets are different. However, DS-TM is not necessarily the best recorded model for less diverse target sets, as in subplot (c) for the Covid'19 incident when the Floods keyword set is used. The F1 scores for DS-TM start at 0.873 in this case and do not vary significantly when more pseudo-labelled target data are added. This is because the target event

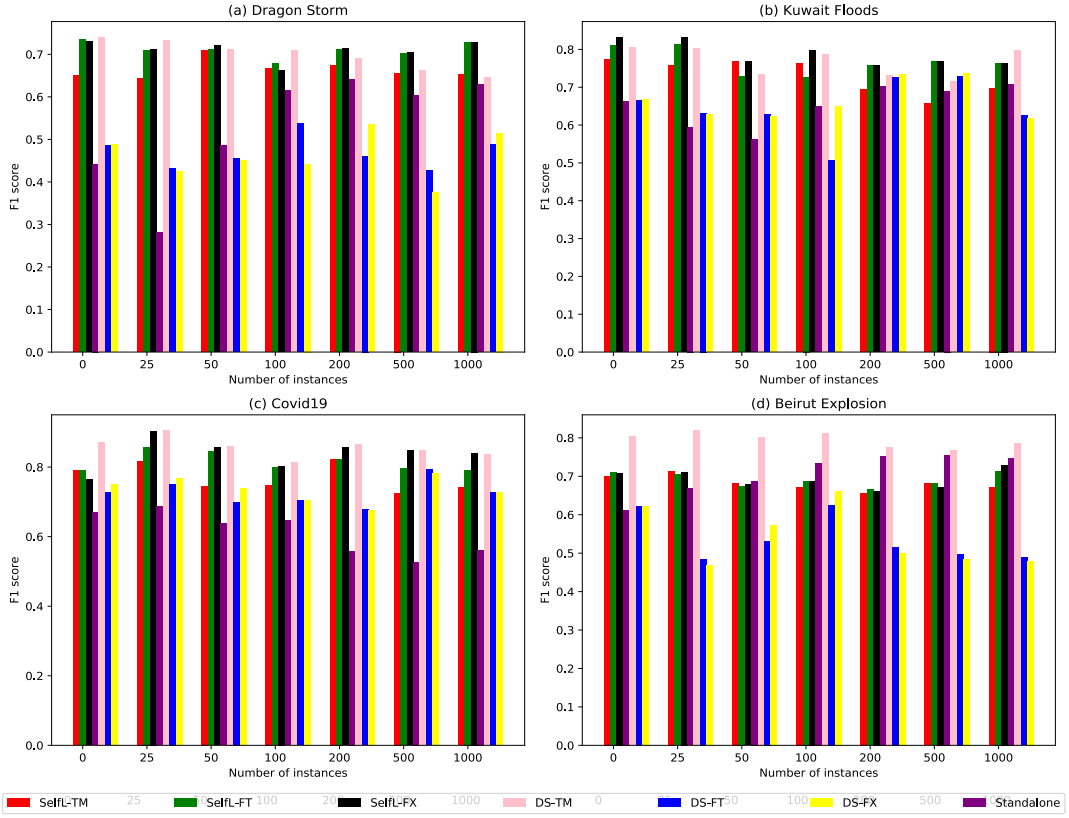

FIGURE 7.3: Results of Arabic domain adaptation models and standalone model in F1 score with varying amounts of incorporated (pseudo-labelled and self-labelled) target data.

differs considerably from the keyword set and source domain; however, it shares a language feature with more than one event in the source data and the keyword set where people use the same Arabic dialect in writing the posted tweets (Kuwaiti and Saudi). The crisis event in subplot (a), Dragon Storm, is considerably different from the keyword set and source domain. For the last case, the source data or the keyword set and the target domains share crisis features such as crisis type, proximal locations and infrastructural damage, as in subplot (b). We note that the DS-TM result for the Kuwait Floods disaster changes in line with the number of the incoming target tweets added to the training data. However, it is recorded as the best domain adaptation model when the added data reaches 2,000 tweets for each class.

The most significant finding here is that DS-TM consistently performs well at the start of the studies (number of instances = 50 per class). For all four settings, the F1 score of DS-TM is always higher than 0.80. This suggests that this domain adaptation model can be used to categorise unseen tweets from current events at the

start of a crisis.

In conclusion, our findings imply that DS-TM is a reliable model that may be applied in real-world scenarios. To properly identify tweets from current events, only 50 pseudo-labelled tweets from both related/informative and not related / not informative classes are required. However, not related / not informative instances can be extracted from tweets sent before the onset of the disaster, thus reducing the amount of time required to gather these examples. We believe that our technique is ideal for real-time crisis scenarios in which humanitarian organisations must make urgent judgments to assist individuals in need.

Statistical analysis

As shown in Table 7.7, the p-value for the factor 'model type' is 0.000000e+00 < 0.05, which rejects the null hypothesis that all seven model types have the same impact on average model performance. On its own, model type definitely generates differences in model performance when classifying tweets related to Covid-19. The p-value for the number of tweets and the combination of both are 1.252120e-71 < 0.05 and 1.216327e-121 < 0.05, respectively. This means that the number of tweets in isolation and the combination of number of tweets and model type constitute a statistical difference on average model performance. However, model type is the most statistically different compared to the remaining factors, as the p-value is the smallest. In this case, a Tukey test is required to determine which group of factors caused the significant difference in the results.

TABLE 7.7: Results of two-way ANOVA test for Arabic tweets.

	sum_sq	**Df**	**F**	**PR(>F)**
C(model type)	8.787645	6	849.192261	0.00E+00
C(number of tweets)	0.674109	6	65.142397	1.25E-71
C(model type, number of tweets)	1.506123	36	24.257315	1.22E-121
Residual	2.450808	1421	NaN	NaN

Model type

In Table 7.8, only DS-FT and DS-FX fail to reject the null hypothesis, indicating that the results of both model types are similar regardless of the number of tweets. All the other groups of model types show a statistical difference on model performance in classifying Arabic Covid'19 Twitter data. This reflects the importance of using the

right model type in the experiment to obtain desirable results.

TABLE 7.8: Model type, Multiple Comparison of Means - Tukey HSD, FWER = 0.05.

Group 1	Group 2	Meandiff	Lower	Upper	Reject
'DS-FT'	'DS-FX'	0.0156	-0.0006	0.0318	False
'DS-FT'	'DS-TM'	0.1312	0.115	0.1475	True
'DS-FT'	'SelfL-FT'	0.0876	0.0714	0.1038	True
'DS-FT'	'SelfL-FX'	0.1124	0.0962	0.1286	True
'DS-FT'	'SelfL-TM'	0.044	0.0278	0.0602	True
'DS-FT'	'Standalone'	-0.1143	-0.1305	-0.098	True
'DS-FX'	'DS-TM'	0.1157	0.0995	0.1319	True
DS-FX'	'SelfL-FT'	0.0721	0.0559	0.0883	True
'DS-FX'	'SelfL-FX'	0.0968	0.0806	0.1131	True
'DS-FX'	'SelfL-TM'	0.0285	0.0123	0.0447	True
'DS-FX'	'Standalone'	-0.1298	-0.146	-0.1136	True
'DS-TM'	'SelfL-FT'	-0.0436	-0.0598	-0.0274	True
'DS-TM'	'SelfL-FX'	-0.0188	-0.035	-0.0026	True
'DS-TM'	'SelfL-TM'	-0.0872	-0.1034	-0.071	True
'DS-TM'	'Standalone'	-0.2455	-0.2617	-0.2293	True
'SelfL-FT'	'SelfL-FX'	0.0248	0.0086	0.041	True
'SelfL-FT'	'SelfL-TM'	-0.0436	-0.0598	-0.0274	True
'SelfL-FT'	'Standalone'	-0.2019	-0.2181	-0.1857	True
'SelfL-FX'	'SelfL-TM'	-0.0684	-0.0846	-0.0522	True
'SelfL-FX'	'Standalone'	-0.2267	-0.2429	-0.2105	True
'SelfL-TM'	'Standalone'	-0.1583	-0.1745	-0.1421	True

Number of tweets

According to Table 7.9, most of the number groups that reject the null hypothesis contain 50 as the number of tweets used to run the experiment. This shows that 50 is the greatest number of tweets that has a significant difference on model performance. On review, we find that the F1 score increases from 25 to 50 and decreases from 50 to any other number. Clearly, then, 50 is the number of tweets required to build a good tweet classifier when testing on Covid'19 data.

