Crisis Detection from Arabic Social Media

By

Alaa Alharbi

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
September 2023

Abstract

Social media (SM) streams such as Twitter provide large quantities of real-time information about emergency events from which valuable information can be extracted to enhance situational awareness and support humanitarian response efforts. The timely extraction of crisis-related SM messages is challenging as it involves processing large quantities of noisy data in real time. Supervised machine learning classifiers are challenged by out-of-distribution learning when classifying unseen (new) crises due to data variations across events. Besides that, it is impractical to label training data from each novel and emerging crisis since obtaining sufficient labelled data is time-consuming and labour-intensive. This thesis addresses the problem of Twitter crisis classification using supervised learning methods to identify crisis-related data and categorising them into different information types in the multi-source (training data from multiple events) setting. Due to Twitter's ubiquity during emergency events in the Arab world, the current research focuses on Arabic Twitter content. We have created and published a large-scale Arabic Twitter corpus of crisis events. The corpus has been analysed and manually labelled. Analysing the content includes investigating the main information categories of conversations posted during a range of crisis events using natural language processing techniques. Building these resources is considered one of this thesis's contributions.

The thesis also investigates the generalisation performance of different supervised classical machine learning and deep learning approaches trained on out-of-crisis data to classify unseen crises. We find that deep neural networks such as LSTM and CNN outperform

the classical machine learning classifiers such as support vector machines and decision trees. We also evaluate different architectures of deep neural networks and several pre-trained text representations (embeddings) learnt from vast amounts of unlabelled text. Results show that BERT-based models are more robust to out-of-distribution target events and remarkably outperform other models on the information classification task. Experiments show that the performance of BERT-based classifiers can be enhanced when training on similar data. Thus, the last contribution of the present study is to propose an instance distance-based data selection approach for adaptation to improve classifiers' performance under a domain shift. Using the BERT embeddings, the method selects a subset of multi-event training data that is most similar to the target event. Results show that fine-tuning a BERT model on a selected subset of data to classify crisis tweets outperforms a model that has been fine-tuned on all available source data.

Acknowledgements

In the name of Allah, the Most Gracious and the Most Merciful

"There is no power in me to do something except through the help of Allah. In Him do I put my trust and to Him do I always turn." (Hood - 88)

I would like to take this opportunity to express my sincere gratitude to several people who supported me during this long journey. My deep gratitude and appreciation go to my supervisor *Prof. Mark Lee* for providing me with guidance, support and inspiring advice while conducting my research. Thank you very much, Mark, for motivating and encouraging me during my years of study.

I thank Dr Peter Hancox and my second supervisor Dr Phillip Smith for their insightful comments and valuable suggestions I received during the thesis group meetings. I also thank the NLP group members and my colleagues and friends for their stimulating discussions and advice.

I am indebted to *Taibah University* in Medina for offering me a fully-funded scholarship to do my PhD in the UK. This work would not have been possible without their financial support. I also thank my colleagues and friends in the College of Computer Science and Engineering at Taibah University for their support.

No words can express my gratitude to my parents for their unconditional support, love, prayers and the sacrifices they have made for me. To my parents: شكراً جزيلاً. I extend my gratitude to my sisters, brothers, nieces and nephews. Heartful thanks and sincere appreciation go to my sister Hadeel. Thanks, Hadeel, for the joyful time and lively discussions. I cannot imagine how this long-term studying experience abroad would be without you. I owe a debt of gratitude to my best friend Thuraya Al-Samarkandi. Thanks, Thuraya, for your prayers, thoughts and wishes.

Publications

The following publications have been generated while developing this thesis, and to an extent have guided the thesis into what it has become.

- **Alharbi, Alaa** and Mark Lee (July 2019). "Crisis Detection from Arabic Tweets". In: *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*. Cardiff, United Kingdom: Corpus Linguistics Conference, pp. 72–79.

- **Alharbi, Alaa** and Mark Lee (Apr. 2021). "Kawarith: an Arabic Twitter Corpus for Crisis Events". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): European Association for Computational Linguistics (EACL), pp. 42–52.

- **Alharbi, Alaa** and Mark Lee (June 2022). "Classifying Arabic Crisis Tweets using Data Selection and Pre-trained Language Models". In: *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools, with Shared Tasks on Quran QA and Fine-Grained Hate Speech Detection*. Marseille, France: LREC, pp. 71-78.

Contents

List of Figures

List of Tables

Acronyms

ANN Artificial Neural Network. 20

BERT Bidirectional Encoder Representation from Transformers. 36

Bi-LSTM Bidirectional Long Short-Term Memory. 18

BoW Bag of Words. 33

CBOW Continuous Bag-Of-Words. 35

CLSTM Convolutional LSTM. 80

CNN Convolutional Neural Network. 30

COVID-19 Corona Virus Disease 2019. 51

CRF Conditional Random Field. 16

DNNs Deep Neural Networks. 6

ELMo Embeddings from Language Models. 19

GRU Gated Recurrent Unit. 32

GloVe Global Vectors for Word Representation. 18

LDA Latent Dirichlet Allocation. 39

LSTM Long Short-Term Memory. 31

ML Machine Learning. 6

MLM Masked Language Model. 36

MSA Modern Standard Arabic. 13

NB Naïve Bayes. 28

NLP Natural Language Processing. 1

NPMI Normalised Pointwise Mutual Information. 68

OOV Out Of Vocabulary. 35

POS Part-Of-Speech. 27

QA Question Answering. 36

RNN Recurrent Neural Network. 31

ReLU Rectified Linear Unit. 79

SM Social Media. 1

SVM Support Vector Machines. 29

TF-IDF Term Frequency-Inverse Document Frequency. 27

USE Universal Sentence Encoder. 20

VSM Vector Space Models. 33

Chapter One

Introduction

1.1 Overview and Motivation

In the last decade, user-generated Social Media (SM) content has been explored by Natural Language Processing (NLP) and data mining researchers as a valuable and accessible source of data (Ritter et al., 2012; Imran et al., 2018; Ahmed et al., 2019; Karami et al., 2020a). Many of these studies have investigated the problem of mining SM (notably microblogging websites) to extract real-world events. Microblogging is an online social networking and broadcast medium that enables users to post short messages in various content formats, including text, image, video and hyperlink. These websites play an important role in fast information diffusion, enabling users to express opinions, comment on news stories, share online news links and report real-world events.

On Twitter[1], which is one of the most popular microblogging services, users interact by posting short messages called 'tweets'. When Twitter launched in 2006, tweets were restricted to 140 characters, but in November 2017, that limit was increased to 280 characters per tweet for all users. Twitter is still growing fast. In March 2012, there were over 140 million active users, with around 340 million messages posted daily;[2] by the first quarter of 2018, Twitter

[1] https://twitter.com
[2] https://blog.twitter.com/official/en_us/a/2012/twitter-turns-six.html

had around 94 million daily active users.[3] There were 298 million annual users in 2018, and the number reached 401 million users in 2022.[4] In February 2022, there were around 217 million daily active users generating around 500 million tweets per day.[5] Users tweet about different aspects of their lives, sharing 'what's happening' with their social networks. By disseminating real-time information, Twitter plays an important role in reporting real-world events, ranging from the mundane and personal to major global happenings. Twitter has attracted increasing attention from both academic and industrial researchers. According to Steiger et al. (2015), 46% of Twitter-related published papers deal with event detection research. Petrovic et al. (2013) found that Twitter covers most of the events mentioned by news agencies, as well as many that are not reported by news agencies. They also showed that Twitter is often first to break incoming news in some cases, such as crisis-related events.

A crisis is a real-world emergency event that occurs at a particular time and location and is characterised by a main topic representing the hazard type (floods, explosions, pandemic etc.). A coherent and rapid understanding of the crisis enables the crisis management teams and the affected communities to respond and recover effectively. However, gaining a situational awareness during emergency incidents is challenging because of their chaotic and rapidly changing nature (Derczynski et al., 2018), which causes anxiety among most stakeholders (Bukar et al., 2020). Hence, crisis responders have utilised microblogging websites (particularly Twitter) as additional sources of up-to-date information to enhance emergency relief and response. Such platforms allow emergency teams to collect information directly from the affected communities to gain a better understanding of the situation.

The huge volume of user-generated Twitter data related to multiple daily events has created a need for automatic event extraction and summarising tools. The information overload makes it difficult for emergency services to process and extract relevant data in a

[3] https://www.statista.com/statistics/1032751/monetizable-daily-active-twitter-users-international/
[4] https://www.businessofapps.com/data/twitter-statistics/
[5] https://www.omnicoreagency.com/twitter-statistics/

timely manner. Time is crucial during emergencies. Crisis coordinators have a rich source of information but limited time. Manual inspection of Twitter data for useful information such as those sent by eyewitnesses and volunteers is challenging and time consuming because the volume of data is overwhelming, and most of the content relates to daily chatter (Java et al., 2007). Detection of disasters as they unfold and extracting informative crisis messages can reduce response time, so mitigating impacts.

This research utilised Twitter data for crisis detection for several reasons. Twitter is popular and accessible, and the enormous volume of data generated daily by Twitter users is available through the public streaming API. Additionally, Twitter offers instant publication, which supports rapid discovery of crises, and uses hashtags (#) that identify posts on a specific topic and can be utilised to collect the dataset. Twitter plays an important role in reporting emergency events. During crises, people and agencies use Twitter as a communication channel to post situational updates, request help, provide aid and search for actionable information (Vieweg et al., 2010; Olteanu et al., 2015; Takahashi et al., 2015). Landwehr and Carley (2014) demonstrated that individuals impacted by the crisis turn to SM to find information. Examples of Twitter's effectiveness during disasters include the 2007 and 2008 wildfires in California (Sutton et al., 2008), tropical storm Cindy in 2017 (J. Kim et al., 2018) and Hurricane Harvey in 2017 (Mihunov et al., 2020). Twitter was used to report the protests that followed the Iranian presidential elections of 2009 (Khondker, 2011; Diriöz, 2013). The Arab Spring (since 2010) has shed light on the role of SM during crises and social protests. For instance, protesters used Twitter to communicate during the Egyptian revolution in February 2011 (Tufekci and Wilson, 2012).

In the Arab World, Twitter has a higher rate of growth and more activity than other SM platforms (Diriöz, 2013). Despite Twitter's ubiquity and effectiveness during emergencies in the Arab World, there is little work investigating crisis detection from Arabic tweets. Besides that, there is no publicly available Arabic multi-type crisis-related dataset. These

opportunities motivated the present research to further investigate the problem of extracting crisis events from Arabic Twitter data.

1.2 Problem Definition and Tasks Description

In this research, we performed cross-domain or cross-event crisis detection from SM. In cross-event classification, models are trained to classify an unseen event. In other words, no data from the target event are included in the training set. We focused on multi-source cross-domain crisis classification, in which the training set includes tweets from different disasters. A domain is defined as a dataset that has been collected from SM for a specific event. Hence, each crisis data represents a distinct domain. In this thesis, we considered two supervised classification tasks, described as follows.

1.2.1 Relevancy Detection

The relevancy detection task is modelled as a binary classification task. It aims to identify crisis-related messages from SM by classifying them as related (on-topic) or not related (off-topic) to a specific event. For example, the tweet: “Flash #floods after heavy rains in Kuwait left one man dead." is related to the Kuwait floods. On the other hand, the tweet: “I will attend a conference on #floods tomorrow." is not related. Using Twitter, the relevancy detection follows the data acquisition process, which crawls candidate crisis messages. Candidate crisis-related tweets are first acquired by one of two methods: collecting posts from users in the affected areas (location-based) or tracking relevant keywords (keyword-based). The location-based sampling is limited to the geotagged posts or those tweets sent by users stating the location in their profiles (Rachunok et al., 2022). According to the Twitter

Platform[6], only ∼1-2% of tweets are geotagged, whereas 30-40% of tweets include profile location information. Most crisis Twitter datasets were built by tracking specific relevant keywords and hashtags. Such a process captures lots of relevant data but can also include irrelevant posts with various distribution across crises. Unrelated posts include advertisements, political views, unrelated personal messages or posts related to other disasters, as we will describe in Chapter 4. Such posts usually exploit trending hashtags to be more visible. Other off-topic tweets were crawled due to the keywords' ambiguity. Relevancy detection is challenging for the cross-event setting due to feature variations across disasters. The class imbalance makes this task more challenging.

1.2.2 Information Classification

This task categorises relevant messages into one or more information categories that support situational awareness and assist people who need help. Examples of information categories include affected people (fatalities, missing, injured and displaced), infrastructure damage, caution, preparations, etc. For example, the tweet: "Eighteen people, mainly schoolchildren, were killed on Thursday by a flash flood in #Jordan." is categorised as affected individuals, while the tweet: "A bridge leading to the site collapsed this morning under the force of the rains" is about infrastructure damage.

This task is modelled as either a multi-class or multi-label problem. In this work, our data has been annotated using a multi-label scheme, as we found that some tweets can communicate more than one humanitarian information category. The method followed to identify these categories will be explained in Chapter 4.

[6] https://developer.twitter.com/en/docs/tutorials/advanced-filtering-for-geo-data

1.3 Research Questions

The main hypothesis of the thesis is: it is possible to automatically identify crisis event messages from SM in real time by Deep Neural Networks ($DNNs$) trained on labelled historical multi-event crisis data, despite the feature distribution gap between the training and testing data. We use Twitter as a SM platform and focus on Arabic content due to Twitter's ubiquity during emergency events in the Arab world. To test this hypothesis, we ask the following questions which identify the scope of the research.

Research Question 1: What are the main information categories of conversations posted during different types of crises on Arabic Twitter data?

There is no accessible Arabic SM corpus for multi-type crisis events. Thus, we create one by collecting a corpus from Twitter for crisis events. In this work, we investigate the main information categories that support humanitarian response efforts to be used to label a multi-event crisis dataset. Hence, we ask the first question to identify the information discussed during different emergency events (floods, explosions, etc.). Because the manual inspection of events' content is hard, we identify the main topics using some NLP techniques and analyse posts belonging to each topic in the light of previously proposed crisis taxonomies.

To test our hypothesis, we explore the generalisation performance of DNNs trained on multi-event crisis data to classify tweets from an unseen crisis. We ask the following question to explore how the generalisability of DNNs compares to the classical (conventional) Machine Learning (ML) classifiers for cross-event classification. We experiment with different DNN architectures and various text representations.

Research Question 2: Using training data from multiple historical events, how does the performance of DNNs compare to that of classical ML classifiers in identifying crisis-related posts and categorising relevant posts into different information types?

To improve the performance of cross-crisis classifiers, we propose an instance-based data selection method to train a classifier on data similar to the target event. The approach uses the K-nearest neighbours algorithm for data selection. We evaluate the performance of our presented method by comparing it with two approaches: the BERT model that learns from all source data and the BERT-based self-training adaptation approach. Finally, we explore whether combining our data selection with a semi-supervised self-training enhances the performance of the data selection technique. Hence, We ask the following questions:

Research Question 3: How do the results of the proposed data selection model compare to the results of the BERT model that learns from all source data?

Research Question 4: How do the results of the proposed data selection model compare to the results of the self-training approach?

Research Question 5: Does combining the proposed selection method with self-training result in performance gain?

1.4 Contributions

The thesis makes the following contributions to the field of crisis detection from SM text:

- Creation and publication of a large-scale Arabic Twitter corpus of crisis events, named Kawarith[7, 8]. The corpus includes ∼1.6M tweets collected during 22 crises and involving several types of hazard. The corpus comprises multi-dialect Arabic tweets, as it was collected from different regions.

- Providing quantitative and qualitative analysis of the corpus content, including investigating the main information categories of conversations posted during various crisis

[7]This is the Romanised form of the Arabic word كوارث, meaning *crises*.

[8]The corpus is available at https://github.com/alaa-a-a/kawarith

events using NLP techniques such as topic detection models.

- Creation and publication of a manually annotated Arabic Twitter dataset of more than 12k messages from seven emergency events.

- Comparing the performance of several classical ML models and DNNs (in different configurations) for cross-event classification to investigate which models can generalise better to classify unseen crisis data when learning from historical crisis events. The evaluation has been conducted for two crisis tasks: binary relevancy detection and multi-label information classification.

- Proposing an instance-based data selection approach to train a classifier on the best matching data for the target crisis instead of using all multi-event source data. Results show that our selection adaptation models outperform the equivalent models that learn from all available source data.

1.5 Thesis Structure

The reminder of this thesis has been organised into the following chapters:

Chapter 2 begins with an overview of the crisis detection problem. Then, it presents the relevant literature. It also describes the supervised classifiers and text representation models used in our experiments.

Chapter 3 highlights our research methodology. It presents an outline of the Twitter dataset creation process and describes the procedure we follow to evaluate the supervised learning models. The last part of the chapter introduces our proposed data selection method.

Chapter 4 introduces and describes our Arabic crisis-related Twitter corpus. It explains the data collection process and discusses the main topics and information categories of conversations posted during various emergency events. Finally, the chapter presents our labelled

dataset used in this study and describes the used annotation scheme. This chapter is based on our published paper (A. Alharbi and M. Lee, 2021).

Chapter 5 evaluates the generalisation performance of several DNN models and classical ML classifiers in identifying Arabic crisis-related tweets and classifying them into information types in the cross-event setting. Different DNN architectures and pre-trained text representations (such as character embeddings) have been evaluated for each classification task. Finally, the chapter concludes with a discussion and error analysis of the best models. This chapter is based on our published work (A. Alharbi and M. Lee, 2019).

Chapter 6 presents the experimental setup of our data selection method. It compares our method against two baselines: training using all available datasets and the self-training approach. The chapter wraps up with the results and discussion. Chapter 6 has been recently published (A. Alharbi and M. Lee, 2022).

Chapter 7 concludes the thesis by outlining the findings of the current study and discusses areas for future work.

Chapter Two

Background and Literature Review

This chapter starts by presenting an overview of crisis detection from Twitter, highlighting the problem of out-of-distribution learning when classifying new crises. Then, we discuss the relevant literature. The related work is divided into three sections. The first one presents the tweet classification approaches in the context of crisis detection. The second section shows the domain adaptation approaches applied in this area. The last part of relevant literature overviews the publicly accessible Twitter crisis-related corpora in the crisis informatics literature. Finally, the chapter briefly presents the supervised learning approaches and text representation models adopted in our experiments.

2.1 Crisis Detection from Twitter Data

Event extraction from Twitter streams poses challenges that differ from traditional media. In particular, traditional text extraction techniques are challenged by the noisy language used in SM, including misspelt words, grammar mistakes, colloquialisms and non-standard acronyms. Because of the imposed character limit, Twitter users tend to use more abbreviations. Users may also post non-informative messages that require some knowledge of the situational context for interpretation by humans. For example, the tweet "حتى سيارتي نفس الشيء #سيول_الكويت" (and my car, the same thing #Kuwait_floods)

does not convey obvious meaning to decide whether it is relevant to the crisis and which information it communicates. Looking at the tweet in context, it is a reply to someone reporting that his car was washed away and damaged by the floods. Thus, we can understand the meaning of the reply. Interpreting such short texts automatically is a hard problem.

Twitter's popularity makes it appealing to spammers who spread propaganda, pornography and advertisements (Benevenuto et al., 2010; Kabakus and Kara, 2017). For advertising, spammers use crisis-related hashtags in their posts to be noticed by large numbers of users. For instance, the following tweet is irrelevant but was captured in the crisis dataset as it included a relevant hashtag.

لا يفوتكم الخصم القوي من باث اند بودي، خصم ٢٠ % علي جميع المنتجات #خصم # سيول ـالكويت

(Don't miss the big discount from Bath and Body, 20% off all products #discount #Kuwait_floods)

We found many tweets like the previous one in our dataset. For similar purposes, some people who are not from the crisis-affected population exploit the popularity of crisis hashtags to request help, ask for money or report lost items, as in the following tweet. Such tweets look like crisis-related messages and are hard to identify automatically.

انشروها في جميع التواصل الاجتماعى جزاكم الله خيرا يحتاج فزعه وعلاج ومساعده لحادث تعرض له اليوم. #يوم ـالجمعة #صوت ـالمطر سيول الأردن

(Please share on all social media. May God reward you. He needs help and treatment as he had an accident today. #Friday #rain_sound Jordan floods)

Like any online SM site, the credibility of information shared on Twitter is always in question because of the (relative) freedom of posting, and there is evidence that Twitter is

used—if sometimes unintentionally—to disseminate rumours, misinformation and false news to large communities (Castillo et al., 2011; Shu et al., 2017). The increasing volume and high-rate data stream of user-generated messages on Twitter create significant computational demand and challenge the data mining techniques that are employed to extract real-time events from a data stream that changes quickly over time.

Crisis detection from Twitter is also challenging because of the lack of labelled data from current events. Supervised learning approaches require in-domain human-labelled datasets for training algorithms to predict outcomes accurately. However, annotated data are unlikely to be available in real-time from emerging crises since obtaining sufficient human-labelled data is time consuming. In contrast, time is a critical factor during mass emergencies.

Researchers proposed to use labelled data from past events to classify new disasters (Imran et al., 2013b; Rudra et al., 2015; C. Caragea et al., 2016). However, supervised methods are challenged by out-of-distribution learning when classifying unseen crises — especially if they are trained on data from cross-type crises due to data variations across such events. Out-of-distribution (covariate shift) refers to the different probability distributions of input features across the training (source) and test (target) data (Ramponi and Plank, 2020). Labelling instances from each possible type of disaster (e.g. flood, wildfire, explosion, etc.) to minimise the feature distribution gap is impractical: such data is hard to collect and expensive to annotate. Training on data from the same disaster type (in-type event) may not improve the model's generalisation.

Textual data varies across SM events in two main aspects: topic and language. Messages from two events of the same type can discuss various topics emerging from the event properties and its distinct aspects such as time, cause, related entities and impact. Such topic variations across events can lead to substantially different feature distributions. Furthermore, the discussed topics can change over time during a crisis. Events on SM are

discussed with varying levels of formality, in various dialects and languages, resulting in a more significant distributional gap across domains. Supervised classifiers' performance drops on test data if it does not follow the training set distribution as many supervised learning algorithms assume (Ramponi and Plank, 2020). Thus, the out-of-distribution problem challenges SM crisis classification models as they learn under distributional shifts. In our work, we evaluate the generalization performance of different classifiers to identify crisis messages in the cross-event setting. We also propose a domain adaptation method to minimize the effect of the out-of-distribution learning, as shown in the following chapters.

2.2 Crisis Detection from Arabic Twitter

The Arabic language refers to a collection of varieties, including a standardised form, Modern Standard Arabic (MSA), and several regional dialects, which are spoken by more than 300 million native speakers (Althobaiti, 2020). Examples of Arabic dialects include the Gulf, Egyptian, Levantine and Maghrebi dialects. The dialects began to appear in a written form with the advent of Web 2.0 (Althobaiti, 2020). SM content, including Twitter, has a strong presence of dialectal Arabic (Shoufan and Alameri, 2015; Alsarsour et al., 2018).

Identifying crises from Arabic SM poses numerous challenges. The Arabic dialects differ in their phonology, morphology and syntax (Chiang et al., 2006). In contrast to MSA, there is no standard orthography for dialectical Arabic (Shoufan and Alameri, 2015). The same words can be written in different forms, usually according to people's pronunciation, resulting in spelling inconsistencies. Some dialectal words can have different spellings, even within the same dialect (Zaidan and Callison-Burch, 2014). Dialects are region-based. Hence, we expect that our corpus includes several dialects as it has been collected during crises that occurred in different areas. The variation of Arabic dialects would result in more covariate

shifts across the data. A classifier trained on event data collected from one region may not perform well when classifying data collected from another area.

Analysing the effect of dialect familiarity on the quality of data annotation on sarcasm detection, Farha and Magdy (2022) showed that annotators performed better in labelling text written in their dialects and dialects they are familiar with. While creating the labelled dataset, we considered this issue by allowing the coders to skip/leave the example if it is hard to judge the tweet because of the dialect variation. For example, one of the annotators left a tweet unjudged because she did not understand the idiom 'رافع الجام على', which is commonly used in Kuwait, meaning to ignore. Spam is prevalent on Arabic Twitter (El-Mawass and Alaboodi, 2016). El-Mawass and Alaboodi (2016) found that about three-quarters of the tweets in trending hashtags in Saudi Arabia, which has the highest number of active Twitter users in the Arab nations, are spam posts and irrelevant to hashtags. Advertising accounts target popular accounts and hashtags on Arabic Twitter to promote their services and products (Mubarak et al., 2020). We expected the same pattern in our corpus, as crisis-related hashtags are usually trending on Twitter. Such spam messages should be filtered out after data collection as they are unrelated to the crisis. Unlike English, there is no capitalisation in the Arabic language, which was used as a feature to identify event-related messages in previous research (Ashktorab et al., 2014), since crisis posts usually include named entities.

This research evaluates the robustness of several widely adopted supervised models and text representations to classify crisis tweets in the multi-dialect setting. Besides that, we experiment with the character-level embeddings to investigate whether they improve the results over word-level embeddings, given that Arabic is a morphologically rich language. Before tweet classification, we pre-process the text to transform the Arabic words into a more uniform sequence. We also create a list of 405 domain-independent multi-dialect Arabic stop words when identifying the main topics in the corpus, as we will show in Chapter 4.

2.3 Related Work

2.3.1 Crisis Detection using Supervised Classifiers

Researchers have shown an increased interest in using supervised learning to extract useful information from SM crisis events to enhance emergency relief and response. To address the data overload problem, some research studies focus on the relevancy detection task, which classifies SM posts into crisis-related (on-topic) or irrelevant (off-topic) (Ashktorab et al., 2014; To et al., 2017; Kersten et al., 2019; Liu et al., 2021). Other studies focused on informative messages identification (Verma et al., 2011; Rudra et al., 2015; C. Caragea et al., 2016; D. Nguyen et al., 2017; Rudra et al., 2018; Neppalli et al., 2018; Derczynski et al., 2018; Madichetty and Sridevi, 2019a; Graf et al., 2020). The informativeness of an SM message during emergencies has been defined based on its relevancy to pre-defined information types or by its usefulness to situational awareness (Rudra et al., 2015; Olteanu et al., 2015). The informativeness identification task has been formulated as a binary classification problem. Several publications went beyond such binary classification tasks by classifying the relevant content into different pre-defined information categories (Imran et al., 2013a; Imran et al., 2013b; Imran et al., 2014; ALRashdi and O'Keefe, 2018; Alam et al., 2019; Madichetty and Sridevi, 2019a). Other existing research only focused on identifying SM messages reporting a specific information type such as the infrastructure damage (Madichetty and Sridevi, 2019b).

Previous studies applied classical ML approaches with handcrafted features to identify messages of interest during disasters (Verma et al., 2011; Imran et al., 2013a; Imran et al., 2013b; Ashktorab et al., 2014; Imran et al., 2014; Parilla-Ferrer et al., 2014; Rudra et al., 2015; Cobo et al., 2015; To et al., 2017; Rudra et al., 2018). Sakaki et al. (2010) developed an earthquake reporting system by processing Twitter data. They used SVM with statistical, keyword and word context features to identify messages reporting earthquake occurrence.

The results showed that the features did not contribute equally to the performance, and the word context attributes had the least contribution. Their study is limited to using only two query terms (earthquake and shaking) to collect tweets that might be relevant to the earthquake. Thus, their system will miss the event-related tweets that do not include these terms. Verma et al. (2011) identified situational awareness tweets using NB and maximum entropy classifiers with handcrafted text-based features such as unigrams, bigrams and POS tags. The authors demonstrated that the models performed well when classifying data from the same event and showed that the maximum entropy model achieved better results. However, their model did not generalise well when classifying cross-event data, particularly when the source and target crises have different types and characteristics. For example, the model produced a poor accuracy (29%) when trained on data from Oklahoma fires to classify the Haiti earthquake tweets.

Imran et al. (2013a) classified Twitter posts from multi disasters into fine-grained classes using an NB model and a set of statistical and text features. The authors only evaluated their model on the same dataset (i.e., the Joplin tornado) and did not show how their proposed model generalises to other crises. In a subsequent study, they described a method for identifying informative tweets and extracting the relative information from them using an NB and Conditional Random Field (CRF) with manually generated features (Imran et al., 2013b). They experimented on two disasters: the Joplin tornado and Hurricane Sandy. Their results show that performance hugely drops in the cross-event setting. The authors considered an adaptation scenario that improved performance by incorporating 10% of test data into the training set. Similarly, Ashktorab et al. (2014) experimented with different classical ML models such as NB, logistic regression and decision trees to find disaster-related tweets. They also used CRF with several handcrafted features to extract actionable information. The models produced low performance when classifying cross-event data. Using the n-gram features, Parilla-Ferrer et al. (2014) detected informative crisis messages. Cobo

et al. (2015) compared the performance of NB, logistic regression, SVM and random forest with user-based and content-based features to classify earthquake tweets and found that random forest outperformed other classifiers. Both studies did not evaluate their work in the cross-event setting.

To enhance models' performance on unseen crises, Rudra et al. (2015, 2018) used vocabulary-independent, low-level lexical and syntactic features to identify tweets reporting situational information. Examples of features include the presence of subjective words, the count of intensifiers, and the use of slang and non-situational words. The authors showed that using these features with an SVM classifier outperformed the same model with BoW features. However, their proposed features are lexicon-based. Thereby, the classifier performance is highly reliant on the quality of lexicons. Besides that, creating such lexicons requires massive manual efforts, and they vary across languages and dialects. J. P. Singh et al. (2017) developed a system to classify flood-related posts as a high or low priority to identify victims who need urgent assistance. Using various linguistic features, they experimented with three ML models. As they focused on one type of disaster, it is unknown how the proposed approach generalises to other types of crises.

As ML models require handcrafted features, researchers leveraged DNNs as they automate the process of feature extraction and can exploit pre-trained text embeddings. Most of these studies used CNN to identify relevant posts (D. T. Nguyen et al., 2016; Burel et al., 2017b; Burel and Alani, 2018; Kersten et al., 2019), informative messages (C. Caragea et al., 2016; D. Nguyen et al., 2017; Burel et al., 2017a; Aipe et al., 2018; Derczynski et al., 2018; Ning et al., 2019) and information categories (D. T. Nguyen et al., 2016; Burel et al., 2017b; Burel and Alani, 2018; Madichetty and Sridevi, 2019a) from crisis events. The CNN architecture used in the reviewed studies is the one proposed by Y. Kim (2014). D. Nguyen et al. (2017) highlighted that CNN performed better than three classical ML approaches: logistic regression, SVM and random forest. They revealed that using crisis embeddings (Imran

et al., 2016), trained on tweets collected during crises, did not always perform better than general-domain embeddings but marginally improved the results on average. ALRashdi and O'Keefe (2018) agreed with the previous conclusion when experimenting with a CNN and Bidirectional Long Short-Term Memory ($Bi\text{-}LSTM$) using domain-agnostic Global Vectors for Word Representation ($GloVe$) embeddings and crisis word embeddings. They found that a Bi-LSTM with GloVe embeddings achieved the highest results. Going further, Neppalli et al. (2018) compared the performance of an NB classifier to two DNN models in identifying informative crisis-related posts. Their results demonstrated that CNN performed slightly better than the GRU model, and they outperformed the NB with different content-based and user-based handcrafted features. Similarly, Madichetty and Muthukumarasamy (2020) highlighted that DNN models outperformed an SVM with low-level lexical and syntactic features for identifying situational awareness tweets.

However, other research studies showed that ML models perform nearly as DNNs. A study conducted by Alam et al. (2019) showed that an SVM with TF-IDF features performed as good as a CNN model using crisis word embeddings in two tasks: event type classification and information type identification, and achieved competitive results on informativeness detection. Using different features, Burel et al. (2017b) and Burel and Alani (2018) concluded that SVM and CNN provide comparable results. They also demonstrated that the random forest algorithm produced competitive results. It is worth noting that different studies used different datasets and training settings (e.g., on-event vs cross-event and multi-source vs one cross-type crisis training set). Hence, different conclusions have been reached when comparing models' performance. There is a lack of comparisons between the proposed and best-performing models as there is no standard dataset on Twitter crisis detection. To the best of our knowledge, no study comprehensively compared classical ML approaches and DNNs of different architectures. Thus, we compare the generalisation ability of three widely adopted classical ML models and several architectures of DNNs (e.g. CNN and RNN) with

different configurations when performing multi-source cross-domain crisis detection. The models will be presented in Section 2.4. Besides that, our work consider two crisis-related tasks: relevancy detection and information categorisation.

Prior research also investigated the generalisation of various word embeddings. Nalluru et al. (2019) proposed to train two RNN architectures using a combination of domain-specific and generic domain-agnostic word embeddings to detect informative tweets during a disaster. They showed that training a different model for each embedding and ensembling their predictions achieved better performance than a model trained on the average of varying word embeddings. H. Li et al. (2018b) evaluated the generalisation of different word embeddings and sentence embeddings for Twitter crisis classification. Their results revealed that GloVe embeddings (Pennington et al., 2014) generalised better than others, and word-level embeddings generally outperformed sentence-level encodings.

Other studies have exploited contextualised text representations to identify crisis messages (Wiegmann et al., 2020; Liu et al., 2021), categorising them into different information types (G. Ma, 2019; Madichetty and Sridevi, 2020) or performed both tasks (Kozlowski et al., 2020). Madichetty and Sridevi (2020) used Embeddings from Language Models ($ELMo$) followed by a dense layer to classify data collected from three disasters into different information categories. Their proposed ELMo classifier outperformed the SVM and CNN models that use BoW and crisis word embeddings, respectively. However, the models have not been evaluated in the cross-event setting. G. Ma (2019) assessed different BERT-based architectures and demonstrated that they yielded approximately 3% higher accuracy than a Bi-LSTM model with GloVe Twitter embeddings. Kozlowski et al. (2020) showed that BERT-based models performed better than DNNs in the cross-event setting. Liu et al. (2021) demonstrated that the LSTM and logistic regression models performed better using BERT embeddings instead of word2vec on crisis relevancy detection and information classification tasks. In contrast, Madichetty and Muthukumarasamy (2020) showed that a CNN model had the worst per-

formance when using the BERT embeddings instead of other pre-trained word embeddings, including word2vec, Glove and crisis word2vec. The drawback of the work conducted by G. Ma (2019) and Liu et al. (2021) is evaluating their approaches once on an unseen randomly chosen set of data by splitting the whole multi-event dataset into training, validation and test sets. Such configurations do not mimic the real scenario where annotated data from a new crisis (or each type of disaster) is unlikely to be available. Wiegmann et al. (2020) experimented with a feed-forward neural network using BERT and Universal Sentence Encoder (USE) embeddings and found that a domain transfer across disaster types results in big performance drops. Our work also assesses the generalisation of different text representations, including word2vec, character embeddings and BERT embeddings. The text representations used in our experiments are highlighted in Section 2.5. We also explore the performance of various BERT-Based models. Unlike the work presented by G. Ma (2019), we perform our experiments in the cross-event setting, supposing that no labelled data is available from the target crisis.