Model type and number of tweets (top 30)

Standalone is the most model type that rejects the null hypothesis; however, the Tukey test does not reveal the best model based on model performance. On review, we can say that standalone does indeed have a significant difference to the other models with the same number of tweets, but it produces the worst results. Unlike standalone, DS-TM with 50 and SelfL-FX with 50 have a significant difference but with the best results. As shown in Table 7.10, the model type in DS-TM with 50 and

TABLE 7.9: Number of tweets, Multiple Comparison of Means - Tukey HSD, FWER = 0.05.

Group 1	Group 2	Meandiff	Lower	Upper	Reject
25	50	0.046	0.0191	0.0729	True
25	100	0.0088	-0.0181	0.0357	False
25	200	-0.0206	-0.0475	0.0063	False
25	500	-0.0121	-0.039	0.0148	False
25	1000	-0.0059	-0.0328	0.021	False
25	2000	-0.0204	-0.0473	0.0065	False
50	100	-0.0372	-0.0641	-0.0103	True
50	200	-0.0666	-0.0935	-0.0397	True
50	500	-0.0581	-0.085	-0.0312	True
50	1000	-0.0519	-0.0788	-0.025	True
50	2000	-0.0664	-0.0932	-0.0395	True
100	200	-0.0294	-0.0563	-0.0025	True
100	500	-0.0209	-0.0477	0.006	False
100	1000	-0.0147	-0.0416	0.0122	False
100	2000	-0.0291	-0.056	-0.0022	True
200	500	0.0085	-0.0184	0.0354	False
200	1000	0.0147	-0.0122	0.0416	False
200	2000	0.0002	-0.0266	0.0271	False
500	1000	0.0062	-0.0207	0.0331	False
500	2000	-0.0083	-0.0352	0.0186	False
1000	2000	-0.0145	-0.0414	0.0124	False

SelfL-FX with 50 rejects the null hypothesis (36 and 37) more than the number of tweets (9 and 8).

7.5 Conclusion

This chapter is the first attempt at a domain adaptation approach for Arabic tweet classification for crisis response using an adaptive distant supervision-based framework to label the unlabelled Arabic target tweets. Our Arabic framework follows the English version in providing a new set of keywords rather than the ones extracted from other events, thereby helping add new features to the training data. This step minimises the gap between source and target domains caused by the domain shift of different crisis events.

In this chapter, we experiment using keyword sets from the same crisis types of the target event. Results show that building a target model using the pseudo-labelled target domain data generated by DS and the available source domain data

TABLE 7.10: Results of Tukey test for (model type and number of tweets) factor.

	Reject 1	**Reject 2**	**total_sum**
500 /'Standalone'	0.0	46.0	46.0
1000 / 'Standalone'	33.0	13.0	46.0
2000 / 'Standalone'	20.0	26.0	46.0
50 /'SelfL-FX'	8.0	37.0	45.0
50 /'DS-TM'	9.0	36.0	45.0
100 /'Standalone'	38.0	6.0	44.0
200 / 'Standalone'	24.0	19.0	43.0
500 / 'DS-FX'	5.0	35.0	40.0
500 / 'DS-FT'	5.0	35.0	40.0
25 / 'Standalone'	11.0	29.0	40.0
50 / 'Standalone'	5.0	32.0	37.0
100 / 'DS-FT'	37.0	0.0	37.0
25 / 'DS-TM'	13.0	23.0	36.0
200 / 'DS-FT'	25.0	11.0	36.0
200 / 'DS-FX'	25.0	11.0	36.0
500 / 'SelfL-FX'	1.0	34.0	35.0
100 / 'DS-TM'	33.0	2.0	35.0
100 / 'SelfL-FX'	33.0	2.0	35.0
50 / 'SelfL-FT'	5.0	30.0	35.0
1000 / 'DS-TM'	28.0	6.0	34.0

always improves model performance (average of 3.7% absolute gain in F1 score). It is also reported the best domain adaptation model on 3 out of 5 datasets.

We also experiment using keyword sets from different crisis types of the target event. The results show that our framework's labelling method (DS) performs better than self-labelling (SelfL) in three different adaptation methods. It always improves the model performance (average of 5.5% absolute gain in F1 score). It is also reported the best domain adaptation model on 7 out of 9 settings. As with the English approach, we conduct further analysis to determine the amount of pseudo-labelled target data needed to build a robust target classifier. The results show that 50 tweets from each class (related and informative and not related or not informative) are sufficient. This outcome perfectly suits our task because it requires only a short time at event onset. Unlike the other methods, it can be considered a general approach without the need to predefine the similarity between the source and target domains. In addition, our framework proves that it can be adopted in any language, even those with limited resources.

Besides the limitations stated in previous chapters, the adaptive method in this

chapter presents its own drawbacks. One is the need to address the problems caused by the nature of the language used in the tweets. In Arabic, in particular, signs are vital because a word with different signs has different meanings. For example, our data show that the word "فجر" has two meanings according to the used sign: while "فُجر" means explosion, "فَجر" means the sunrise. To solve this issue, we can add tweets that contain the undesirable word to the irrelevant or not informative class. In doing so, the classification model with the absence of word signs can learn the meaning of the shaped words from the context of the tweets. For future work, we can use keyword sets from both the same and different crisis types to the target event to increase the accuracy of our automated labelling process. We would also look to use our framework for tweets in other languages, such as French and Spanish.

Chapter 8

Conclusion

In Chapter 1, we addressed the context in which this research is valuable: primarily by classifying tweets to simplify and improve the work of humanitarian organisations to make quick and correct key decisions in the name of helping people in need during crises. Our proposed method is able to build a good and reliable classifier using a relatively low number of tweets (50 relevant tweets) from the current events, which gives the opportunity for organisations to respond in short time (minutes or hours) after the event onset. We discussed the relevant research topics in Chapter 2: tweet classification for crisis response, distant supervision and transfer learning, including domain adaptation applications. We outlined the experimental setup used in this thesis, including the datasets, classification models and evaluation metrics, in Chapter 4.

Section 8.1 outlines the contributions of this thesis in response to the research questions introduced in Chapter 1. Finally, Section 8.2 discusses some future research directions.

8.1 Summary of Contributions

We proposed a simple yet effective distant supervision-based framework to automatically label tweets from crisis events to improve the performance of tweet classification models. Our main contributions are summarised below.