Most studies on crisis classification have been limited to English SM messages. Relatively little research has considered other languages. Some researchers have focused on cross-lingual classification (Khare et al., 2018). Other studies have utilised multi-lingual word embeddings for different crisis-related classification tasks (Lorini et al., 2019; Torres, 2019; Ray Chowdhury et al., 2020). Other work has classified crisis tweets written in both English and Hindi (Rudra et al., 2018; Madichetty and Muthukumarasamy, 2020), Spanish (Cobo et al., 2015) and French (Kozlowski et al., 2020). There is very little published research on detecting Arabic crisis tweets. Alabbas et al. (2017) experimented with several classical ML classifiers (e.g., SVM, NB and decision trees) and an Artificial Neural Network (ANN) to extract high-risk flood-related Arabic posts. Adel and Y. Wang (2020b) proposed a feature-based approach to identify Arabic tweets related to famine, cholera and refugees. Alsudias and Rayson (2021) experimented with several ML and deep learning methods to

classify Arabic influenza and COVID-19 posts into multi-label categories.

The work considered Arabic crisis data has two limitations. Firstly, each study focused on one or two domains (crisis types), such as floods or pandemics. Secondly, they evaluated their models using small datasets, while DNNs require vast amounts of data to perform well. In our work, we will address the problem of identifying Arabic crisis posts from multiple types of disasters using a larger dataset in the multi-source setting, in which training data contains tweets from different events.

2.3.2 Crisis Detection using Domain Adaptation Methods

Supervised ML models assume that the source training and target test data are independent and identically distributed (sampled from the same distribution). Their performance drops (or is not guaranteed) on unseen target data if the distribution of that data differs from the source distribution. Domain adaptation methods are proposed to mitigate changes in distribution (domain/dataset shift) between the source and target domains.

As defined by Ramponi and Plank (2020), a domain (D) is denoted as $D = \{X, P(X)\}$, where $P(X)$ is the marginal probability distribution over that feature space X. A task (T), such as text classification, is denoted as $T = \{Y, P(Y|X)\}$, where Y is the label space, and $P(Y|X)$ is the conditional probability distribution learnt from the labelled training examples. The goal of domain adaptation is to enhance the generalisation ability of a trained model to a target domain if $P_S(X) \neq P_T(X)$ by learning on both source (D_s) and target (D_t) domains.

Domain adaptation is a special case of transfer learning called transductive transfer learning (Ruder, 2019). In domain adaptation, the source and target tasks T_s and T_t are the same, whereas the marginal probability distributions P_s and P_t differ across the source D_s

and D_t. The shift in the marginal probability distribution between the source and target domains is called a covariate shift (Ramponi and Plank, 2020). Several domain adaptation approaches have been proposed, including adversarial training, self-labelling, co-training, data selection and autoencoder-based models.

Domain adaptation can be categorised into two types: supervised and unsupervised. Supervised adaptation leverages a limited amount of labelled target data and more significant amounts of labelled source data. Unsupervised domain adaptation, which applies to most real-world scenarios, handles the domain shift by learning from labelled source data and unlabelled target domain. Knowledge can be transferred from a single source domain or multiple source domains. The latter is called a multi-source domain adaptation. In this research, we propose an unsupervised multi-source domain adaptation approach.

Previous research have adopted domain adaptation approaches to improve the generalisation of supervised models trained on past crisis data to classify unseen new crises (H. Li et al., 2018a; Alam et al., 2018a; Mazloom et al., 2019; Q. Chen et al., 2020). They showed that learning from both the labelled source and unlabelled target data is better than learning only from source labelled data. Several studies adopted an unsupervised domain adaptation approach using self-training. Their work showed that an iterative self-training improved the performance of a NB classifier (H. Li et al., 2015, 2017, 2018a). H. Li et al. (2018c) demonstrated that the self-training strategy outperformed a feature-based correlation alignment method. Mazloom et al. (2018) proposed a hybrid feature-instance domain adaptation method using matrix factorisation and the k-nearest neighbours algorithm to learn a NB classifier for the target event. The work was extended by Mazloom et al. (2019) who combined the feature-instance approach with the self-training algorithm presented by H. Li et al. (2017). H. Li et al. (2021) used self-training with CNN and BERT models and highlighted that self-training improved the performance of the DNN models. For retraining the base classifier, they used a soft-labelling strategy.

Alam et al. (2018a) proposed an approach based on adversarial training and graph embeddings in a single deep learning framework. The adversarial training minimises the distribution shift across domains, whereas graph-based learning encodes similarity between source and target instances. Krishnan et al. (2020) created a multi-task domain adversarial attention network based on a shared Bi-LSTM layer to filter Twitter posts for crisis analytics under domain shift. The tasks are relevancy, priority level, sentiment and factoid detection. Q. Chen et al. (2020) used a BERT-based adversarial model to classify tweets gathered during a disaster into different information categories. ALRashdi and O'Keefe (2019) proposed to use a distant supervision-based framework to label the data from emerging disasters. The pseudo-labelled target data is then used with labelled data from past crises of a similar type to train a classifier. X. Li and D. Caragea (2020) explored the use of the domain reconstruction classification approach on disaster tweets, which aims at reducing the covariate shift between source and target data distributions using an autoencoder. The authors showed that this approach outperformed the domain adaptation method proposed by Alam et al. (2018a).

We thought there was room for improvement by exploring some other techniques to improve our results. Thus, we contribute to this line of research on crisis detection using domain adaptation methods by adopting a selection-based approach that leverages pre-trained language models. Recent works on domain adaptation show that training on a domain similar to the target data results in performance gains for various NLP tasks. Ruder et al. (2017) explored the performance of several domain similarity metrics on different text representations for data selection in the sentiment analysis context. The authors also proposed a subset-level data selection approach that outperforms instance-level selection. In the same vein, Guo et al. (2020) studied different domain distance measures and presented a bandit-based multi-source domain adaptation model for sentiment classification. X. Ma et al. (2019) presented a domain adaptation method based on data selection and curriculum learning to fine-tune BERT models for text classification. Leveraging pre-trained language

model representations, Aharoni and Goldberg (2020) proposed data selection approaches for multi-domain machine translation using cosine similarity in embedding space. In this thesis, we adopt a data selection approach to train a classifier on the best matching data for the target emergency event instead of using all multi-event source data, as shown in Chapter 6.

2.3.3 Twitter Datasets for Crisis Events

In the publicly available Twitter crisis-related corpora in the crisis informatics literature, manually labelled datasets enable supervised machine learning techniques to extract messages of interest, including actionable information that contributes to SA. Such corpora can serve different purposes, ranging from historical data analysis to crisis forecasting.

Most of the published Twitter crisis datasets were written in English. Imran et al. (2013a, 2013b) built two annotated datasets labelled for two tasks: identifying informative messages that support awareness and assigning these to information types such as donations and cautions. The first dataset, ISCRAM2013, consists of tweets about the Joplin tornado, and the second comprises tweets collected during the Joplin tornado and Hurricane Sandy. One of the largest accessible and labelled crisis datasets is CrisisLex, which incorporates two collections: CrisisLexT6 (Olteanu et al., 2014) and CrisisLexT26 (Olteanu et al., 2015). CrisisLexT6 contains 60K English tweets from six emergency events. The messages were annotated by relatedness to the event in question (relevant vs not irrelevant). CrisisLexT26 includes tweets collected during 26 crises annotated in terms of informativeness (informative vs uninformative), information source and information type, and most subsequent studies have followed the CrisisLex taxonomies.

Imran et al. (2016) released CrisisNLP, a corpus of ∼52K labelled tweets collected during 19 crisis events between 2013 and 2015. The tweets were manually annotated by

volunteers and paid workers in terms of information type. Most messages in CrisisLex and CrisisNLP were written in English. Still, they contain tweets written in other languages, such as Italian, French and Spanish. Alam et al. (2018b) published CrisisMMD, a multimodal dataset of ∼16K English tweets with attached images gathered from seven natural disasters. Posts were labelled according to informativeness, information categories and damage severity. TREC-IS4[1] (McCreadie et al., 2020) provided Twitter datasets from 48 past emergency events, manually annotated by information types and priority levels. Alam et al. (2021a) published HumAID, a large human-labelled English Twitter dataset sampled from 19 disaster events.

There exist some labelled datasets written in other languages. Cobo et al. (2015) gathered ∼2K Spanish tweets from the Chilean earthquake in 2010, manually labelled by relatedness to the event. Cresci et al. (2015) published the SoSItalyT4 dataset, comprising ∼5.6K Italian tweets collected during four natural disasters in Italy, which were labelled for damage assessment. Kozlowski et al. (2020) built a dataset of ∼13K French tweets collected during various ecological crises. Alsudias and Rayson (2019) released a disease-related Arabic dataset of 1266 tweets tagged by the information source. In subsequent work, Alsudias and Rayson (2020b) collected Arabic tweets related to COVID-19 and influenza and manually labelled them according to an Arabic Infectious Diseases Ontology. Hamoui et al. (2020) presented the FloDusTA dataset, which includes ∼9k tweets from floods, dust storms, and traffic accident events. The messages were labelled by event type and time of occurrence (historical, immediate, future or irrelevant).

In the crisis informatics context, accessible unlabelled corpora have been utilised for various purposes, including prevalent topic extraction, social analytics and public sentiment assessment during crisis events. Such large published Twitter corpora can only be reused by reassembling the data from tweet IDs. Twitter's Developer Policy does not allow the

[1] http://dcs.gla.ac.uk/_richardm/TRECIS/

distribution of tweet contents for large-scale datasets[2]. Examples include 6M geo-tagged tweet IDs from Hurricane Sandy (H. Wang et al., 2015), ∼7M English tweets from Hurricane Harvey (Phillips, 2017) and 35M tweet IDs related to Hurricanes Irma and Harvey (Littman, 2017). Alam et al. (2018c) also created a Twitter corpus of more than 8M message IDs collected in 2017 from Hurricanes Irma, Maria and Harvey. Research has recently focused on COVID-19, and several studies on crisis informatics have published large-scale Twitter datasets collected during this pandemic. While some of these are limited to a single language such as English (Lamsal, 2020; Gupta et al., 2020) or Arabic (Alsudias and Rayson, 2020a; Alqurashi et al., 2020; Haouari et al., 2021a; Addawood, 2020), others have created multilingual datasets (E. Chen et al., 2020; Banda et al., 2021; Qazi et al., 2020; L. Singh et al., 2020; Alshaabi et al., 2021; Uniyal and Agarwal, 2021). Liu et al. (2020) released EPIC, an epidemic corpus comprising ∼30M tweets related to several diseases.

The thesis contributed to this body of research by creating and publishing a large-scale crisis-related Arabic Twitter corpus as we focused on Arabic crisis detection. To the best of our knowledge, there is no accessible large Arabic Twitter corpus of multi-crisis events. Alsudias and Rayson (2020b) only focused on diseases and their dataset is small. Hamoui et al. (2020) considered two natural disasters: floods and dust storms. Unlike our dataset, which was collected from crises that occurred in different Middle Eastern countries, Hamoui et al. (2020) limited their collection to tweets from Saudi Arabia. Thus, the tweets are expected to be written in the Gulf dialect. Two Arabic datasets created by Alabbas et al. (2017) and Adel and Y. Wang (2020a, 2020b) are unavailable to the research community. The former research focused on flood events, while the latter limited their data collection to two topics: famine and cholera in Yemen and refugees in Syria. Unlike these datasets, our corpus contains tweets from 22 crises involving different hazard types, and we annotated part of the corpus by relatedness to the event and information types.

[2] https://developer.twitter.com/en/developer-terms/agreement-and-policy

2.4 Supervised Machine Learning (ML) Models

Supervised learning is an ML task that learns a mapping function between a set of input and output variables given some annotated examples, and applying the mapping to make predictions about unseen data (Cunningham et al., 2008). We categorise the supervised classification ML models into two main types: classical ML and DNNs or deep learning. Deep learning is a subfield of ML that learns meaningful representations of input data through multiple successive neural layers or DNNs. Classical ML algorithms are called shallow learning as they do not provide such layered representations of data (N. K. Chauhan and K. Singh, 2018; Janiesch et al., 2021).

Feature engineering is crucial to classical ML models. Their performance can heavily depend on how the features are extracted and selected. Examples of such features include Term Frequency-Inverse Document Frequency (TF-IDF) vectors, Part-Of-Speech (POS) tagging (Janiesch et al., 2021) and the length of text. Thus, these learning methods may require human expertise and intervention for feature representation. Unlike classical ML algorithms, DNN models do not require handcrafted features as they can learn discriminative features automatically with their deep structures (Janiesch et al., 2021) but they need large training data to achieve good performance. Below, we review the classical ML and DNN models that we experimented with in this research.

2.4.1 Classical ML Models

2.4.1.1 Naïve Bayes

A Naïve Bayes (NB) is a probabilistic classification model based on applying Bayes' theorem. For a given instance χ, the classifier $h(\chi)$ predicts the class y as follows:

$$h(\chi) = argmax_y P(y|\chi) = argmax_y \frac{P(\chi|y)P(y)}{P(\chi)} \tag{2.1}$$

As the denominator is constant for different values of y, $P(\chi)$ can be ignored. Hence, the algorithm must calculate $P(y)$ and $P(\chi|y)$, called the prior and the posterior, respectively. While the prior can be estimated easily from a corpus, calculating the conditional probability $P(\chi|y)$ of an example for a given class could be very complex. Hence, an NB model assumes that features are conditionally independent given the class. Thus, the calculation of the conditional probability is simplified as follows:

$$P(\chi|y) = \prod_{i=1}^{n} P(x_i|y) \tag{2.2}$$

where x_i represents the ith feature in χ, and n is the number of features. Despite the unrealistic assumption of independence, the classifier achieves good performance on several NLP tasks (Rish, 2001). NB models differ according to the assumption they make regarding the feature distribution. For instance, the multinomial NB is commonly used for discrete multinomial distribution such as counts of words, while the Bernoulli NB is used with discrete binary features such as the presence of a word in a document.

2.4.1.2 Support Vector Machines

Support Vector Machines (SVM) (Boser et al., 1992) is a non-probabilistic binary classifier that is commonly applied for NLP classification and regression problems. The algorithm separates the instances of each class based on a decision boundary. The decision boundary is the hyperplane that maximises the margin between the two classes. Data points that specify the margin of largest gap between the categories are called support vectors. SVM predicts the new samples according to which sides of the margin they fall on. If data samples are not linearly separable in the initial finite-dimensional space, SVM maps the instances to points in a higher dimension feature space, where the data become linearly separable. The algorithm uses some kernel functions to avoid the complex calculations of points' coordinates in a high-dimensional space. Hence, kernel functions map points in the original space to the distances between these points in the new representation space (Cristianini and Scholkopf, 2002). SVM does not natively support multi-class problems, but can solve such problems using One-vs-One or One-vs-Rest approaches. The former divides a multi-class problem into one binary classification task per each pair of classes, while the latter divides the problem into one binary classification task per class.

2.4.1.3 Decision Trees

Decision trees (Breiman et al., 1984) are nonlinear predictive algorithms for classification and regression tasks. A tree is built through binary recursive partitioning. Decision trees learn decision rules inferred from the training instances by recursively splitting the dataset into subsets based on the instances' features. A tree's nodes represent some tests that examine the input variables. These nodes are linked by edges that determine the tests' outcomes. Starting from the root, the algorithm continues to evaluate the input features by following some branches until reaching a conclusion about the target variables, which are represented

by leaves. The following branches are chosen based on the test outcome of passed nodes. For classification trees, the output variables represent discrete classes. In regression problems, the output variables can take continuous or ordered discrete values.

2.4.2 Deep Neural Network (DNN) Models

DNN models learn complex relationships between input and output data using DNNs (N. K. Chauhan and K. Singh, 2018). A DNN is an ANN consisting of multiple layers. Text should be transformed into numeric tensors before being fed into the DNN models. The transformation can be done in different levels: words or n-grams of consecutive characters or words (Chollet, 2017). DNNs learn through their own errors by finding a set of model parameters that minimises a pre-defined loss function. Such deep models have been successfully applied to NLP. In the following, we provide an overview of two main categories of DNN architectures.

2.4.2.1 Convolutional Neural Network

A Convolutional Neural Network (CNN) (LeCun et al., 1998) is a special type of feedforward neural network. CNNs have proven successful in document classification tasks as they are able to extract position-invariant patterns from text sequences. Such learnt features boost the model's generalisation power. A CNN mainly comprises stacked convolutional and pooling layers. A convolutional layer applies a filter (known as a kernel) to each possible window of n words in the input sequences to create feature maps that represent the extracted patterns. Each filter is applied by performing a dot product multiplication between a window-sized patch of the input and the filter (a weight matrix). Pooling layers shrink the size of feature maps and overcome possible overfitting. Different pooling operations can be used for subsampling the feature dimensionality, including maximum pooling, average

pooling and sum pooling.

2.4.2.2 Recurrent Neural Networks

A Recurrent Neural Network (RNN) is a deep neural model that can handle input sequences of variable-length using a recurrent hidden state. It iterates through the input elements and keeps a state of information about what has been processed in prior time-steps. The activation of the recurrent hidden state (h_t) at each time depends on the previous hidden state, as shown in the following equation.

$$h_t = \sigma(W_h h_{t-1} + W_x \chi_t + b) \tag{2.3}$$

Where χ_t, h_{t-1} and b denote the current input, the previous hidden state and the bias, respectively. W_h and W_x are weight matrices; and σ is an activation function such as a logistic sigmoid or a tangent function. Nevertheless, standard RNNs cannot handle long-term dependencies as they may suffer from the vanishing gradient problem, which hinders learning from lengthy data sequences (Bengio et al., 1994). The gradients include information that is used to update the RNN parameters, and as the gradient gets smaller, the parameter updates become insignificant, implying that no learning occurs. To solve the vanishing gradient problem, various RNNs architectures have been proposed, as described in the following.

2.4.2.2.1 Long Short-Term Memory

Long Short-Term Memory ($LSTM$) networks were introduced by Hochreiter and Schmidhuber (1997) to address the long-term dependencies by incorporating a 'gate' into the recurrent unit. Variants of LSTM models have been proposed. The commonly used LSTM (Gers et al., 2000) has a memory unit that represents the cell state c_t and three gates: input gate i_t, forget gate f_t and output gate o_t. The LSTM unit can be illustrated as follows:

$$
\begin{aligned}
i_t &= \sigma(W_i h_{t-1} + W_i \chi_t + b_i), \\
f_t &= \sigma(W_f h_{t-1} + W_f \chi_t + b_f), \\
g_t &= \tanh(W_g h_{t-1} + W_g \chi_t + b_g), \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot g_t, \\
o_t &= \sigma(W_o h_{t-1} + W_o \chi_t + b_o), \\
h_t &= o_t \cdot \tanh(c_t)
\end{aligned} \tag{2.4}
$$

where h_t denotes the output of the LSTM unit and the dot operator represents the point-wise multiplication of vectors. The input gate determines what information is to be kept in memory, whereas the forget gate decides what information is to be discarded from the memory unit. The output gate determines the value of the next hidden state.

2.4.2.2.2 Gated Recurrent Unit

Gated Recurrent Unit (*GRU*) (Cho et al., 2014) is a simplified variant of LSTM as it has a smaller number of parameters. Thus, GRUs train faster than LSTMs. A GRU layer consists of two gates: an update gate z_t and a reset gate r_t. The update gate allows the model to determine how much information (from past time steps) is to be transmitted to the subsequent time step, whereas the reset gate controls how much of the previous information to forget. The following expressions illustrate how the GRU handles input sequences.

$$
\begin{aligned}
z_t &= \sigma(W_z h_{t-1} + W_z \chi_t + b_z), \\
r_t &= \sigma(W_r h_{t-1} + W_r \chi_t + b_r), \\
g_t &= \tanh(W_g (r_t \cdot h_{t-1}) + W_g \chi_t + b_g) \\
h_t &= (1 - z_t) \cdot h_{t-1} + z_t \cdot g_t
\end{aligned} \tag{2.5}
$$

2.5 Text Representation

Text should be encoded into a form that can be handled by ML models. Different vector-based text representations have been proposed to obtain machine readable encodings of variable-length input texts, while incorporating their semantic information. In this section, we reviewed the text representations that we experimented with in this research. In Vector Space Models (VSM), items (such as words, sentences or documents) are represented as vectors in a multi-dimensional semantic space, in which related items tend to occur in the proximity of each other (Pilehvar and Camacho-Collados, 2021). In other words, vector-based representations can capture the notion of semantic similarity.

2.5.1 Count-Based Representations

Count-based representations are commonly used in NLP and information retrieval. They are simple VSMs that represent text based on word occurrence and co-occurrence frequency. Bag of Words (BoW) is an example of count-based representations. It represents the text as a bag of its terms. BoW was proposed based on the assumption that the frequency of terms in a text is a good indicator of its relevance to a given query (Salton et al., 1975; Turney and Pantel, 2010). This model represent a text such as a sentence or a document as a vector, whose dimension is the entire vocabulary. Two pieces of text tend to have similar meanings if they share same words, taking into account their frequencies. The BoW model does not preserve the word order. Hence, frequencies of n-grams (usually $n \geq 3$) are usually used as features. The BoW model supposes that all terms play equal role in text representation. However, not all words in text are equally as important. Thus, variants of BoW models with different term weighting mechanisms have been proposed such as TF-IDF representation. The TF-IDF model weights down the terms that occur in most of the documents.

If the corpus is very large, the BoW representation would result in a sparse feature matrix as the vectors would contain many zeros. To solve this, feature selection and dimensionality reduction approaches can be applied. Such models ignore any syntactic or semantic relationships between terms or term groups. They do not capture the correlation between closely related words such as "crisis" and "crises". Despite these limitations, the BoW model has been successfully applied with classical ML approaches for text classification problems.

2.5.2 Pre-trained Embeddings

Transfer learning leverages data from other tasks or domains to improve the performance of learning models for the target task. Recent work in inductive transfer learning, in which source and target tasks differ, proposes methods to generate pre-trained text representations (embeddings) learnt from huge amounts of unlabelled text. Such representations are then utilised for downstream NLP tasks that have smaller labelled datasets.

Pre-trained embeddings are effective methods for encoding the semantic information in textual documents. The term 'embedding' refers to compact vector representations learnt using neural networks (Pilehvar and Camacho-Collados, 2021). The idea behind such representations is based on distributional hypothesis (Harris, 1954), which states that words that appear in same contexts tend to have similar meanings.

2.5.2.1 Word2vec

Word2vec (Mikolov et al., 2013) provides low-dimensional vector representations of words, learnt from large corpora in an unsupervised or self-supervised manner using a feedforward neural network model. The goal of word2vec is to map words that occur in similar contexts to the similar vector space.

Two word2vec models were proposed: Continuous Bag-Of-Words ($CBOW$) and Skip-gram. CBOW seeks to predict the target word given its surrounding words, whereas the Skip-gram model aims at finding the context words based on the target word. Both models are trained by iterating through the words in the corpus and consider a window of size n for each target word to make a prediction. Word2vec models generate a d-dimension continuous vector for each word, whose meaning can be inferred by its relation to other words. The representation can capture syntactic regularities in language, and find multiple degrees of similarity between words using simple algebraic operations.

Word2vec is a static text representation. It learns one vector for each token/word, which combines all different senses of the word. Besides that, word2vec models cannot effectively handle Out Of Vocabulary (OOV) as they are assigned random numerical vectors.

2.5.2.2 Character Embeddings

To tackle the problem of unseen words, character-level representations, which consider the morphological structure of words, have been proposed. Such models derive representations of words from their morphemes (Lazaridou et al., 2013; Botha and Blunsom, 2014) or constituent character n-grams such as FastText (Bojanowski et al., 2017). FastText is based on the Skip-gram model. It breaks words into character n-grams and learns their vector representations. Hence, it can infer the representation of an OOV word by summing the vectors of its constituent character n-gram. Bojanowski et al. (2017) evaluated their proposed text representations on word similarity and analogy tasks. FastText showed significant improvement for morphologically rich languages over word-level representations that do not consider subword information. The authors also demonstrated that FastText outperformed other character-level methods relying on morphological analysis. As with word2vec, FastText embeddings do not consider the context. Some words can have similar constituent n-grams

but different meanings (Pilehvar and Camacho-Collados, 2021).

2.5.2.3 Contextualised Embeddings

Contextualised embeddings consider the surrounding words when generating the representation for a given target word. Unlike the static word2vec representations, a word in contextualised embeddings can have varied representations based on its context. Contextualised embeddings can be categorised into two architectures: RNN-based and Transformer-based models. Transformer-based models have the advantage of capturing the distant contexts of a target word due to their self-attention mechanisms (Pilehvar and Camacho-Collados, 2021). In our work, we utilised one of the Transformer-based models: Bidirectional Encoder Representation from Transformers ($BERT$) (Devlin et al., 2019) embeddings.

BERT is a deep bidirectional Transformer encoder that was developed by researchers at Google AI Language. It uses a Masked Language Model (MLM) pre-training objective to produces context-sensitive embeddings. The MLM is trained by randomly replacing some of the input tokens with a special token, and then predicting those masked tokens based on left and right context words. BERT is also pretrained using a next sentence prediction task to enhance performance for certain tasks, such as Question Answering (QA), that require learning the relationships between two sentences. BERT is trained to minimise a loss of a linear combination of both MLM and next sentence prediction. Prior to training, words are segmented into subword tokens using a tokenisation algorithm such as WordPiece tokeniser. Generating embeddings for subwords instead of words can decrease the size of the vocabulary and allow the model to deal with OOV words (Pilehvar and Camacho-Collados, 2021).

BERT was released in two variants: Base and Large. The former uses 12 encoder layers with a hidden size of 768, whereas Large version has 24 layers of encoder with a hidden size of 1024. The BERT model can be fine-tuned on downstream NLP tasks with

minimal modifications to its architecture, or its embeddings can be extracted from one or more of the layers and be leveraged as input features for task-specific models. Recently, BERT has achieved state-of-the-art results on a wide variety of NLP tasks.

2.6 Conclusion

This chapter presented the relevant literature. It discussed the related work, including the crisis classification techniques, the domain adaptation approaches adopted in the crisis informatics field and the developed Twitter crisis-related datasets. The chapter also reviewed the models and text representations we experimented with in this study. Section 2.4 described several classical ML and DNN models. The last section overviewed different text representation techniques, including count-based representations and pre-trained embedding models. In the subsequent chapter, we present our research methodology.

Chapter Three

Methodology

This chapter describes our research methodology. Section 3.1 summarises the Twitter dataset creation process and highlights the used topic identification technique. The second part of the chapter is related to crisis tweet classification. It first describes the methodology followed to evaluate the supervised classifiers for crisis tweet classification. Then, it briefly presents our proposed domain adaptation approach.

3.1 Building the Dataset

3.1.1 Overview of the Dataset Creation Process

The first contribution of this research is creating a gold-standard Arabic Twitter dataset that we used for crisis detection. In our data collection, we considered high- to medium-risk crises that are most likely to trigger notable Twitter activity. These crises left several people displaced or dead and resulted in substantial property and infrastructure damage. We focused more on flooding crises, as floods frequently occur in the Middle East between October and December after heavy rain and subsequent flash flooding. For example, in October and November 2018, heavy rainfall caused severe flooding in various Middle Eastern

countries, including Saudi Arabia, Kuwait, Jordan, Qatar and Iran[1]. According to civil defence authorities in Saudi Arabia, 1,480 individuals were rescued, 30 died and 3,865 were evacuated during floods that occurred in the period between 19 October and 14 November[2]. In Jordan, the flash flood on 9 November left at least 12 people dead and 29 injured[3]. On the same day, Kuwait had heavy rain that resulted in infrastructure and property damage, leaving at least one person dead[4]. Such risky and impactful disasters are included in our corpus. We learnt about events from Twitter trending topics and news.

The corpus data were collected iteratively using the Twitter search API[5] by tracking selected keywords and hashtags used as query terms. The list of query terms was updated frequently to include new relevant hashtags found in the collected data. The data collection process will be described in detail in the following chapter. We also provided quantitative and qualitative analysis of the collected conversations during crises on Arabic Twitter. Analysing tweets about emergency events offers an opportunity to understand their characteristics, thereby making appropriate decisions based on the quality (usefulness) of shared information. Looking at the conversation size provides insight into the attention a topic receives. We found that more than half of the tweets in the corpus were duplicates, indicating that crisis posts obtain a high number of shares and receive considerable attention.

We removed the duplicates and identified the main themes of discussion (information types) at the event and message levels to answer the first research question. We used word cloud and the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic modelling technique. The LDA model will be briefly described in the following section. Investigating the main topics helps to compose accurate annotation instructions and examine whether the dataset includes valuable information that can be used for situational awareness. In our

[1] https://floodlist.com/asia/flooding-iran-iraq-and-kuwait-november-2018
[2] https://sabq.org/saudia/jgvvgz
[3] https://floodlist.com/asia/jordan-flash-floods-november-2018
[4] https://floodlist.com/asia/jordan-flash-floods-november-2018
[5] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

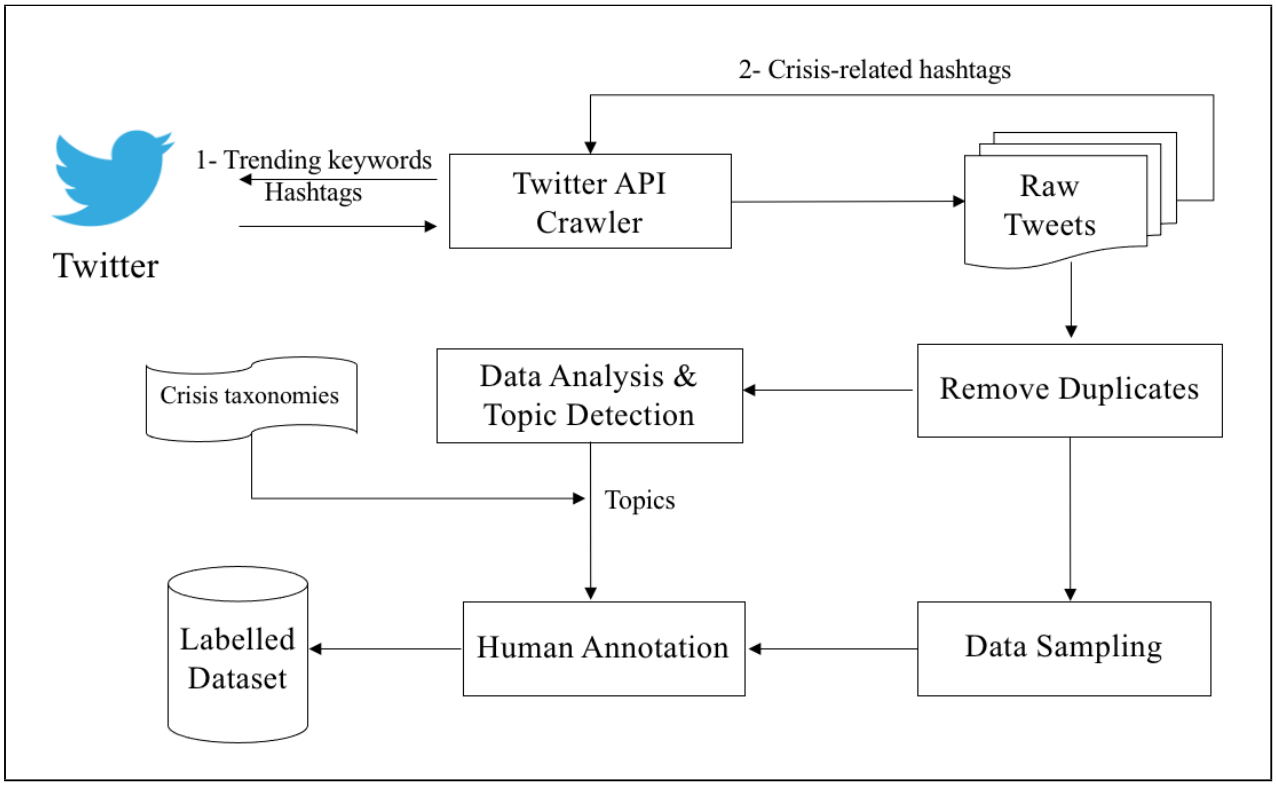

Figure 3.1: Dataset creation scheme

work, tweets belonging to each topic were manually analysed. We proposed a multi-label annotation scheme in light of extracted topics and taxonomies presented in previous studies to categorise messages into different information types (Olteanu et al., 2015; Imran et al., 2016; Sit et al., 2019). We asked volunteers to label samples of data from seven emergency events. In this study, 21 native Arabic speakers participated in data annotation. The coders received their primary and secondary education in an Arabic-speaking country, including Saudi Arabia, Egypt and Syria. All annotators are older than 20 years. The author and one volunteer analysed and explored the topics, while the 21 participated in data labelling. They first decided whether the tweet was related to the crisis and then selected the appropriate information type the tweet communicates. The dataset serves as a gold standard for an Arabic Twitter crisis detection task. Figre 3.1 summarises the labelled dataset creation process, which will be elaborated in Chapter 4.

3.1.2 Latent Dirichlet Allocation

We were inspired by previous studies that used LDA models to analyse SM emergency events (Kireyev et al., 2009; Sit et al., 2019; Karami et al., 2020b; Alam et al., 2020). Using data from Hurricane Irma, Sit et al. (2019) showed that LDA could be used to identify latent fine-grained categories of tweets such as damages, warnings and critiques. Karami et al. (2020b) employed the LDA technique to discover the negative public concerns during the 2015 South Carolina flood. Alam et al. (2020) utilised the LDA modelling to unfold topical patterns over time. They demonstrated that LDA-generated topics disclosed public issues during emergencies.

LDA is a generative probabilistic model commonly used to uncover hidden themes (latent topics) in text collections. The model makes three main assumptions: a BoW, exchangeability for documents in a corpus and the known number of topics (Blei et al., 2003). LDA assumes that all the documents in the corpus share the same topics but with different proportions (Blei, 2012). Hence, a document can belong to multiple topics, where each topic is represented by a distribution over a fixed vocabulary. The LDA's generative process defines a joint probability distribution over the observed variables (words of the documents) and hidden variables (topic structure). The joint distribution of the variables is equivalent to the following:

$$P(\beta_{1:k}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{n=1}^{N} P(Z_{d,n}|\theta_d)P(W_{d,n}|\beta_{1:k}, Z_{d,n}) \text{ (Blei, 2012)} \tag{3.1}$$

$\beta_{1:K}$ denotes the k topics, where each topic is a distribution over words. The topic mixture for document d is θ_d, where $\theta_{d,k}$ shows the topic proportion for topic k in the dth document. The topic assignments for document d are Z_d, whereas $Z_{d,n}$ represents the topic assignment for the nth word in d. Such assignment depends on the document-level topic

proportions θ_d. The set of words for the dth document are W_d, while $W_{d,n}$ is the nth word in that document.