Chapter 3 represented the first attempt in the field of TCFCR to search for the best neural network and word embedding to build a good tweet classifier. The main goal was to find the best model to start our research. We achieved this goal by

investigating the effect of using domain-specific and general word embeddings with two deep learning architectures: BiLSTM and CNN. We showed that BiLSTM with GloVe produced the highest F1 score among the other classifiers. As a result, we used this model in conducting experiments for the rest of our contributions (for English) in this thesis.

Chapter 5 introduced a novel framework to answer the first research question from Chapter 1. The framework was utilised to produce automatically labelled data from new events to be used in addition to the available human-labelled data in training crisis-related classifiers. The main objective here was to solve the problem of the lack of labelled data as a means to enhance the performance of the tweet classification models. We showed that our framework can produce good-quality automatically labelled training data compared to the human-labelled data. Furthermore, automatically labelled training data may be substituted for a proportion of the manually labelled training data with little effect on model performance, demonstrating that automatically labelled data can be utilised when hand-labelled data are unavailable.

In Chapter 6, to respond to the second research question presented in Chapter 1, we combined our framework from Chapter 5 with the adaptation method to automatically label unseen tweets from emerging incidents to be adopted in the training phase. The main goal here was to minimise the gaps between source and target domains by using common crisis-type keywords. These keywords were then expanded to include new linguistically similar keywords from an external knowledge base. Our framework achieved this goal by identifying tweets with these new keywords, which brought target-specific features into the training data. Our two-stage approach – labelling and adaptation – boosts the tweet classifiers' performance by a value ranges from 0.1% to 11.4% and outperformed the state-of-art domain adaption method for TCFCR (iterative self-training). It is also suitable for practical real-time situations because it only requires 50 automatically labelled tweets to perform well.

Chapter 7 adapted the domain adaptation method introduced in Chapter 6 to automatically labelled Arabic tweets rather than English from the current disasters. This was achieved by changing some of the details in the pseudo-labelling stage from the labelling method. Our goal was to overcome the issues of low-resource

languages in applying solutions to domain shifts between source and target data. We accomplished this and answered our third research question by using clusters instead of manually labelled tweets and using an Arabic-specific resource, Almaany, to extend the initial keyword list. Results showed that our adaptive method always improves the model performance (average of 3.7% absolute gain in F1 score).We also ran experiments to use keyword sets from different crisis types to the target incident. As a result, we found that our framework can classify unseen tweets from a given disaster using a keyword set from different disasters. Results showed that adoptive DS-TM always improves model performance (average of 5.5% absolute gain in F1 score). To this end, we can say that our original and adaptive domain adaptation approaches represent robust models to classify tweets from emerging events, even for languages with limited resources. We hope that leveraging automatically labelled data will accelerate the research on classifying Arabic tweets in crisis response.

We also noted that languages differ in many aspects, and the distribution of the classes in the corpora or the dataset is not similar. For example, most of the tweets posted in Arabic, unlike in English, reflect political opinions after a crisis or advertising for a product or service, which increases the number of irrelevant or not informative tweets. In addition, selecting training data is an essential step in training crisis-related classifiers. The source data should be similar to the target data in terms of crisis type [132]. However, the different dialects that come with languages like Arabic should be considered when choosing the data to train crisis-related classifiers. Moreover, the availability of resources varies from one language to another, and language-based specific word embeddings and external resources are lacking for crisis response in Arabic. Ultimately, the nature of the language plays a vital role in detecting tweets during a disaster.

8.2 Future Work

In Chapter 5, we defined the areas where our framework is not applicable for English tweets, including rare events like Volcano incidents. For these events, labelled data are not available, which restricts the application of our original framework. An interesting future approach would be to use our adaptive framework, which utilises

clustering on unlabelled data and treats these events as events from low-resource languages.

We also highlighted our work's limitations, including weak keywords in the labelling method without any restrictions, which introduces noise to the automatically labelled data generated by our framework. We believe that adding some constraints in using such keywords can improve the labelling process. The noise caused when using the distant supervision technique in Chapter 6 can also be reduced by applying co-training, tri-training or active learning to domain adaptation models that use DS to generate pseudo-labels from target data using unseen tweets and choose the agreed labels.

Our work in this thesis presents other future research directions. Expanding our methods to include multi-class classification tasks for information types could be helpful. The authors in [131] use our framework presented in Chapter 5 as a baseline in their experiments. They have shown the effectiveness of our proposed distant supervision method in classifying tweets in different categories based on their information types. Our proposed methods can also be applied to tweet classification for other purposes, such as identifying tweets from terrorist events or cyberbullying activities. These topics can have unique keywords that express such events or actions. They can also be used to classify texts rather than tweets, such as emails, text messages, medical texts, resident feedback texts, and other complex people-generated texts. We believe that tweets share features with these ill-formed texts, which points to the potential of our methods to identify specific events, behaviors or feelings expressed on these communication platforms. In addition, the proposed methods in this thesis can be tested to address NLP problems besides tweet classification, including sentiment analysis. In our opinion, this could be achieved by combining two keyword sets: one for given topics and the other for sentiment expressions. Although the adaptive domain adaptation method introduced in Chapter 7 aimed to improve the performance of classifying Arabic tweets from emerging events, this method is flexible enough to be extended to other low-resource languages like Spanish.

We have also recognised several future works in tweet classification for crisis response. Several problems can be addressed in the future for classification models,

such as misinformation labelling, metadata extraction, onsite event detection and cross-language models. For training data, more approaches can be considered to include crisis images and crisis video in the training data as well as crisis tweet texts; in our opinion, this could be used in multi-source models to improve situational awareness during crises. For transfer learning, zero-shot learning or few-shot learning can be applied to solve the problem of classifying unseen crisis types. Finally, building a language model for crisis response could be of great value in enhancing tweet classifier performance.

Appendix A

Initial and Expanded Keyword Lists

Table A.1 gives examples of the initial and the final keyword lists used in our distant-based framework mentioned in Chapter 5 to pseudo-label tweets from new or emerging crisis events in Chapters 5, 6, and 7. Note that the external resource for English tweets is FrameNet and Almaany dictionary for the Arabic tweets.

[

TABLE A.1: Examples of original keyword list contains 10 strong keywords and the expanded keyword list which uses these important keywords as seeds in external resources to raise the number of words on the final list.

Keyword sets	**Initial keyword list**	**Expanded keyword list**	**Target event**
Arabic Floods events (Kuwait and Hafer-albatin)	مطر ، غرق ،اخلاء، سيل ، بحر ، طرق، الطقس ، قرب ، صور ، حفظ.	مطر ، رذاذ ،غرق ،غطس ،غاص ،غط ،انغمس ،رسب ، ،اخلاء ،ترحيل ،اجلاء ، جلاء ،تفريغ ،افراغ ،اخراج ،ابعاد ،اطلاق ، طرد ، حدث ،اتى ،تم ، وقع ،حصل ،نشا ، نتج، بدا ،نجم ،انتصب ،انبثق ،صدر ،نبت ، سيل، تدفق ،طوفان ،فيضان ،مسيل ،انصباب ،انهمار ،تدفق ،جريان ، سيلان ، بحر ،يم ،غمر ،نوفل ،طرق ،اتجاه ،جهة ، وجهة ،درب، سبيل ،منهج ،الطقس ،جو ،قرب ،دنو ،اقتراب ،كثب، محاذاة ،تلامس ،مجاورة ، اتصال ، صور ، حفظ ،تخزين ،حماية ،حراسة ،وقاية ، صيانة ،منع ، رقابة ،حجز ،توفير ،دافع ، رصد ، اعتنى	Jordan Floods