LDA infers the hidden topics by using the joint distribution to calculate the conditional (posterior) distribution of the hidden variables given the documents:

$$P(\beta_{1:k}, \theta_{1:D}, Z_{1:D}|W1:D) = \frac{P(\beta_{1:k}, \theta_{1:D}, Z_{1:D}, W1:D)}{P_(W_{1:D})} \text{ (Blei, 2012)} \tag{3.2}$$

3.2 Crisis Tweet Classification

3.2.1 Evaluation of Supervised Models

Models are challenged by out-of-distribution learning when classifying a new crisis using data from other crisis events due to their different features, such as location names. The generalisation performance of classical ML models depends mainly on how the features are extracted and selected. On the other hand, DNNs do not require feature engineering because their deep structures allow them to learn discriminative features automatically. DNN techniques outperformed ML on many Arabic Twitter classification problems such as sentiment analysis (Nassif et al., 2021), emotion detection (Baali and Ghneim, 2019), and hate speech recognition (Al-Hassan and Al-Dossari, 2022). Nevertheless, such models need vast amounts of training data to learn and generalise well, which is expensive to collect and annotate. Our dataset might not have been of adequate size for achieving good performance with DNNs. Thus, we ask the second research question to explore the generalisation ability of supervised classical ML and DNNs when classifying new crises using data from historical events. We investigate which models are most suitable for Arabic Twitter crisis classification tasks and explore whether different pre-trained embeddings can enhance the models' generalisation.

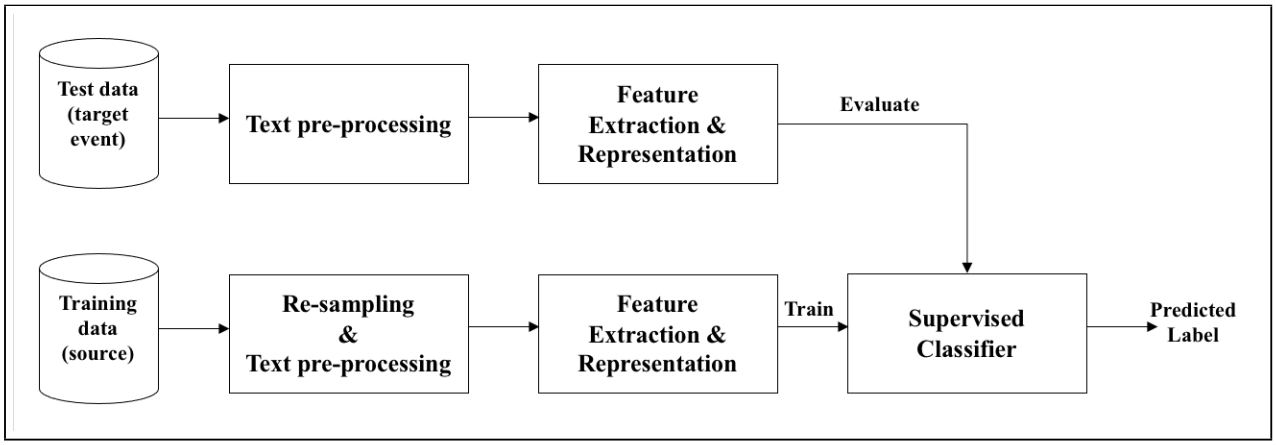

Figure 3.2: Pipeline for crisis tweet classification

To answer the second research question, we conducted several experiments using different classical ML classifiers and DNNs to perform cross-domain crisis detection. We experimented with three models for the classical ML: NB, SVM and decision trees. They were selected among other classical ML approaches because they were widely adopted for crisis classification tasks, particularly the NB and SVM (Verma et al., 2011; Imran et al., 2013a; Ashktorab et al., 2014; Burel et al., 2017b; Alabbas et al., 2017; Alam et al., 2019). Besides that, Alabbas et al. (2017) showed that they produced good results when classifying in-domain (i.e. flood-related) Arabic tweets. We also experimented with a random forest classifier (as it was also widely used) and found that it produced similar or sometimes lower performance than decision trees. Thus, we decided to stick to the three ML models. Regarding the DNNs, we evaluated different DNN architectures, including CNN, LSTM and GRU. We also assessed the performance of convolutional LSTM, which combines the CNN and LSTM. In this model, the output of the CNN was fed into an LSTM layer. The CNN learns the most significant high-level features from the input text, while the LSTM has the advantage of processing these features in sequence. The models were described in Section 2.4. Figure 3.2 outlines the classification framework.

Our dataset, as shown in Chapter 4, has skewed class distribution. Learning from

imbalanced datasets negatively impacts the classification process (Ali et al., 2019). The majority classes bias the classification models towards themselves, resulting in poor performance and misleading accuracy results. The classifiers would predict the dominant category and ignore the minority or underrepresented class. Thus, we first performed re-sampling techniques to balance the training data and alleviate problems associated with data skew when evaluating different supervised models. We also reported the performance using appropriate metrics, such as the Macro-averaged F1 that considers the unweighted mean for each class.

After data re-sampling, we pre-processed the text as we will describe in Chapter 5. We used word n-gram and character n-gram features with classical ML classifiers. As for DNNs, we used pre-trained word2vec (CBOW) embeddings. We also experimented with FastText character embeddings as they perform well for morphologically rich languages such as Arabic (Bojanowski et al., 2017). We also investigated whether DNNs can perform well without using pre-trained word embeddings that transfer the knowledge (text representation learnt from massive unlabelled data) to downstream tasks by allowing the models to learn word embeddings from the training dataset. Finally, we experimented with contextualised embeddings. We conducted two experiments with the BERT models. In the first experiment, the embeddings were fine-tuned during training with a linear classification layer on top of BERT. In the second experiment, we combined BERT with the best-performing DNN architecture and explored whether they enhanced the results. In other words, the DNN (i.e. CNN or RNN) was stacked on top of BERT, so that the BERT representations from the final hidden state were fed into the DNN as features. Text representations were described in Section 2.5.

As mentioned in the introduction, we performed two supervised crisis-related classification tasks: relevancy detection and information categorisation. The former is a binary classification problem, while the latter is a multi-label problem. We found that training a classifier for each task and performing two subsequent classification steps as depicted in

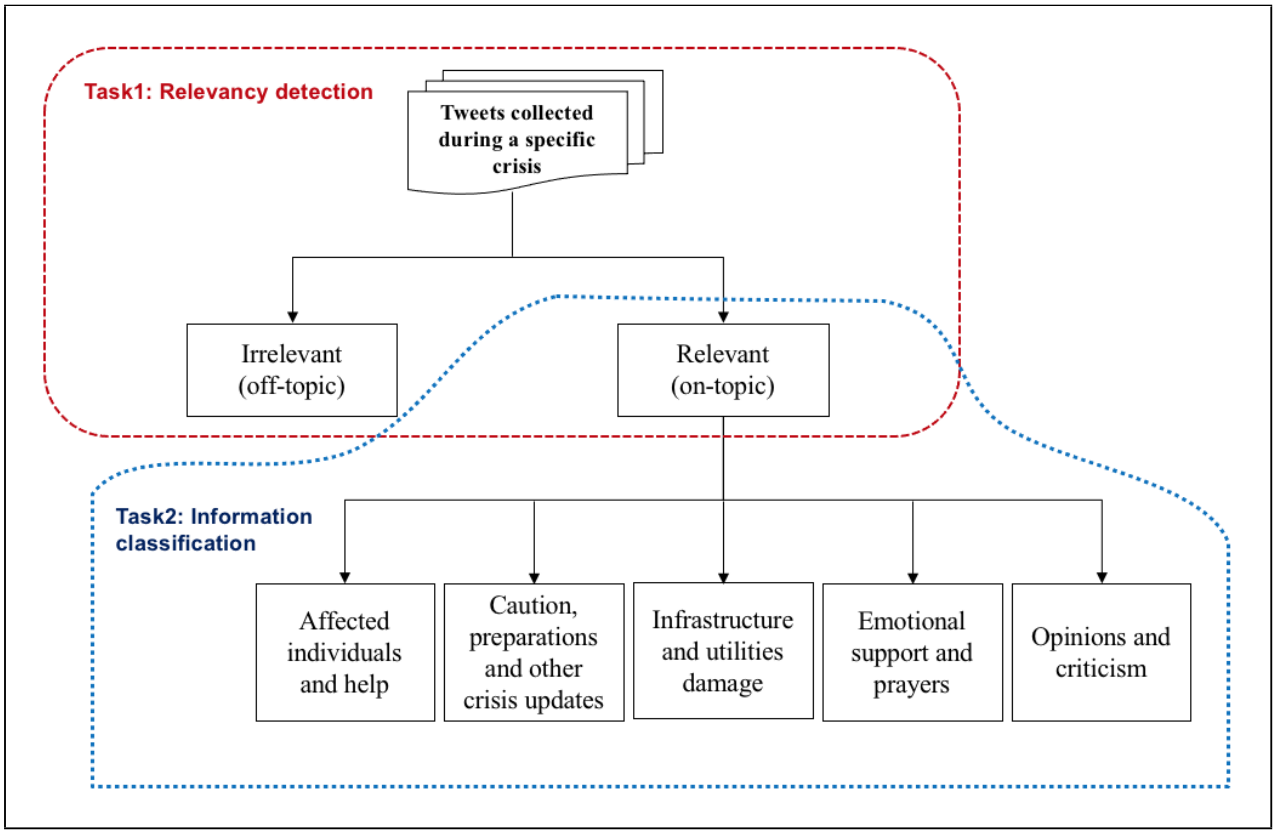

Figure 3.3: Flowchart for the two crisis detection tasks: relevancy detection and information categorisation

Figure 3.3 instead of one that filters out irrelevant posts while categorising messages results in more accurate performance, particularly in identifying spam and opinions. By training a classifier for relevancy detection, we also avoid tagging a message as irrelevant with another informative class because we have a multi-label dataset. In this study, the described models will be evaluated for the two tasks.

In order to mimic a real scenario, we assume we are given labelled data from historical crises (multi-source/multi-event datasets) and only unlabelled data from an emerging emergency event, representing the source and target set, respectively. Thus, we evaluate the classifiers' performance using the leave-one-event-out setting. In this setting, the evaluation is performed by choosing one target event as the test set and the remaining events for training. If we have four crises: A, B, C and D, we will perform four experiments. In each

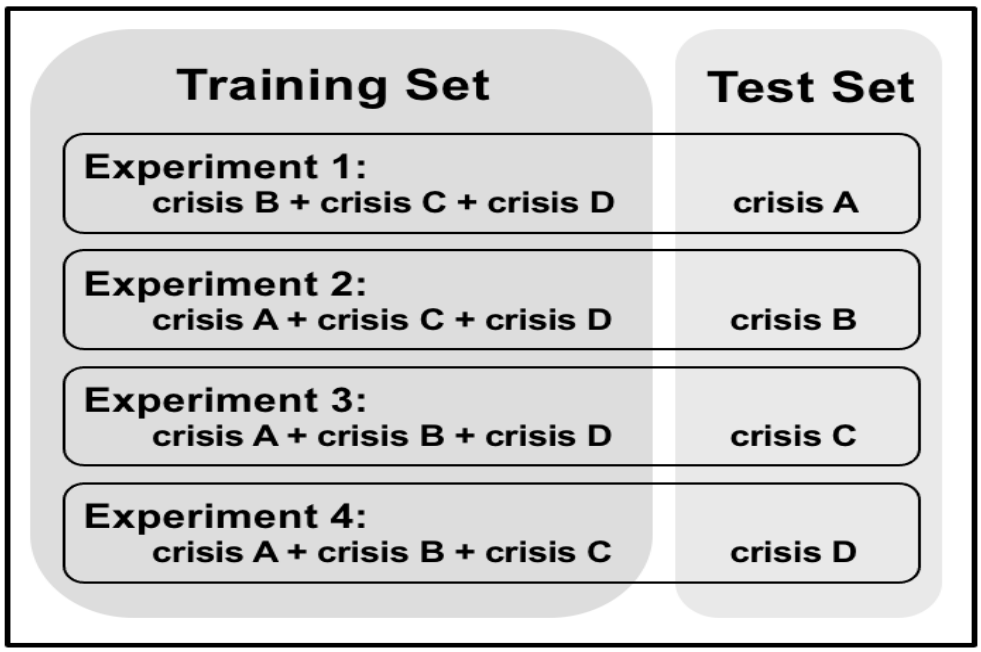

Figure 3.4: Leave-one-event-out evaluation setting

experiment, one event is left out for testing, and the other events are used for training. For instance, if crisis A is selected as a target event, the source data will include B, C and D. The leave-one-event-out strategy is illustrated in Figure 3.4. The details of the experiments, including the models' settings, pre-processing techniques, the used embeddings and results, will be presented in Chapter 5.

3.2.2 Data Selection Approach

One of the main contributions of this thesis is proposing a domain adaptation technique to enhance the models' performance for cross-event classification. In light of the evaluation results from the previously described experiments, we adopted a data selection technique to train the model on a subset of multi-event source data that is most similar to the target crisis. Training a classifier using examples that are dissimilar to the target data can adversely affect the model performance.

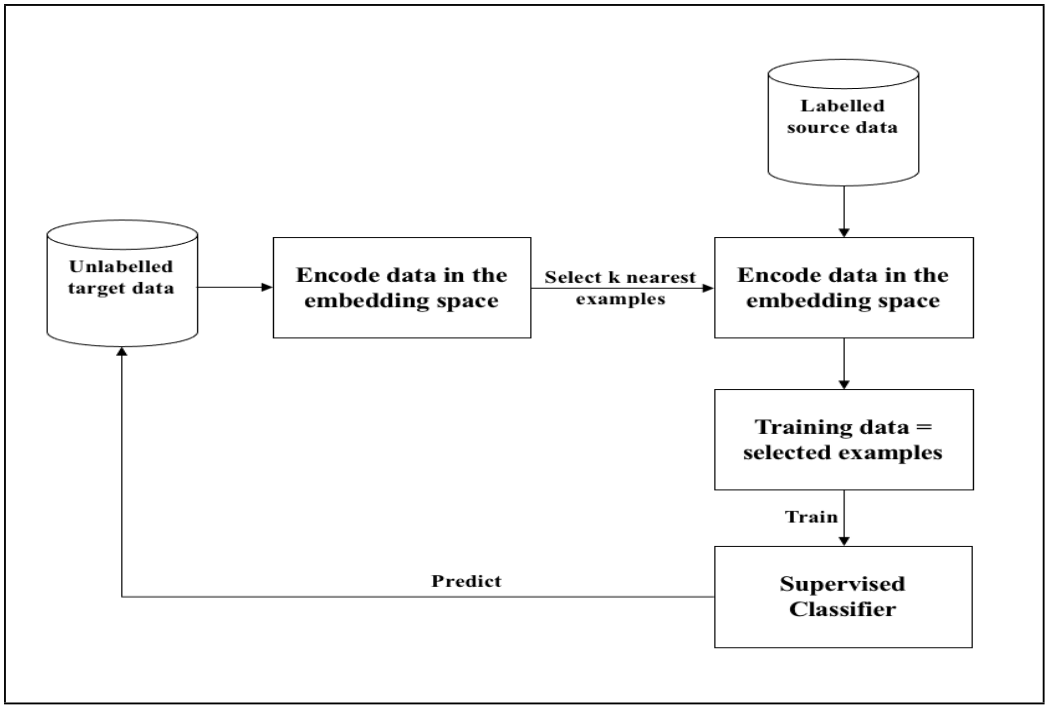

Figure 3.5: A framework of the data selection approach

We exploited the contextualised representation produced from pre-trained language models to encode our data and select the most similar instances based on the document similarity in the embedding space. In this work, we used the K-nearest neighbours algorithm for data selection. For each tweet in the target data, the algorithm selects the (k) closest (most similar) documents from the source data and adds them to the training set. Hence, our adaptation is an instance distance-based data selection method. Finally, the selected subset is used to train the best-performing models for each task. To answer the third research question, we compare the performance of our selection adaptation models to the equivalent models that learn from all available source data. The evaluation was conducted on the two crisis-related tasks using the same leave-one-event-out evaluation strategy. Figure 3.5 outlines the data selection method. Our presented adaptation strategy is unsupervised as it does not require any labelled instances from the target domain. It can also be utilised during the early hours of a disaster when small unlabelled data is available from the target event.