Continuation of Table A.1

Keyword sets	**Initial keyword list**	**Expanded keyword list**	**Target event**
English Floods events except Colorado Floods	'flood', 'bigwet', 'hit', 'help', 'affect', 'relief', 'kill', 'dead', 'people', 'water'	'flood', 'crowd' 'flock', 'hail', 'parade', 'pelt', 'pour', 'rain', 'roll', 'shower', 'stream', 'swarm', 'teem', 'throng', 'troop', 'bigwet', 'hit', 'attract', 'cast', 'catapult', 'chuck', 'drag', 'draw', 'drive', 'drop', 'fling', 'force', 'haul', 'hurl', 'impel', 'jerk', 'knock', 'launch', 'lift', 'move', 'nudge', 'pitch', 'press', 'propel', 'pull', 'punt', 'push', 'rake', 'roll', 'run', 'scoot', 'shove', 'slam', 'slide', 'stick', 'throw', 'thrust', 'toss', 'transfer', 'tug', 'wrench', 'wrest', 'yank', 'help', 'abet', 'aid', 'assist', 'assistance', 'cater', 'help out', 'helpful', 'serve', 'succor', 'affect', 'effect', 'impact', 'influence', 'power', 'relief', 'kill', 'annihilate', 'annihilation', 'asphyxiate', 'assassin', 'assassinate', 'assassination', 'behead', 'beheading', 'blood bath', 'bloodshed', 'butcher', 'butchery', 'carnage', 'crucifixion', 'crucify', 'deadly', 'decapitate', 'decapitation', 'destroy', 'dispatch', 'do in', 'drown', 'eliminate', 'euthanasia', 'euthanize', 'exterminate', 'extermination', 'fatal', 'fatality', 'fratricide', 'garrote', 'genocide', 'holocaust', 'homicide', 'immolation', 'infanticide', 'killer',	Colorado Floods

Continuation of Table A.1

Keyword sets	Initial keyword list	Expanded keyword list	Target event
		'killing', 'lethal', 'liquidate', 'liquidation', 'liquidator', 'lynch', 'massacre', 'massacre', 'matricide', 'murder', 'murderer', 'patricide', 'pogrom', 'regicide', 'shooting', 'silence', 'slaughter', 'slaughter', 'slaughterer', 'slay', 'slayer', 'slaying', 'smother', 'smothering', 'starve', 'suffocate', 'suffocation', 'suicide', 'suicide', 'take someone life', 'take out', 'terminate','dead', 'alive', 'deceased', 'dirt nap', 'late', 'life', 'lifeless', 'live', 'living', 'nonliving', 'undead', 'undead', 'death', 'asphyxiate', 'croak it', 'croak', 'decease', 'demise', 'die', 'drown', 'end', 'expire', 'kick the bucket', 'mortal', 'mortality', 'pass away', 'perish', 'starvation', 'starve', 'suffocate', 'suffocation', 'terminator', 'toll', 'people', 'water', 'clammy', 'damp', 'dewy', 'drenched', 'humid', 'moist', 'moistened', 'saturated', 'soaked', 'soaking', 'sodden', 'soggy', 'sopping', 'sweaty', 'waterlogged', 'wet'	

Continuation of Table A.1

Keyword sets	Initial keyword list	Expanded keyword list	Target event
Arabic Explosion events (Beirut and Cairo)	سبب، فجر،انقذ ،عدد، تعد،اسعف، جثة، طوارئ، كارثة	اوجد ، سبب ،أوجد ،أفضى الى، احدث، أحدث، أنتج، انتج ،لفجر فجر ،تدفق ،أراق ،أنبثق ،خرج ، دفق ، سال ، فاض ،نبع ،هراق ، جمهر ،راع ،كثر ،عدد ،جمهور، رعاع، عام ،كثره ، اسعف ،أسعف ،إسعاف ،انقذ ، انقاذ ،إنقاذ، اغاث ،أغاث ،اسعف ،أسعف ،عون ،غوث إتعد ،تأهب ،تاهب، تحضر ،تجهز ، تهيأ ،اعد ،أعد ،تحمس ،نشط ،تشمر ،حشد ،واقعة ،نكب ،نكبة ،كرب ،خطب ، ابتلأ،ابتلا، أصاب ،اصاب ،حدث ، كارثة ،ابتلاء،إبتلاء مصيبة ،ازمة ،حادثة ،حادث ، إمتحن ،امتحن ،محنة ،أجاح ،اجاح ،جائحة ،وقع ، هلاك ،جثة ،طوارئ	Covid’19

Continuation of Table A.1

Keyword sets	**Initial keyword list**	**Expanded keyword list**	**Target event**
English Typhoon events except Hagypit Typhoon	'typhoon', 'hurricane', 'tornado'.,'hit', 'help', 'pray', 'victim', 'donate', 'affect', 'relief'	'typhoon','hurricane','tornado','hit', 'attract', 'cast', 'catapult', 'chuck', 'drag', 'draw', 'drive', 'drop', 'fling', 'force', 'haul', 'hurl', 'impel', 'jerk', 'knock', 'launch', 'lift', 'move', 'nudge', 'pitch', 'press', 'propel', 'pull', 'punt', 'push', 'rake', 'roll', 'run', 'scoot', 'shove', 'slam', 'slide', 'stick', 'throw', 'thrust', 'toss', 'transfer', 'tug', 'wrench', 'wrest', 'yank', 'help', 'abet', 'aid', 'assist', 'assistance', 'cater', 'help out', 'helpful', 'serve', 'succor', 'pray', 'anoint', 'baptism', 'baptize', 'bar mitzvah', 'bless', 'blessing', 'christen', 'christening', 'circumcise', 'circumcision', 'communion', 'confession', 'confirm', 'confirmation', 'consecrate', 'consecration', 'eucharist', 'evensong', 'exercise', 'initiate', 'initiation', 'mass', 'ordain', 'order', 'ordination', 'prayer', 'rite of passage', 'rite', 'ritual', 'sacrament', 'sacrifice', 'sacrifice', 'service', 'unction', 'vesper', 'vigil', 'worship','victim', 'accident', 'apocalypse', 'befall', 'betide', 'calamitous', 'calamity', 'casualty', 'cataclysm', 'catastrophe', 'catastrophic', 'crisis', 'debacle', 'disaster', 'disastrous', 'incident','mischance', 'misfortune', 'mishap', 'suffer', 'tragedy', 'donate', 'advance', 'bequeath' 'bequest', 'charity', 'confer upon', 'contribute', 'contribution', 'donate', 'donation', 'donor', 'endow', 'fob off', 'foist', 'gift', 'give out', 'give', 'hand in', 'hand out', 'hand over', 'hand', 'leave', 'pass out', 'pass', 'treat', 'volunteer', 'will', 'affect', 'effect', 'impact', 'influence', 'power', 'relief'	Hagypit Typhoon

Appendix B

Statistical analysis of results of DA for English tweets

Tables B.1, B.2 , B.3 and B.4 show the results for two-way ANOVA test for English tweets.

TABLE B.1: Results of two-way ANOVA test for English tweets.

	sum_sq	**Df**	**F**	**PR(>F)**
C(model type)	7.570986	6	891.628465	0.00E+00
C(number of tweets)	0.093237	5	13.176558	1.61E-12
C(model type, number of tweets)	2.475226	30	58.301049	2.67E-211
Residual	1.723711	1218	NaN	NaN

TABLE B.2: Number of tweets, Multiple Comparison of Means - Tukey HSD, FWER = 0.05.

Group 1	**Group 2**	**Meandiff**	**Lower**	**Upper**	**Reject**
25	50	0.0071	-0.0199	0.034	False
25	100	0.0157	-0.0113	0.0427	False
25	200	0.016	-0.011	0.043	False
25	500	0.0188	-0.0082	0.0458	False
25	1000	0.027	0.0001	0.054	True
50	100	0.0086	-0.0183	0.0356	False
50	200	0.0089	-0.0181	0.0359	False
50	500	0.0118	-0.0152	0.0387	False
50	1000	0.02	-0.007	0.047	False
100	200	0.0003	-0.0267	0.0273	False
100	500	0.0031	-0.0239	0.0301	False
100	1000	0.0113	-0.0156	0.0383	False
200	500	0.0028	-0.0242	0.0298	False
200	1000	0.0111	-0.0159	0.038	False
500	1000	0.0082	-0.0188	0.0352	False

TABLE B.3: Model type, Multiple Comparison of Means - Tukey HSD,FWER = 0.05.

Group 1	Group 2	Meandiff	Lower	Upper	Reject
'DS-FT'	'DS-FX'	0.0181	-0.0001	0.0363	False
'DS-FT'	'DS-TM'	0.0836	0.0653	0.1018	True
'DS-FT'	'SelfL-FT'	-0.129	-0.1473	-0.1108	True
'DS-FT'	'SelfL-FX'	-0.109	-0.1272	-0.0907	True
'DS-FT'	'SelfL-TM'	-0.1156	-0.1339	-0.0974	True
'DS-FT'	'Standalone'	-0.1053	-0.1235	-0.0871	True
'DS-FX'	'DS-TM'	0.0654	0.0472	0.0837	True
'DS-FX'	'SelfL-FT'	-0.1472	-0.1654	-0.1289	True
'DS-FX'	'SelfL-FX'	-0.1271	-0.1453	-0.1089	True
'DS-FX'	'SelfL-TM'	-0.1337	-0.152	-0.1155	True
'DS-FX'	'Standalone'	-0.1234	-0.1416	-0.1052	True
'DS-TM'	'SelfL-FT'	-0.2126	-0.2308	-0.1944	True
'DS-TM'	'SelfL-FX'	-0.1925	-0.2107	-0.1743	True
'DS-TM'	'SelfL-TM'	-0.1992	-0.2174	-0.181	True
'DS-TM'	'Standalone'	-0.1889	-0.2071	-0.1707	True
'SelfL-FT'	'SelfL-FX'	0.0201	0.0019	0.0383	True
'SelfL-FT'	'SelfL-TM'	0.0134	-0.0048	0.0316	False
'SelfL-FT'	'Standalone'	0.0237	0.0055	0.0419	True
'SelfL-FX'	'SelfL-TM'	-0.0067	-0.0249	0.0115	False
'SelfL-FX'	'Standalone'	0.0037	-0.0146	0.0219	False
'SelfL-TM'	'Standalone'	0.0103	-0.0079	0.0285	False

TABLE B.4: Results of Tukey test for (model type and number of tweets) factor.

	Reject 1	Reject 2	total_sum
100 / 'Standalone'	35.0	6.0	41.0
100 / 'SelfL-FT'	37.0	3.0	40.0
50 / 'Standalone'	7.0	33.0	40.0
50 / 'DS-TM'	11.0	25.0	36.0
25 / 'DS-FT'	16.0	19.0	35.0
200 / 'DS-FT'	22.0	13.0	35.0
1000 / 'DS-FT'	28.0	7.0	35.0
500 / 'DS-TM'	4.0	30.0	34.0
200 / 'DS-FX'	23.0	11.0	34.0
500 / 'Standalone'	0.0	34.0	34.0
1000 / 'DS-FX'	27.0	7.0	34.0
100 / 'DS-TM'	34.0	0.0	34.0
100 / 'DS-FT'	33.0	0.0	33.0
1000 / 'Standalone'	22.0	11.0	33.0
200 / 'DS-TM'	21.0	12.0	33.0
500 / 'DS-FX'	4.0	29.0	33.0
500 / 'DS-FT'	4.0	29.0	33.0
100 / 'DS-FX'	33.0	0.0	33.0
25 / 'DS-TM'	15.0	18.0	33.0
50 / 'DS-FX'	11.0	22.0	33.0

Appendix C

Results of DS-A for Arabic Tweets

Table C.1 shows the results of DS-A – DS without expanding the initial keyword list via Almaany when the keyword set related to the same crisis type to the target event. Refer to Table 7.4 for the settings and Table 7.5 for results of DS, with expanding the keyword list.

It is clear from the results in Tables 7.5 and C.1 that using our framework without distant supervision drops the performance for all the DS models except for setting 5. This setting represents Hafer-albatin Floods event data, which has 5 out of 10 top Floods keywords including powerful ones ("مطر" and "غرق"). In addition, less than 39% of the expanded keyword list is shared with the target data.

TABLE C.1: Experimental results in F1 score for DS-A models tested on 5 crisis events from the same crisis type as the keywords set (without expanding the initial keyword list).

Model/Setting	S1	S2	S3	S4	S5
SSL-DS-TM	0.738 (0.025)	0.788(0.021)	0.811 (0.012)	0.613 (0.020)	0.751 (0.015)
SSL-DS-FX	0.591 (0.000)	0.594 (0.000)	0.654 (0.000)	0.647 (0.009)	0.731 (0.000)
SSL-DS-FT	0.600 (0.0160)	0.587 (0.0130)	0.679 (0.034)	0.612 (0.072)	0.729 (0.029)

Bibliography

[1] Diab Abuaiadah. "Using bisect k-means clustering technique in the analysis of Arabic documents". In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 15.3 (2016), pp. 1–13.

[2] Diab Abuaiadah, Dileep Rajendran, and Mustafa Jarrar. "Clustering Arabic tweets for sentiment analysis". In: *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2017, pp. 449–456.

[3] Ghadah Adel and Yuping Wang. "Arabic twitter corpus for crisis response messages classification". In: *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*. 2019, pp. 498–503.

[4] Ghadah Adel and Yuping Wang. "Detecting and Classifying Humanitarian Crisis in Arabic Tweets". In: *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE. 2020, pp. 269–274.

[5] Waleed Alabbas, Haider M al-Khateeb, Ali Mansour, Gregory Epiphaniou, and Ingo Frommholz. "Classification of colloquial Arabic tweets in real-time to detect high-risk floods". In: *2017 International Conference On Social Media, Wearable And Web Analytics (Social Media)*. IEEE. 2017, pp. 1–8.

[6] Firoj Alam, Shafiq Joty, and Muhammad Imran. "Domain adaptation with adversarial training and graph embeddings". In: *arXiv preprint arXiv:1805.05151* (2018).

[7] Firoj Alam, Ferda Ofli, and Muhammad Imran. "Crisismmd: Multimodal twitter datasets from natural disasters". In: *Twelfth international AAAI conference on web and social media*. 2018.

[8] Alaa Alharbi and Mark Lee. "Crisis detection from Arabic tweets". In: *Proceedings of the 3rd workshop on arabic corpus linguistics*. 2019, pp. 72–79.

[9] Alaa Alharbi and Mark Lee. "Kawarith: an Arabic Twitter corpus for crisis events". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. 2021, pp. 42–52.

[10] James Allan, Ron Papka, and Victor Lavrenko. "On-line new event detection and tracking". In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, pp. 37–45.