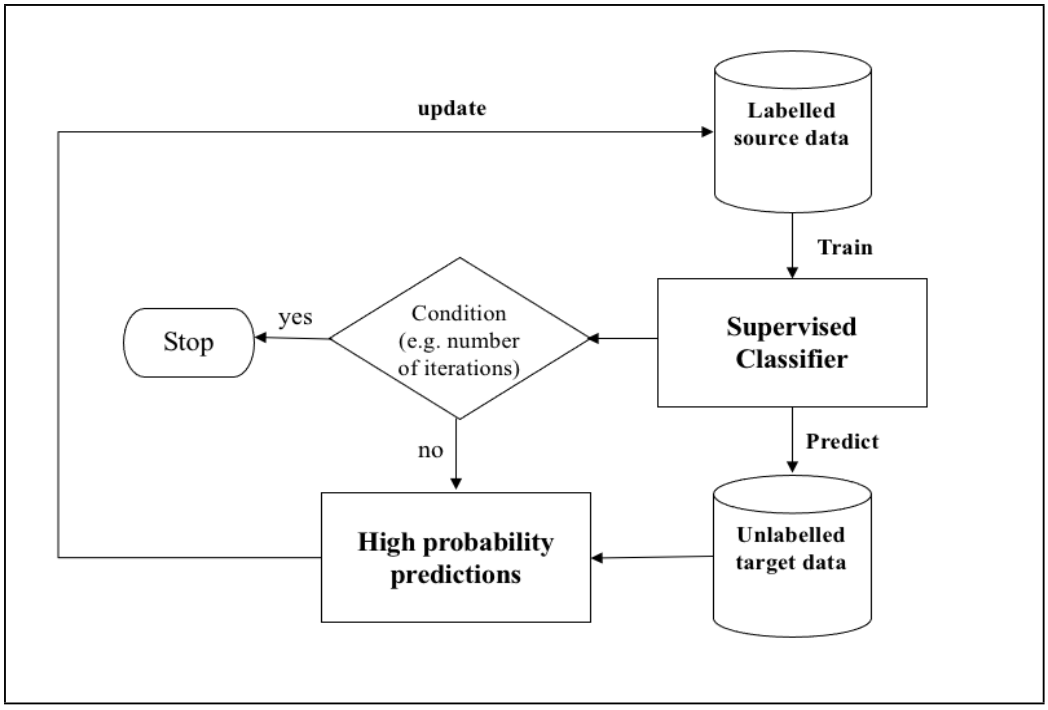

Figure 3.6: The self-training adaptation approach

We also explored how the results of the data selection method compare to the results of a self-training adaptation approach (**RQ4**). Self-training is a semi-supervised learning approach that learns a base classifier from source data and then uses that classifier to label the unseen target data. Predictions with high probability are added to the training data. The classifier is then retrained by utilising the source and pseudo-labelled target data. The self-training continues for a fixed number of iterations or until convergence. Figure 3.6 depicts the self-training technique. Self-training produced good results for English Twitter crisis detection (H. Li et al., 2021). Finally, we explored the effect of combining the data selection approach with self-training to answer the last research question (**RQ5**). Chapter 6 will present the selection method in detail, along with the configuration for the self-training approach. It will also show the experiments' results to answer the presented research questions.

3.3 Conclusion

This chapter highlighted the research methodology. It began by providing a summary of the dataset creation process. Then, it presented our classification pipeline and described how we evaluated the classifiers to identify crisis tweets and categorise them into different information types. Finally, the chapter provided an overview of the instance-based data selection approach, which is the last contribution of this thesis. The subsequent chapter explains how we collected, analysed and annotated the crisis-related dataset.

Chapter Four

Kawarith Twitter Corpus

This chapter introduces our Arabic crisis-related Twitter corpus, named Kawarith. The chapter also presents a quantitative and qualitative analysis of the corpus content, including investigating the main topics and information categories of conversations posted during a range of crisis events. These information types were used to label a subset of the data. The labelling process will be described in Section 4.6. This chapter was published in our paper titled: "Kawarith: an Arabic Twitter corpus for crisis events" (A. Alharbi and M. Lee, 2021).

4.1 Crisis Events and Data Collection

The Kawarith corpus comprises Arabic tweets from 22 crisis events that occurred between October 2018 and September 2020. Kawarith encompasses a wide range of hazard types, including floods, bombing, shootings, wildfires, pandemics, sandstorms and explosions. Table 4.1 lists these crises by date; flood events occurring in the same area are referenced by location and year of occurrence. The corpus is expected to include tweets written in different dialects because these crises occurred in various Arabic speaking regions. Previous studies have revealed that Arabic dialects are strongly present in SM (Alsarsour et al., 2018). However, many messages, especially those sent from news and organisation accounts, are written in MSA.

As we mentioned in the last chapter, we collected the corpus data iteratively using the Twitter search API. First, we observed the news accounts and trending topics in the Middle East using Twitter API to learn about new crises. During each crisis, we began by using trending crisis-related hashtags or keywords as query terms. We noticed that most crisis-related trends are the crisis location, such as Kuwait or the hazard type, such as floods. Hence, if no relevant trends were found during this initial data collection phase, we used the API to search Twitter using a logical *AND* combination of the terms *hazard type* and *crisis location*. For example, we used the query (سيول AND عدن) "Aden AND floods" to begin gathering data for Aden floods. Additionally, as an alternative search term, we linked the two terms in hashtag form, as we observed that people tended to use crisis-related hashtags like #سيول_الأردن "#Jordan_floods" and #جائحة_كورونا "#corona_pandemic" for the Jordan floods and Corona Virus Disease 2019 (*COVID-19*) crises, respectively. This first step led us to crawl an initial set of Twitter messages, which was manually inspected to identify any new hashtags that related strongly to the event. The dataset was then expanded by tracking these hashtags, and this step was repeated until no new relevant hashtags could be found. Finally, we updated our query to include all manually selected keywords linked by logical *OR* to extract crisis messages in the following timeframe, which we set to 24 hours. Concurrently, we updated the query with any new relevant keywords emerging as trends on Twitter. Keywords can be in the form of hashtags, phrases or single words. The data collection process is illustrated in Figure 4.1. Examples of used query terms are listed in Table 4.2.

We adopted a cautious approach to keyword selection, often using event-specific terms such as relevant named entities rather than general hazard descriptors like أمطارغزيرة "heavy rain" to reduce false positives, especially for flood events, which usually occurred simultaneously. Terms such as country name hashtags were generally disregarded, especially if the event had little impact on that country. The decision to use such terms as search queries was generally based on recently retrieved tweets; a candidate term was added to the query if it retrieved event-relevant messages. Importantly, as we followed a keyword-based collection, tweets that did not include the query terms were missed. However, we are satisfied that our data captured the main aspects of the crises.

For COVID-19, we tracked only nine keywords referring to the event by name because the event has triggered many other topics (such as conspiracy theories and the world economy) that were not immediately relevant to our purposes. As our study focused on building an Arabic dataset, data collection was confined to tweets that Twitter tagged as Arabic, and this language parameter necessarily excluded tweets in other languages. In Lebanon, for example, people also tweeted in Arabizi (Romanised Arabic), English and other languages, which may account for the relatively small volume of Arabic data crawled for those events despite their severity and impact.

Data collection continued from the first day of a crisis until the end, which we chose to define as the point at which it no longer triggered conversations on Twitter and related keywords no longer appeared in the Twitter trending list for that geographical area. We treated long-term crises like the COVID-19 pandemic as exceptions to this rule. In the case of COVID-19, data collection was delayed until near the peak of the epidemic in the Middle East. The goal was to obtain representative rather than comprehensive samples. COVID-19 has lasted for a long time, and collecting data during the epidemic's peak will result in crawling more relevant tweets, such as those about new cases. Lists of the query terms and collection dates have been included in the published data. In total, we collected 1,658,795 unique tweets from 22 crisis events. Apart from COVID-19, which was global, the crises were specific to eleven different countries[1], as displayed in Table 4.1.

4.2 Tweet-related and User-related Statistics

The Twitter search API supports search of tweets published in the previous seven days[2]. As tweets matching different queries within the same timeframe might be captured on multiple occasions during the iterative collection process, we removed redundant posts (ID-based duplicates) from the corpus and retained only messages with unique IDs. Table 4.1 displays the number of unique tweets

[1]Dragon storms have affected several countries, but we focused our collection on the Egyptian Twitter content.

[2]https://developer.twitter.com/en/docs/twitter-api/search-overview

Table 4.1: List of crises sorted by date, with tweets and users statistics

Year	Crisis name	Country	Start date	# of tweets	# of unique authors	# of tweets by verified users
2018	Jordan floods	Jordan	25/10/18	8493	5376	452
2018	Kuwait floods-18	Kuwait	04/11/18	34315	20285	637
2018	Qurayyat floods	Saudi Arabia	10/11/18	9731	6781	176
2018	Hafr Albatin floods-18	Saudi Arabia	14/11/18	6069	4218	105
2018	Leeth floods	Saudi Arabia	23/11/18	9596	6170	99
2019	Khartoum massacre	Sudan	03/06/19	12305	6811	50
2019	Cairo bombing	Egypt	04/08/19	2018	1320	182
2019	Lebanon wildfires	Lebanon	13/10/19	8585	5907	100
2019	Egypt floods	Egypt	21/10/19	10938	4138	51
2019	Hafr Albatin floods-19	Saudi Arabia	25/10/19	14546	8398	120
2019	Karbala massacre	Iraq	28/10/19	11961	6593	328
2019	Dubai floods	United Emirates	10/11/19	2480	1983	75
2019	COVID-19	Worldwide	01/12/19	775169	345381	16295
2019	Lebanon floods	Lebanon	09/12/19	8415	5272	148
2019	Kuwait floods-19	Kuwait	15/12/19	25491	15566	312
2020	Dragon storms	Egypt	12/03/20	92014	49037	1479
2020	Aden floods	Yemen	21/04/20	37019	10638	147
2020	Oman floods	Oman	30/05/20	80673	25240	755
2020	Ta'if floods	Saudi Arabia	24/07/20	25424	13524	69
2020	Beirut explosion	Lebanon	04/08/20	307795	158427	7584
2020	Syria wildfires	Syria	03/09/20	22632	15162	167
2020	Sudan floods	Sudan	04/09/20	153126	96257	815
Total				1,658,795	812,484	30,146

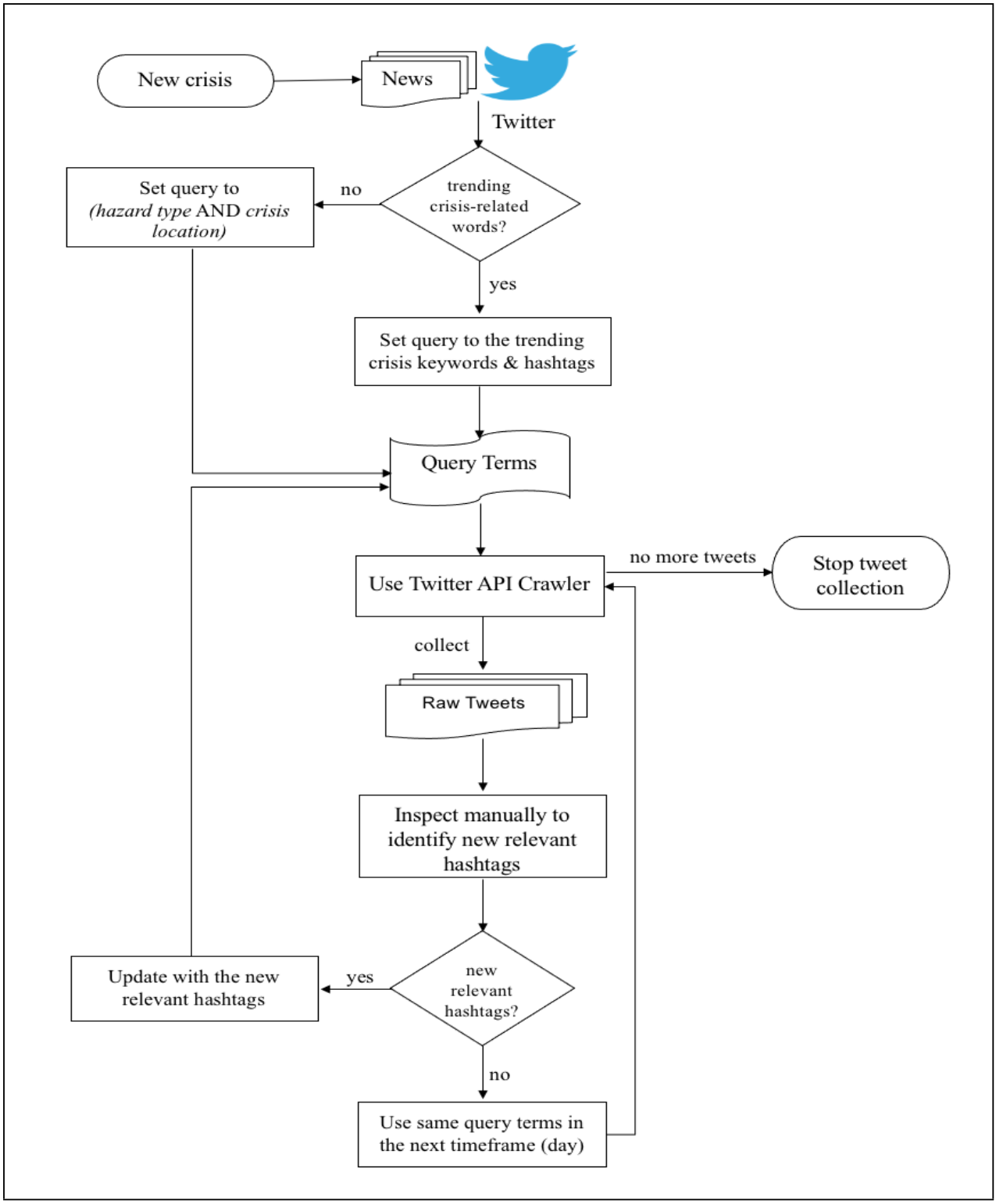

Figure 4.1: Data collection approach

Table 4.2: List of crises with examples of query terms used to collect the data

Crisis name	Examples of queries
Jordan floods	#سيول_الاردن ، #سيول_البحر_الميت
Kuwait floods-18	#الكويت_تغرق ، #امطار_الكويت
Qurayyat floods	#سيول_القريات ، #قرية_غطي
Hafr Albatin floods-18	#مفقودي_شعيب_فليج ، #مفقود_شمال_الحفر
Leeth floods	#سيل_وادي_الليث ، #الليث_تغرق
Khartoum massacre	#مدن_السودان_تنتفض ، #اعتصام_القيادة_العامة
Cairo bombing	#معهد_الاورام ، #انفجار_المنيل
Lebanon wildfires	#لبنان ، #لبنان_يحترق
Egypt floods	#مصر_بتغرق ، #مصر_غرقت
Hafr Albatin floods-19	#حفرالباطن_الان ، #امطار_حفرالباطن
Karbala massacre	#كربلاء_تستغيث ، #كربلاء_تباد
Dubai floods	#امطار_الامارات ، #دبي_تغرق
Coronavirus disease	#كوفيد١٩ ، #جائحة_كورونا
Lebanon floods	#لبنان_يغرق
Kuwait floods-19	#مطر_الكويت ، #سيول_الكويت
Dragon storms	#منخفض_التنين ، نهر النيل
Aden floods	#عدن_تغرق_بسيول_الامطار ، #انقاذ_عدن
Oman floods	#منخفض_محافظه_ظفار ، #عمان_مستعدة
Ta'if floods	امطار الطايف ، #الشفا
Beirut explosion	#انفجار_بيروت ، #انفجار_المرفأ
Syria wildfires	#سوريا_تحترق ، #حرائق_سوريا
Sudan floods	#السودان_فيضان٢٠٢٠ ، #السودان_يغرق

and unique authors for each crisis, along with the number of tweets sent by verified accounts. In total, only ∼1.8% of messages were generated by such accounts, indicating that few crisis-related tweets were published by public interest accounts (e.g. media, government) which are typically verified.

We found a strong Pearson correlation of 0.76 between tweets posted by verified accounts and those that included URLs. The Cairo bombing returned the highest percentages of both (9.02% of tweets sent by verified accounts and 24.43% of tweets containing URLs). This suggests that many of the tweets related to this event were generated by authentic news accounts rather than by the general public, who might have little to share about an instantaneous and focalised event of this kind. On average, only 7.9% of the corpus tweets include URLs. Overall, the corpus contains 40175 unique links, excluding links pointing to other posts in quote tweets.

4.3 Content Redundancy

We explored the amount of duplicated content in the corpus. Two tweets were considered content-based duplicates if they exhibited a matching sequence of tokens (words or emojis). To identify duplicate content, we first cleaned the tweet text by removing 'RT', URL, user name, punctuation and special characters. This pre-processing also cleaned diacritics (short vowels) and elongation. Then, we automatically filter out the content-based duplicates.

The pre-processing revealed that more than half of the tweets in the corpus were duplicates. The author and a native Arabic speaker volunteer inspected the duplicates manually and found that most of them were retweets. Other identical messages included shared news, emergency updates and instructions. We anticipated that this content was received and copied from different sources. We also found that tweets expressing emotional support included similar common prayers and condolence phrases. In addition, we observed that many nearly identical tweets were spam that contained similar text (tokens), with shared shortened links referring to the same URL or to URLs with similar content. Spammers habitually exploit trending hashtags to advertise and spread

malicious content. A Twitter post is either a new message or a retweet. Non-duplicates are corpus messages with unique text, whether new or retweeted. It is worth noting that we included retweets while collecting the data for two reasons. We noticed that the search API could miss many event-related tweets but captured them as retweets when we did not exclude them. In the future, we will analyse the factors impacting retweets during crises and explore which people pass the crisis-related tweets. Hence, we decided to collect both and filter out the content-based duplicates afterwards. Table 4.3 shows the percentage of content-based duplicates in Kawarith by crisis, along with the number of new messages and retweets.

4.4 Content-related Statistics

After eliminating the content-based duplicates, we calculated the number of words, sentences and unique hashtags for each event in the dataset. Kawarith includes 9,280,833 words and 1,046,579 sentences, as shown in Table 4.4. Considering the event location, the table also displays the main dialect for each crisis data. We followed the taxonomy created by Althobaiti (2020), who categorised the 36 Arabic dialects identified by Ethnologue (Eberhard et al., 2019) into seven main groups: Arabian Peninsula, Mesopotamian (Iraqi), Levantine (Shami), Maghrebi, Central Asian, Egyptian and Central & Northeast African. The Kawarith corpus covers five Arabic dialects.