[11] Almaany.com, ed. *Almaany Dictionary*. https://www.almaany.com/, last accessed 2022-06-23.

[12] Tahani Almanie, Alanoud Aldayel, Ghaida Alkanhal, Lama Alesmail, Manal Almutlaq, and Ruba Althunayan. "Saudi mood: a real-time informative tool for visualizing emotions in Saudi Arabia using twitter". In: *2018 21st Saudi Computer Society National Computer Conference (NCC)*. IEEE. 2018, pp. 1–6.

[13] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.

[14] Saed Alqaraleh and Merve Işik. "Efficient Turkish tweet classification system for crisis response". In: *Turkish Journal of Electrical Engineering & Computer Sciences* 28.6 (2020), pp. 3168–3182.

[15] Mohammed Alqmase, Husni Al-Muhtaseb, and Habib Rabaan. "Sports-fanaticism formalism for sentiment analysis in Arabic text". In: *Social Network Analysis and Mining* 11.1 (2021), pp. 1–24.

[16] Reem ALRashdi and Simon O'Keefe. "Deep learning and word embeddings for tweet classification for crisis response". In: *arXiv preprint arXiv:1903.11024* (2019).

[17] Reem Alrashdi and Simon O'Keefe. "Automatic labeling of tweets for crisis response using distant supervision". In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 418–425.

[18] Reem ALRashdi and Simon O'Keefe. "Robust Domain Adaptation Approach for Tweet Classification for Crisis Response". In: *Innovation in Information Systems and Technologies to Support Learning Research: Proceedings of EMENA-ISTL 2019* 7 (2019), p. 124.

[19] Meshrif Alruily. "Classification of Arabic Tweets: A Review". In: *Electronics* 10.10 (2021), p. 1143.

[20] Samah M Alzanin and Aqil M Azmi. "Rumor detection in Arabic tweets using semi-supervised and unsupervised expectation–maximization". In: *Knowledge-Based Systems* 185 (2019), p. 104945.

[21] Abdullah Aref, Rana Husni Al Mahmoud, Khaled Taha, Mahmoud Al-Sharif, et al. "Hate speech detection of arabic short text". In: *Comput. Sci. Inf. Technol* (2020), pp. 81–94.

[22] Mohammed Matuq Ashi, Muazzam Ahmed Siddiqui, and Farrukh Nadeem. "Pre-trained word embeddings for Arabic aspect-based sentiment analysis of airline tweets". In: *International Conference on Advanced Intelligent Systems and Informatics*. Springer. 2018, pp. 241–251.

[23] B Athira, Josette Jones, Sumam Mary Idicula, Anand Kulanthaivel, and Enming Zhang. "Annotating and detecting topics in social media forum and modelling the annotation to derive directions-a case study". In: *Journal of Big Data* 8.1 (2021), pp. 1–23.

[24] Mahmoud Al-Ayyoub, Safa Bani Essa, and Izzat Alsmadi. "Lexicon-based sentiment analysis of arabic tweets". In: *International Journal of Social Network Mining* 2.2 (2015), pp. 101–114.

[25] Collin F Baker, Charles J Fillmore, and John B Lowe. "The berkeley framenet project". In: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. 1998.

[26] Ranjan Kumar Behera, Monalisa Jena, Santanu Kumar Rath, and Sanjay Misra. "Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data". In: *Information Processing & Management* 58.1 (2021), p. 102435.

[27] EBK Bholowalia. "Means: A Clustering Technique based on Elbow Method and K-Means in WSN". In: *Int. J. Comput. Appl* 105 (2014), p. 17.

[28] Aritz Bilbao-Jayo and Aitor Almeida. "Automatic political discourse analysis with multi-scale convolutional neural networks and contextual data". In: *International Journal of Distributed Sensor Networks* 14.11 (2018), p. 1550147718811827.

[29] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information". In: *Transactions of the association for computational linguistics* 5 (2017), pp. 135–146.

[30] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge". In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008, pp. 1247–1250.

[31] 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 15th ISCRAM Conference*. Rochester, NY, USA, 20-23 May 2018.

[32] Grégoire Burel, Hassan Saif, Miriam Fernandez, and Harith Alani. "On semantics and deep learning for event detection in crisis situations". In: *Workshop on Semantic Deep Learning (SemDeep)*. ESWC 2017. Portoroz, Slovenia, 28 May 2017.

[33] Kai Cao, Xiang Li, and Ralph Grishman. "Improving event detection with dependency regularization". In: *Proceedings of the International Conference Recent Advances in Natural Language Processing*. 2015, pp. 78–83.

[34] 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.

[35] Victor Diogho Heuer de Carvalho, Thyago Celso Cavalcante Nepomuceno, and Ana Paula Cabral Seixas Costa. "An Automated Corpus Annotation Experiment in Brazilian Portuguese for Sentiment Analysis in Public Security". In: *International Conference on Decision Support System Technology*. Springer. 2020, pp. 99–111.

[36] Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. "Automatically labeled data generation for large scale event extraction". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017, pp. 409–419.

[37] David Chiang, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef. "Parsing arabic dialects". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006, pp. 369–376.

[38] Seong Eun Cho, Kyujin Jung, and Han Woo Park. "Social media use during Japan's 2011 earthquake: how Twitter transforms the locus of crisis communication". In: *Media International Australia* 149.1 (2013), pp. 28–40.

[39] Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. "On identifying hashtags in disaster twitter data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01. 2020, pp. 498–506.

[40] Kenneth Church and Patrick Hanks. "Word association norms, mutual information, and lexicography". In: *Computational linguistics* 16.1 (1990), pp. 22–29.

[41] Alfredo Cobo, Denis Parra, and Jaime Navón. "Identifying relevant messages in a twitter-based citizen channel for natural disaster situations". In: *Proceedings of the 24th international conference on world wide web*. 2015, pp. 1189–1194.

[42] Mark Craven, Johan Kumlien, et al. "Constructing biological knowledge bases by extracting information from text sources." In: *ISMB*. Vol. 1999. 1999, pp. 77–86.

[43] Stefano Cresci, Maurizio Tesconi, Andrea Cimino, and Felice Dell'Orletta. "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.

[44] Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. "Detecting perceived emotions in hurricane disasters". In: *arXiv preprint arXiv:2004.14299* (2020).

[45] Lucas Satria Aji Dharma and Edi Winarko. "Classifying Natural Disaster Tweet using a Convolutional Neural Network and BERT Embedding". In: *2022 2nd International Conference on Information Technology and Education (ICIT&E)*. IEEE. 2022, pp. 23–30.

[46] Aarzoo Dhiman and Durga Toshniwal. "An approximate model for event detection from Twitter data". In: *IEEE Access* 8 (2020), pp. 122168–122184.

[47] Tulsee Doshi, Emma Marriott, and Jay Patel. "CS224N Final Project: Detecting Key Needs in Crisis". In: (2017).

[48] Christopher Ifeanyi Eke, Azah Anir Norman, and Liyana Shuib. "Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model". In: *IEEE Access* 9 (2021), pp. 48501–48518.

[49] Ronald Aylmer Fisher. "Statistical methods for research workers". In: *Breakthroughs in statistics*. Springer, 1992, pp. 66–70.

[50] 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.

[51] Yasmeen George, Shanika Karunasekera, Aaron Harwood, and Kwan Hui Lim. "Real-time spatio-temporal event detection on geotagged social media". In: *Journal of Big Data* 8.1 (2021), pp. 1–28.

[52] Alec Go, Richa Bhayani, and Lei Huang. "Twitter sentiment classification using distant supervision". In: *CS224N project report, Stanford* 1.12 (2009), p. 2009.

[53] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[54] Imane Guellil, Ahsan Adeel, Faical Azouaou, and Amir Hussain. "Sentialg: Automated corpus annotation for algerian sentiment analysis". In: *International conference on brain inspired cognitive systems*. Springer. 2018, pp. 557–567.

[55] Imane Guellil, Faical Azouaou, and Francisco Chiclana. "ArAutoSenti: automatic annotation and new tendencies for sentiment classification of Arabic messages". In: *Social Network Analysis and Mining* 10.1 (2020), pp. 1–20.

[56] Maria Habib, Mohammad Faris, Alaa Alomari, and Hossam Faris. "AltibbiVec: A Word Embedding Model for Medical and Health Applications in the Arabic Language". In: *IEEE Access* 9 (2021), pp. 133875–133888.

[57] Aimad Hakkoum and Said Raghay. "Semantic Q&A System on the Qur'an". In: *Arabian Journal for Science and Engineering* 41.12 (2016), pp. 5205–5214.

[58] Btool Hamoui, Mourad Mars, and Khaled Almotairi. "FloDusTA: Saudi tweets dataset for flood, dust storm, and traffic accident events". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. 2020, pp. 1391–1396.

[59] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory". In: *Neural computation* 9.8 (1997), pp. 1735–1780.

[60] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. "Knowledge-based weak supervision for information extraction of overlapping relations". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*. 2011, pp. 541–550.

[61] Lamia Al-Horaibi and Muhammad Badruddin Khan. "Sentiment analysis of arabic tweets using semantic resources". In: *International Journal of Computing & Information Sciences* 12.2 (2016), p. 149.

[62] Guellil Imane, Darwish Kareem, and Azouaou Faical. "A set of parameters for automatically annotating a Sentiment Arabic Corpus". In: *International Journal of Web Information Systems* 15.5 (2019), pp. 594–615. URL: https://doi.org/10.1108/IJWIS-03-2019-0008.

[63] Muhammad Imran, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier. "Extracting information nuggets from disaster-Related messages in social media." In: *Iscram* 201.3 (2013), pp. 791–801.

[64] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. "Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages". In: *arXiv preprint arXiv:1605.05894* (2016).

[65] Karen Sparck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of documentation* 28.1 (1972), pp. 11–21. URL: https://doi.org/10.1108/eb026526.

[66] Ibrahim Kaibi, Hassan Satori, et al. "A comparative evaluation of word embeddings techniques for twitter sentiment analysis". In: *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)*. IEEE. 2019, pp. 1–4.

[67] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[68] Efsun Sarioglu Kayi, Linyong Nan, Bohan Qu, Mona Diab, and Kathleen Mckeown. "Detecting Urgency Status of Crisis Tweets: A Transfer Learning

Approach for Low Resource Languages". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 4693–4703.

[69] Jens Kersten, Anna Kruspe, Matti Wiegmann, and Friederike Klan. "Robust filtering of crisis-related tweets". In: *ISCRAM 2019 conference proceedings-16th international conference on information systems for crisis response and management*. Valencia, Spanien, May 2019.

[70] Prashant Khare. "Identifying and Processing Crisis Information from Social Media". PhD thesis. 2020.

[71] Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. URL: https://aclanthology.org/D14-1181.

[72] Papa Moryba Kouate. *Evaluation metrics for classification*. Ed. by Towards Data Science. https://towardsdatascience.com/evaluation-metrics-for-classification-1dc9945bee2, last accessed on 2022-06-23. Sept. 2020.

[73] Diego Kozlowski, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane. "A three-level classification of French tweets in ecological crises". In: *Information Processing & Management* 57.5 (2020), p. 102284.

[74] Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. "An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media". In: *Informatics*. Vol. 8. 1. Multidisciplinary Digital Publishing Institute. 2021, p. 19.

[75] Shir Meir Lador. "What metrics should be used for evaluating a model on an imbalanced data set". In: *Towards Data Science* 5 (2017).

[76] James Lane. "The 10 most spoken languages in the world". In: *Babbel Magazine* 6 (2019).

[77] Dong-Hyun Lee et al. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks". In: *Workshop on challenges in representation learning, ICML*. Vol. 3. 2. 2013, p. 896.

[78] Gang Li, Cathy Wu, and K Vijay-Shanker. "Noise reduction methods for distantly supervised biomedical relation extraction". In: *BioNLP 2017*. 2017, pp. 184–193.

[79] Hongmin Li. *Domain adaptation approaches for classifying social media crisis data*. Kansas State University, 2021.

[80] Hongmin Li, Doina Caragea, and Cornelia Caragea. "Combining Self-training with Deep Learning for Disaster Tweet Classification". In: *The 18th International Conference on Information Systems for Crisis Response and Management (ISCRAM 2021)*. 2021.

[81] Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. "Disaster response aided by tweet classification with a domain adaptation approach". In: *Journal of Contingencies and Crisis Management* 26.1 (2018), pp. 16–27.

[82] Hongmin Li, Oleksandra Sopova, Doina Caragea, and Cornelia Caragea. "Domain adaptation for crisis data using correlation alignment and self-training". In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 10.4 (2018), pp. 1–20.

[83] Xukun Li and Doina Caragea. "Domain adaptation with reconstruction for disaster tweet classification". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 1561–1564.

[84] Junhua Liu, Trisha Singhal, Lucienne TM Blessing, Kristin L Wood, and Kwan Hui Lim. "Crisisbert: a robust transformer for crisis classification and contextual crisis embedding". In: *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*. 2021, pp. 133–141.

[85] Walid Magdy, Hassan Sajjad, Tarek El-Ganainy, and Fabrizio Sebastiani. "Distant supervision for tweet classification using youtube labels". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 9. 1. 2015, pp. 638–641.

[86] Micol Marchetti-Bowick and Nathanael Chambers. *Learning for microblogs with distant supervision: Political forecasting with twitter*. Tech. rep. MICROSOFT CORP SAN FRANCISCO CA, 2012.

[87] Rawan N Al-Matham and Hend S Al-Khalifa. "Synoextractor: a novel pipeline for Arabic synonym extraction using Word2Vec word embeddings". In: *Complexity* 2021 (2021).

[88] Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. "Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection". In: *arXiv preprint arXiv:2103.14916* (2021).

[89] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 2009, pp. 1003–1011.

[90] Salman Mohammed, Nimesh Ghelani, and Jimmy Lin. "Distant supervision for topic classification of tweets in curated streams". In: *arXiv preprint arXiv:1704.06726* (2017).

[91] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. "Comparative study of word embedding methods in topic segmentation". In: *Procedia computer science* 112 (2017), pp. 340–349.

[92] Dat Tien Nguyen, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra. "Robust classification of crisis-related data on social networks using convolutional neural networks". In: *Eleventh international AAAI conference on web and social media*. 2017.

[93] Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. "Applications of online deep learning for crisis response using social media information". In: *arXiv preprint arXiv:1610.01030* (2016).

[94] Kan Nishida. *K-means clustering - deciding how many clusters to build*. Ed. by Exploratory. https://exploratory.io/note/kanaugust/K-Means-Clustering-Finding-the-optimal-K-Number-of-Clusters-yAp2MbM7bk, last accessed 2022-06-23.