As some tweets might be written in MSA or Arabic dialects that differ from the dialect spoken in the crisis area, we manually annotated tweets from events with different dialect groups to check whether they were written in the regional dialect, MSA or another dialect. As the Kawarith events were categorised into five main dialects, we randomly sampled 200 tweets from the first crisis belonging to each dialect category. We asked annotators to decide whether the tweet was written in MSA or dialect and choose which one. The selected events were the Jordan floods, Kuwait floods-18, Khartoum massacre, Cairo bombing and Karbala massacre, which were categorised into Levantine, Arabian Peninsula, Central & Northeast African, Egyptian and Mesopotamian dialects, respectively. We measure inter-rater agreement with Cohen's Kappa, resulting in $k = 0.9$. In cases

Table 4.3: Kawarith content redundancy statistics

Crisis name	# of new messages	# of retweets	# of messages with unique text	% of duplicates
Jordan floods	2379	6114	2383	71.94%
Kuwait floods-18	5504	28811	6139	82.11%
Qurayyat floods	903	8828	885	90.91%
Hafr Albatin floods-18	734	5335	786	87.05%
Leeth floods	1898	7698	1945	79.73%
Khartoum massacre	974	11331	1296	89.47%
Cairo bombing	747	1271	711	64.77%
Lebanon wildfires	1353	7232	3122	63.63%
Egypt floods	2207	8731	2384	78.20%
Hafr Albatin floods-19	1475	13071	2023	86.09%
Karbala massacre	2147	9814	1880	84.28%
Dubai floods	416	2064	383	84.56%
Coronavirus disease	189697	585472	250980	67.62%
Lebanon floods	2899	5516	3275	61.08%
Kuwait floods-19	5947	19544	6984	72.60%
Dragon storms	23125	68889	21815	76.29%
Aden floods	6640	30379	6274	83.05%
Oman floods	15843	64830	18224	77.41%
Ta'if floods	3910	21514	4612	81.86%
Beirut explosion	54956	252839	63408	79.40%
Syria wildfires	6459	16173	6160	72.78%
Sudan floods	45702	107424	23577	84.60%

Table 4.4: Kawarith content-related statistics

Crisis name	# of words	# of sentences	# of unique hashtags	Dialect based on geographical location
Jordan floods	48042	6214	447	Levantine
Kuwait floods-18	127261	19550	1077	Arabian Peninsula
Qurayyat floods	14504	2104	248	Arabian Peninsula
Hafr Albatin floods-18	14272	2097	256	Arabian Peninsula
Leeth floods	36375	4184	359	Arabian Peninsula
Khartoum massacre	34811	5297	241	Central and Northeast African
Cairo bombing	11141	1484	79	Egyptian
Lebanon wildfires	65275	9580	600	Levantine
Egypt floods	46022	6930	417	Egyptian
Hafr Albatin floods-19	31639	5373	454	Arabian Peninsula
Karbala massacre	41624	5477	399	Mesopotamian
Dubai floods	5423	793	126	Arabian Peninsula
Coronavirus disease	5683873	662311	33568	multi-dialect
Lebanon floods	57488	8540	397	Levantine
Kuwait floods-19	112123	20561	702	Arabian Peninsula
Dragon storms	431767	57212	3520	Egyptian
Aden floods	162964	18709	883	Arabian Peninsula
Oman floods	377670	51216	2514	Arabian Peninsula
Ta'if floods	75913	14479	1332	Arabian Peninsula
Beirut explosion	1369063	63035	8354	Levantine
Syria wildfires	109782	18240	1331	Levantine
Sudan floods	471843	69407	3012	Central and Northeast African
Total	9,280,833	1,046,579	60316	5 dialects

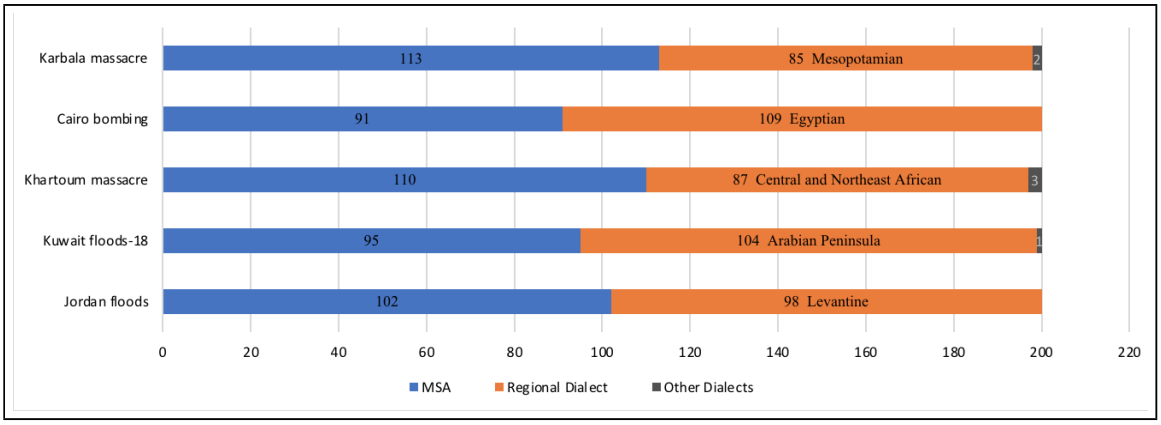

Figure 4.2: Distribution of dialects for tweets sampled from five crises

when the two annotators disagreed, the tweet was judged by the author.

Figure 4.2 shows the distribution of dialects for the tweets sampled from the five crises. We found that nearly half of the tweets for each event were written in MSA, while the rest were dialectal. Most dialectal tweets for each crisis were written in the regional dialect. In the Khartoum massacre data, we found two tweets written in dialects of the Arabian Peninsula and one in the Egyptian dialect. Two tweets in the sampled Karbala massacre data were written in the Levantine and Arabian Peninsula dialects, and we found one tweet written in the Egyptian dialect in the Kuwait-18 sample. Figure 4.3 depicts the distribution of the annotated thousand tweets by Arabic dialects. Table 4.5 show example tweets from the dataset for each dialect group. The dialectal words and phrases were marked in bold. The table also displays an example of a tweet written in MSA. It is worth noting that the tweet includes spelling mistakes (marked in bold), such as misspelling the different forms of alef, which we will consider in the pre-processing phase while we extract the topic words.

Figure 4.4 displays the distribution of the dialects in the Kawarith corpus. We generalised the observation about the percentages of MSA tweets (nearly 50%) and the regional dialects. We excluded the COVID-19 data as it is a global event, i.e., not limited to a specific Middle Eastern region. Many factors contributed to the varying dialects' percentages, including the crisis impact (i.e. severe crises will trigger more conversations), the dialect spread, the number of speakers of each

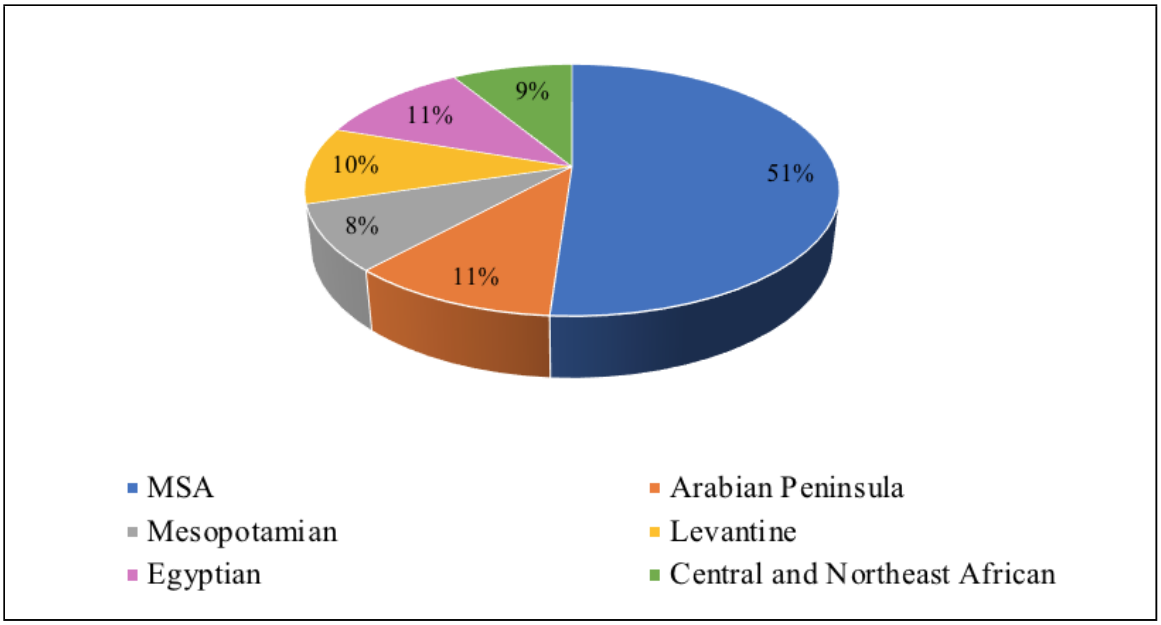

Figure 4.3: Distribution of 1000 tweets (sampled from five crises) by Arabic dialects

Table 4.5: Dialects with example tweets

Dialect	Tweet
Arabian Peninsula	**يالربع** والله تكسرت **جام** سيارتي الامامي من الحصى حتى الليت الامامي ماسلم تشطب **ابي** تعويض **شلون** الليت غالي #سيول_الكويت
Mesopotamian	**شكم** أم **هسه گلبها محروگ** على ابنها **الي** استشهد **شكم** طفل **حيعيش** بدون أب **شكم** واحد استشهد وهو **جان** عنده احلام وطموحات وحبيبة يريدها. الله يرحمهم ويصبر كل واحد فقد شخص عزيز عليه #كربلاء
Levantine	**مرق** على الحادثة الماضية اسبوعين و ولا مسؤول تحرك و عمل **اشي** محرز لحل المشكلة كل **الي** صار اعلان عطلة و مقابلات تافهة و **هي** عدد الوفيات و الاصابات ارتفع اليوم لنفس السبب خسرنا **كتير** لتى اللهم حوالينا ولا علينا
Egyptian	**اللي** قريب من القصر ويقدر يتبرع بالدم يتبرع **علشان** بنك الدم في الوقت **ده** من اليوم بيبقى **مافيهوش** دم **خالص** حرفيا #معهد_الاورام
Central and Northeast African	وصلني تلفون من السودان الإنترنت مقطوع تماما وفي شك **انو** الرسايل الدولية **برضو** مقطوعة بس جزئية الرسائل **ما مؤكده**
MSA	**اعلان** حالة الطوارئ في مدينة العقبة تحذيرا من سيول **قادمه** يرجى **اخذ** الحيطة من الجميع، وخصوصا من يسكنون في التسوية **اسفل** الطوابق **الارضية** ، **اكثروا** من الاستغفار **واكثروا** من التكبير فالله **اكبر** ،

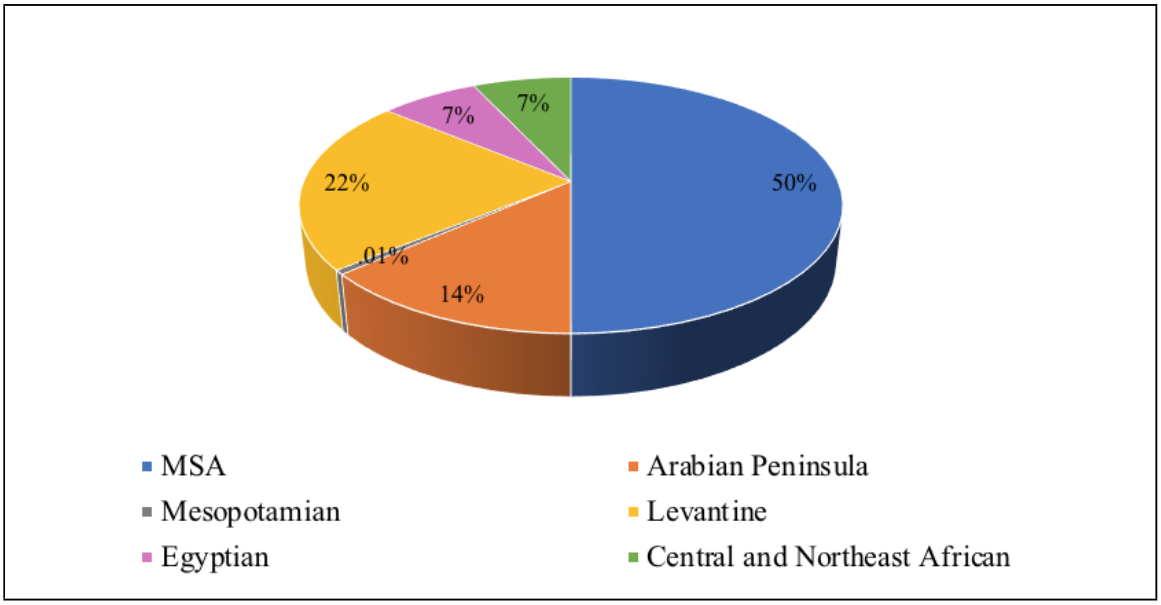

Figure 4.4: Distribution of corpus tweets by Arabic dialects

dialect and the number of active Twitter users per country[3]. In our study, we selected the crises based on their impacts, neither the location nor the dialect. For future work, we aim to include events occurring in North Africa to collect tweets written in the Maghrebi dialect.

4.5 Prevalent Topics

This section inspects keywords and main topics in the corpus to gain insights into users' communication during emergency events. It also identifies the main information types of conversations shared during crises in the Twitter Arabic content and ensures data usefulness for situational awareness. Identifying the primary information will help to propose accurate and clear annotation instructions. To achieve this, we employed word cloud and LDA topic modelling techniques. Prior to topic exploration, tweets were cleaned as explained below.

[3]https://www.arabsocialmediareport.com/Twitter/LineChart.html

4.5.1 Data Pre-processing

We conducted two main pre-processing steps. First, we removed noise by eliminating URLs, user names, punctuations, emojis, diacritics (short vowels), elongation and stop words, as we are interested in the words. We also omitted hashtags from the vocabulary, as these were used as query terms to collect the data and therefore occurred with greater frequency. The second step involved four types of letter normalisation: different forms of alef (آ، إ، أ) were normalised to bare alef ا, alef maqsora (ى) to ya (ي), wāw mahmoza (ؤ) to wāw (و), and ta marbouta (ة) to ha (ه). People often misspell words including these letters. Hence, we performed this step to mitigate the spelling mistakes. After normalisation, words that have been written in different forms, such as حافلة (bus) and حافله would be spelled the same way. Otherwise, they will be considered two different words.

Stop words were removed from the vocabulary because of their high frequency of occurrence without adding meaningful content to the domain in question. For this purpose, we employed Arabic stop words from the NLTK toolkit (Bird, 2006) and Alrefaie's repository[4], which contain 243 and 750 stop words from MSA and classical Arabic, respectively. We found that many of the dialectal stop words in our corpus are not used in MSA, as Arabic-speaking people also tweet in their dialects. To the best of our knowledge, there is no available domain-independent multi-dialect Arabic stop word list. Hence, we created such a list by collecting samples of tweets from the countries in our crisis list (see Table 4.1) and manually identifying the dialectal stop words from the most frequent words in each sample.

Using the Mo3jam dictionary[5], we added synonyms in other dialects, taking account of spelling variations. For example, the word لِسَعْ (lsʕ[6]) "not yet" which blends لساعته "to this moment" and حتى هذه الساعة "until this moment" (Aldrsoni, 2012), also takes the form لسّاته (lsAth). Arabic speakers tend to adopt a phonological system of spelling when writing non-MSA words, and the former could also be written as لِسّه (lsh), لسّا (lsA) or لسى (lsý). We also included common

[4] https://github.com/mohataher/arabic-stop-words
[5] https://en.mo3jam.com/
[6] We used Habash-Soudi-Buckwalter transliteration scheme (Habash et al., 2007)

misspellings of frequently occurring words such as أصلن (ÂSln) "ever" for the word أصلاً (ÂSlA). It is important to note that as we discarded diacritics, homographic stop words that share spellings with commonly used non-stop words were disregarded. For example, to avoid filtering out the word دُوَل, which translates as 'countries', we ignored the word دُوْل (dwl), which means 'these/those' in Egyptian and Higazi dialects. The final stop word list included 405 multi-dialect Arabic stop words[7]. Adding these words to the NLTK and Alrefaie's lists, 1177 (1098 after letter normalisation) words were identified as stop words to be filtered out before extracting keywords and underlying topics.

4.5.2 Most Frequent Words

We used word clouds to visualise the text at event level to gain some sense of the most frequent unigrams and bigrams and find the shared themes across crises. First, we exhibit the most frequent words of each event before eliminating the content-based duplicates to investigate the most popular topics. Figure 4.5 depicts word clouds of the top 200 words associated with the crisis events. In general, the diagrams show that the most frequently occurring terms are location names. For many events, terms related to emotional support and prayers show a high rate of occurrence. A closer look reveals words about crisis impact, whether related to the environment, such as البنية التحتية "infrastructure", or services such as مياه "water", اكسجين "oxygen" and كهرباء "electricity". The diagrams display many individual names; most of them are victims, officials and politicians. Many political terms appear in events that occurred in countries witnessing civil unrest, such as Lebanon and Yemen. Observation suggests that one crisis can be discussed using data from another; for example, #لبنان_ينتفض (which relates to the Lebanese protests) is the second most frequent hashtag in the Lebanese floods data. For that reason, it is helpful to identify messages in terms of crisis type following data collection.