[95] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. "Crisislex: A lexicon for collecting and filtering microblogged communications in crises". In: *Eighth international AAAI conference on weblogs and social media*. 2014.

[96] 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.

[97] Leysia Palen and Kenneth M Anderson. "Crisis informatics—New data for extraordinary times". In: *Science* 353.6296 (2016), pp. 224–225.

[98] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[99] Nikolaos Panagiotou, Ioannis Katakis, and Dimitrios Gunopulos. "Detecting events in online social networks: Definitions, trends and challenges". In: *Solving Large Scale Learning Tasks. Challenges and Algorithms*. Springer, 2016, pp. 42–84.

[100] Nayan Ranjan Paul, Deepak Sahoo, and Rakesh Chandra Balabantaray. "Classification of crisis-related data on Twitter using a deep learning-based framework". In: *Multimedia Tools and Applications* (2022), pp. 1–21.

[101] 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.

[102] Linda Plotnick, Starr Roxanne Hiltz, Jane A Kushma, and Andrea H Tapia. "Red Tape: Attitudes and Issues Related to Use of Social Media by US County-Level Emergency Managers." In: *ISCRAM*. 2015.

[103] Philips Kokoh Prasetyo, Ming Gao, Ee-Peng Lim, and Christie Napa Scollon. "Social sensing for urban crisis management: The case of singapore haze". In: *International conference on social informatics*. Springer. 2013, pp. 478–491.

[104] Jianfeng Qu, Dantong Ouyang, Wen Hua, Yuxin Ye, and Ximing Li. "Distant supervision for neural relation extraction integrated with word attention and property features". In: *Neural Networks* 100 (2018), pp. 59–69.

[105] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. "Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake". In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 2011, pp. 25–34.

[106] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *arXiv preprint arXiv:1910.10683* (2019).

[107] Christian Reuter, Thomas Ludwig, Therese Friberg, Sylvia Pratzler-Wanczura, and Alexis Gizikis. "Social media and emergency services?: Interview study on current and potential use in 7 European countries". In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 7.2 (2015), pp. 36–58.

[108] Sebastian Riedel, Limin Yao, and Andrew McCallum. "Modeling relations and their mentions without labeled text". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2010, pp. 148–163.

[109] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. "Transfer learning in natural language processing". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*. 2019, pp. 15–18.

[110] Radwa MK Saeed, Sherine Rady, and Tarek F Gharib. "An ensemble approach for spam detection in Arabic opinion texts". In: *Journal of King Saud University-Computer and Information Sciences* 34.1 (2022), pp. 1407–1416.

[111] 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.

[112] Arun Kumar Sangaiah, Ahmed E Fakhry, Mohamed Abdel-Basset, and Ibrahim El-henawy. "Arabic text clustering using improved clustering algorithms with dimensionality reduction". In: *Cluster Computing* 22.2 (2019), pp. 4535–4549.

[113] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[114] Qianzi Shen, Zijian Wang, and Yaoru Sun. "Sentiment analysis of movie reviews based on cnn-blstm". In: *International Conference on Intelligence Science*. Springer. 2017, pp. 164–171.

[115] 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* (2019).

[116] Luke Sloan, Jeffrey Morgan, William Housley, Matthew Williams, Adam Edwards, Pete Burnap, and Omer Rana. "Knowing the Tweeters: Deriving sociologically relevant demographics from Twitter". In: *Sociological research online* 18.3 (2013), pp. 74–84.

[117] Rion Snow, Daniel Jurafsky, and Andrew Ng. "Learning syntactic patterns for automatic hypernym discovery". In: *Advances in neural information processing systems* 17 (2004).

[118] Kate Starbird, Leysia Palen, Amanda L Hughes, and Sarah Vieweg. "Chatter on the red: what hazards threat reveals about the social life of microblogged information". In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 2010, pp. 241–250.

[119] Peng Su, Gang Li, Cathy Wu, and K Vijay-Shanker. "Using distant supervision to augment manually annotated data for relation extraction". In: *PloS one* 14.7 (2019), e0216913.

[120] Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. "Reducing wrong labels in distant supervision for relation extraction". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2012, pp. 721–729.

[121] Sakirin Tam, Rachid Ben Said, and Ö Özgür Tanriöver. "A ConvBiLSTM deep learning model-based approach for Twitter sentiment classification". In: *IEEE Access* 9 (2021), pp. 41283–41293.

[122] Andrea H Tapia and Kathleen Moore. "Good enough is good enough: Overcoming disaster response organizations' slow social media data adoption". In: *Computer supported cooperative work (CSCW)* 23.4 (2014), pp. 483–512.

[123] Robert Thomson, Naoya Ito, Hinako Suda, Fangyu Lin, Yafei Liu, Ryo Hayasaka, Ryuzo Isochi, and Zhou Wang. "Trusting tweets: The Fukushima disaster and information source credibility on Twitter." In: *Iscram*. 2012.

[124] Hien To, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. “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.

[125] Lisa Torrey and Jude Shavlik. “Transfer learning”. In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[126] Ibtissam Touahri and Azzeddine Mazroui. “Deep analysis of an Arabic sentiment classification system based on lexical resource expansion and custom approaches building”. In: *International Journal of Speech Technology* 24.1 (2021), pp. 109–126.

[127] Zeynep Tufekci and Christopher Wilson. “Social media and the decision to participate in political protest: Observations from Tahrir Square”. In: *Journal of communication* 62.2 (2012), pp. 363–379.

[128] Sudha Verma, Sarah Vieweg, William Corvey, Leysia Palen, James Martin, Martha Palmer, Aaron Schram, and Kenneth Anderson. “Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 1. 2011, pp. 385–392.

[129] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. “Microblogging during two natural hazards events: what twitter may contribute to situational awareness”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2010, pp. 1079–1088.

[130] Sarah Elizabeth Vieweg. “Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications”. PhD thesis. University of Colorado at Boulder, 2012.

[131] Junaid Abdul Wahid, Lei Shi, Yufei Gao, Bei Yang, Lin Wei, Yongcai Tao, Shabir Hussain, Muhammad Ayoub, and Imam Yagoub. “Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response”. In: *Expert Systems with Applications* 195 (2022), p. 116562.

[132] Congcong Wang, Paul Nulty, and David Lillis. "Crisis Domain Adaptation Using Sequence-to-sequence Transformers". In: *arXiv preprint arXiv:2110.08015* (2021).

[133] Mei Wang and Weihong Deng. "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312 (2018), pp. 135–153.

[134] Si Si Mar Win. "Automated text annotation for social media data during natural disasters". PhD thesis. MERAL Portal, 2018.

[135] 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.

[136] Zhilin Yang, Ruslan Salakhutdinov, and William W Cohen. "Transfer learning for sequence tagging with hierarchical recurrent networks". In: *arXiv preprint arXiv:1703.06345* (2017).

[137] Kiran Zahra, Muhammad Imran, and Frank O Ostermann. "Automatic identification of eyewitness messages on twitter during disasters". In: *Information processing & management* 57.1 (2020), p. 102107.

[138] Ying Zeng, Yansong Feng, Rong Ma, Zheng Wang, Rui Yan, Chongde Shi, and Dongyan Zhao. "Scale up event extraction learning via automatic training data generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[139] Shanshan Zhang and Slobodan Vucetic. "Semi-supervised discovery of informative tweets during the emerging disasters". In: *arXiv preprint arXiv:1610.03750* (2016).

[140] Wu Zheng and Catherine Blake. "Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles". In: *Journal of biomedical informatics* 57 (2015), pp. 134–144.