Other visible words are relevant to hazard response such as الدفاع المدني "civil defence" and أرقام الطوارئ "emergency contact numbers". The diagrams for flood events reveal many weather-

[7] https://github.com/alaa-a-a/multi-dialect-arabic-stop-words

Figure 4.5: Word clouds showing the top 200 words from Kawarith

related terms. Human-induced crises such as the Cairo bombing, Karbala massacre and Beirut explosion share many common event-independent words such as دماء "blood", مستشفيات "hospitals", مصابين "injured people" and تبرع "donation", along with crisis-specific terms. Regarding COVID-19, prevalent terms include وزارة الصحة "Ministry of Health", الإجراءات الاحترازية "prevention measures", إصابة جديدة "new case" and الصحة العالمية "World Health Organisation".

After duplicate removal, the top terms have also been visualised to check if there are notable changes in predominant words. In general, eliminating duplicates unfolded more terms such as named entities. Prayers and location names still populate the diagrams. We noticed a shift in word frequencies for some events. For instance, the phrase وزارة الأشغال "Ministry of Public Works" shrinks in size after duplicate removal in the Lebanon floods and Kuwait floods-19, indicating that messages discussing this topic have received a high number of retweets. Unlike crises involving social movements or shootings, the Khartoum massacre generated words related to internet blockage. Following duplicate filtration, visualisation revealed further hazard-related words. A similar pattern has been found in the Leeth floods data, confirming that a single topic may dominate the event dataset because of the duplicated content.

Retweeted messages or content duplicates are not necessarily relevant to the crisis, as spam messages associated with crisis-related hashtags sometimes attract a large number of shares, and word clouds may include irrelevant terms (e.g. advertisements as in Ta'if floods). In the case of the Dragon storms, phrases about invoking blessings upon the prophet Mohammed (peace be upon him) populate the diagram because the event occurred on Friday. This confirms the importance of removing duplicates and irrelevant posts from crisis data in pursuit of meaningful insights. Figure 4.6 shows the frequent words associated with the Leeth floods, Khartoum massacre, Lebanon floods and Dragon storms before and after duplicate removal, as they show the importance of duplicate filtration.

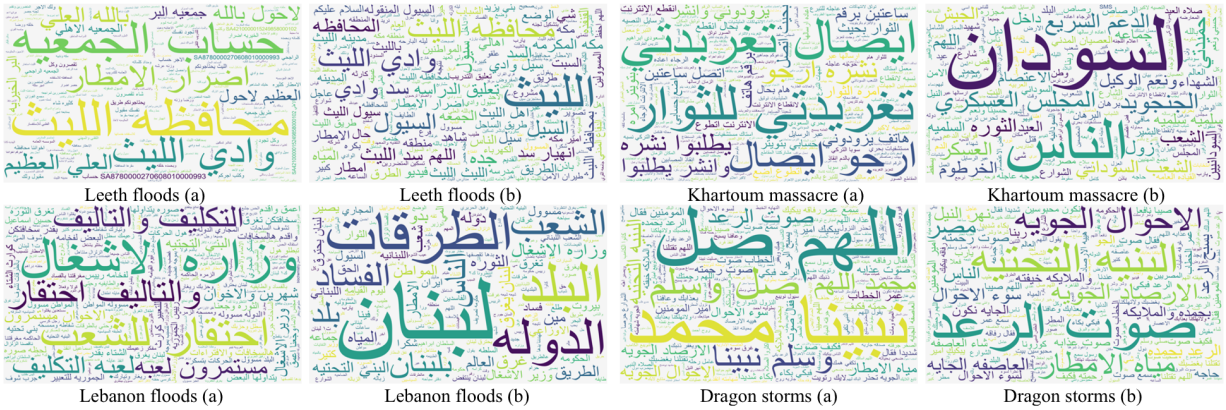

Figure 4.6: Word clouds for four crises: (a) for all data and (b) after duplicate removal

4.5.3 LDA Topics and Content Categorisation

We also investigated the main topics discussed during crises using the LDA modelling technique. We ran the LDA model for ten crises, as the process of topic extraction and interpretation is time-consuming. The selected crises are Jordan floods, Kuwait floods-18, Karbala massacre, COVID-19, Dragon storms, Aden floods, Oman floods, Beirut explosion, Syria wildfires and Sudan floods. The selected list of event data covers different Arabic dialects and hazard types. In this research, we used the scikit-learn library (Pedregosa et al., 2011). Text in the corpus was stemmed using the Farasa[8] stemmer (Mubarak, 2018) before performing the described pre-processing steps. Then, we tokenised the text into words. We filtered out terms that appear in more than 85% of documents or in less than ten documents. For reproducibility, the random state was set to 101.

Choosing the optimal number of topics (k) is crucial in LDA models as it determines the quality of the generated themes. Selecting a small k would result in broad topics, whereas choosing large values would lead to fragmented or uninterpretable topics (Syed and Spruit, 2017). Different metrics have been proposed to evaluate the quality of LDA models, including the coherence measures. A topic is described as coherent if its top words are semantically related. In this work, we used the C_v coherence measure (Röder et al., 2015; Syed and Spruit, 2017) as an indicator to identify the optimal number of topics per event. Röder et al. (2015) demonstrated that the C_v

[8] https://pypi.org/project/farasapy/

measure highly correlates with the human interpretability of topic words. The C_v measures how the top scoring terms in each topic (we set to 20) are semantically similar to each other by using the Normalised Pointwise Mutual Information ($NPMI$) and the cosine similarity (Röder et al., 2015). The value of C_v ranges from 0 (minimum) to 1 (maximum). We evaluated four LDA models for each crisis with the proposed topic numbers: [10, 15, 20, 25], which were empirically chosen. Then, we selected the model that retained the highest coherence value. We obtained C_v scores ranging from 0.6 to 0.79, indicating coherent topics. The pre-processing and using the dialectal stop words improved the total C_v scores (for the ten events) by 3.7% on average.

To assign interpretations to topics, we manually analysed examples of tweets belonging to each topic to explore their meanings in context. Table 4.6 shows the interpretation of each topic extracted from the Jordanian flood data. Analysis of extracted topics and example tweets from different emergency events yielded the following findings regarding the prominent topics.

Most of the broad information types used for situational awareness and coordinating response efforts as reported in previous work (Olteanu et al., 2015; Imran et al., 2016; Sit et al., 2019) appear in our corpus as topics, varying in frequency across events. One type of these topics that we found in each crisis data is related to the affected population and rescue effort, as in topics 5, 8, 11 and 12 in Table 4.6. Most of the posts belonging to this theme are reports about affected people (fatalities, missing, trapped, injured, evacuated, infected and survivors) that were generated by organisations or individuals. Other messages of this category are sent by individuals who request help or report damage to personal property.

Another type of information that supports situational awareness is infrastructure damage, as in topic number 15 in Table 4.6. This topic is found in most of the natural disaster data in the Kawarith corpus. We found topics about interrupted services such as power outages in the Aden flood data. Other extracted topics can be categorised as warnings and precautions. Tweets exhibiting this topic were usually posted by organisations. Examples of these topics include weather warnings (as in topic 10), an announcement of school closure, disease prevention measures and best practices.

Table 4.6: Topics extracted from the Jordan floods data using the LDA model

#	Top 10 words per topic	# of posts	Interpretation
1	مسؤول ازمه ادار سد وطني مدرسه مدير مواطن دفع خاص "responsible crisis manage dam national school principal citizen pay private"	94	The responsibility for the deaths of 15 pupils from a private school, while they were on a school outing during the event
2	وطن منطقه ضحيه ديوان ملكي مملكه اعلن حذر طقس روح "homeland region victim court royal kingdom announced caution weather soul"	81	Warnings about weather conditions and news about mourning of victims.
3	كويت غرق مطر ماء بتراء انباط امطار شغل بناء رياض "Kuwait drown rain water Petra Nabataean rains work building Riad(person name)"	151	The heavy rains that hit the city of Petra and some mentions of the Kuwait floods
4	اللهم سقي رحم عذاب اجعل غرق خير شر هدم كويت "oh_god watering mercy torment make drowning good evil destroy Kuwait"	250	Supplications said when feeling worried during heavy rains
5	عمل اردن وادي موقع اخلاء حكومه امنيه جنوب دفاع مدني "work Jordan valley site evacuation government security south civil defence"	125	The evacuation of civilians near the Jordan valley by civil defence authorities
6	مصاب حفظ وطن اردن راجع اهل شهيد رحم اراد الله "injured save homeland Jordan return family martyr have mercy wanted Allah"	279	Prayers for victims, emotional support for their families and supplications
7	كويت لطف رحمه شعب يارب الله احفظ اردن اهل اللهم "Kuwait kindness mercy people Lord Allah save Jordan people oh_God"	342	Prayers to save the people of Jordan and Kuwait during floods

Continued on next page

Table 4.6 – continued from previous page

#	Top 10 words per topic	# of posts	Interpretation
8	اردن وزير رزاز سائح اردنيه حكومه رساله عاجل جديد اسرائيلي "Jordan minister Razzaz(politician) tourist Jordanian government message breaking new Israeli"	136	Evacuation of tourists from the city of Patera and news related to the role of the prime minister during the event
9	دوله تحتي بنون مات فساد مطر بلد ناس اردن اراد "country infra children died corruption rain land people Jordan wanted"	145	The corruption that led to this catastrophe
10	حل سوسنه حال جويه شكر جو اخبار معني بتراء اراد "solution Assawsana(newspaper) weather condition thank news meaning Petra wanted"	211	Weather updates from the news
11	اراد حق طفل باد ساره فيديو عمل رجل انقذ اصاب "wanted justice child want Sarah(victim) video work man rescue hit"	91	The rescue efforts after the flash flooding. Posts on this topic include the hashtag: #نريد_حق_سارة_والأطفال "#justiceForSarahAndChildren"
12	حال اراد جويه ضحيه ارتفاع بيت الله عاجل وفاه مملكه "status wanted weather victim elevation house God breaking death kingdom"	100	Reports of deaths, most of which were sent by news accounts
13	سد واله عاجل عمليه دفاع مدني مادب بحث مفقود اردن "dam Wala(dam name) breaking operation defence civil Madaba(city) search missing Jordan"	189	Civil defence members search for missing people in the Wala district in Madaba and information about the state of the Wala dam

Continued on next page

Table 4.6 – continued from previous page

#	Top 10 words per topic	# of posts	Interpretation
14	بنيه تحتيه مناخي تغير كارثه وجه حق مر طلع استودع "infrastructure climate change disaster side right pass rise bid farewell"	121	Discussions about climate change and criticism of infrastructure failure
15	بحر ميت طريق فاجع جسر اردن اتجاه روضه امن طالب "dead sea road catastrophic bridge Jordan direction kindergarten security student"	68	Information about the infrastructure damage (collapsed bridges and road closures) caused by the dead sea flood

Words related to emotional support and prayers emerge as topics in each crisis data. Examples include topics 4, 6 and 7 (Table 4.6). We found topics showing thanks and gratitude for volunteers and response teams, as in the Kuwait floods-18 and Oman floods. Some of the derived topics represent public opinion and general discussion, as shown in topics 1 and 14. Most of the views are criticism and sarcasm, and the conversations sometimes drift from the crisis topic to politics, as we found in the Beirut explosion data. Other extracted topics describe relevant events and consequences such as authority resignations and travel restrictions during the COVID-19 pandemic. In-domain fine-grained topics were identified from COVID-19 data including disease spread, COVID-19 symptoms, vaccine, volunteering, prevention measures and other relevant discussions regarding the virus impact.

A crisis-unrelated topic emerged from the COVID-19 data. The top words of this theme are related to advertisement and online shopping, including offer, order, application, shop, coupon, discount and price. Similar themes have been found in the Dragon storms and Syria wildfires data, which include crisis irrelevant terms. Documents belonging to these topics are mostly spam, and they form about 7% and 4% of each corpus, respectively. Terms related to the COVID-19 spread appear as a topic in the Oman flood data. Messages on this topic provide updates about the Coronavirus situation in Oman and include flood-related hashtags. These messages were published during the beginning of the flood crisis. Generally, we found that COVID-19 has been mentioned

in many of the subsequent events data. The Sudan flood has also been mentioned using the Syria wildfires hashtags, particularly in emotional support tweets and prayers. Similarly, some users referenced the Kuwait floods in their prayers during the Jordan floods (see topic 7 in Table 4.6).

We investigated the main concerns during the crises of different types (Jordan floods, Karbala massacre, COVID-19, Dragon storms, Beirut explosion and Syria wildfires) by looking at the topic distribution. Considering the dominant topic per document, we counted the number of tweets belonging to each topic. Empathy and prayers represent the most popular theme in the Jordan floods, Syria wildfires and Beirut explosion, followed by topics related to the affected population and rescue effort. Most tweets in the Dragon storms data are cautions and warnings (16.4%) and reports about infrastructure damage (20%). Despite being small in number in the Jordanian data, posts about infrastructure damage generally received a high number of retweets. A tweet on this topic has been shared 33 times on average during the Jordan floods. Most messages of the Karbala massacre express sect- and politics-related opinions, and they retain a high number of retweets (average of 44 shares per message). For the natural disaster data, public opinion tweets obtain a low number of average retweets per post. Disease spread (situation reports) is the predominant and most shared topic in the COVID-19 corpus, followed by prevention measures and cautions.

In this study, we propose a coarse-grained annotation scheme to categorise messages as different information types based on manual interpretation of prevalent topics and in light of earlier taxonomies. A multi-label scheme has been employed as a tweet can communicate different information types. For example, the message below includes warning updates about affected individuals (the first sentence) and weather conditions (the second sentence). Looking at the topic distribution for this tweet, the dominant topic is the second one shown in Table 4.6, which mainly describes weather warnings. The topic also includes some victim-related words, as some of the documents belonging to this topic mention updates about both weather and victims. The tweet also belongs to topic 12, which shows death reports.

الأردن #عاجل وفاة طفلة بسبب السيول الجارفة اليوم الجمعة وفقدان عدد من الأشخاص في عدة مناطق. وتشهد الأردن ومناطق واسعة من #السعودية موجة من الأمطار الغزيرة وتشكلا

للسيول. ويتوقع اشتداد الحالة الجوية هذا المساء وغدا السبت. #سيول ـالاردن #وسم #غدق #طقس

"#Jordan #Breaking A child has died in flash flooding this Friday and several people are missing in many areas. Jordan and wide regions of Saudi are witnessing heavy rainfall leading to floods. Severe weather conditions are expected this evening and tomorrow. #Jordan_flood #wasm #Ghadag #weather"

Table 4.7: Labels with example tweets

Label	Tweet
Affected individuals & help	عشرة عناصر من اطفاء بيروت مفقودين. "Ten members of Beirut firefighters are missing." قسم الطوارئ في مستشفى اوتيل ديو يستغيث ويطلب للتبرع بالدم. "The Emergency Department at the Hôtel-Dieu hospital calls for help and appeals for blood donations."
Infrastructure & utilities damage	قطع المياه عن محافظة القاهرة بالكامل لسوء الأحوال الجوية ـ بوابة الشروق. "Water is cut off in Cairo governorate due to bad weather - Al Shorouk Gate."
Caution, preparations & other crisis updates	الأرصاد تحذر أمطار وسيول وانخفاض في درجات الحرارة ورمال وأتربة. وتعطيل للدراسة الخميس بسبب الأحوال الجوية. "The Meteorological Department warns of rains, floods, temperature drop, sand and dust. And schools are closing on Thursday due to weather conditions."
Emotional support, prayers & supplications	اللهم احفظ مصر وأهلها. "May Allah save Egypt and its people."
Opinions & criticism	الاستقالات لا تكفي، وانما المطلوب محاسبة ومحاكمة كل مسئول مهمل. "Resignations are not enough, what is required is accountability and trial for every negligent official."
Irrelevant	في أعماقنا رعد وبرق وعواصف وأمطار لا تشير إليها الأرصاد الجوية. "Deep inside us are thunder, lightning, storms and rain that have never been detected by weather forecast."

Unlike previous work taxonomies (Olteanu et al., 2015; Imran et al., 2016), we did not consider the 'other useful information' class since it is subjective to decide the information's usefulness. Instead, we introduced 'other crisis updates' as a "catchall" category for other information that varies across crises, such as flood level and emergency location. We observed that such updates were usually mentioned as caution. Hence, we merged these two categories. As a few messages related to donation and volunteering in most events, we combined this category with affected individuals, as donations meant to help the affected population. We also tagged opinions, supplications and prayers. These may not be useful to humanitarian responders or contribute to situational awareness but can be used for other purposes, including opinion mining and measurement of event impact. Table 4.7 shows some example tweets from each category. In the case of COVID-19, a tweet was classified as either relevant or irrelevant to the event[9]. The following section describes the manual labelling process.

4.6 Manual Annotation and Inter-rater Reliability

Tweets from seven crises were manually labelled to automatically identify information categories by ML algorithms. The seven crises selected for annotation are the Jordan floods, Kuwait floods-18, Hafr Albatin floods-19, the Cairo bombing, the Dragon storms, the Beirut explosion and COVID-19. We focused on flood and explosion events as frequent occurrences in the Middle East and considered COVID-19 as it is an impactful global pandemic.

The data were annotated by volunteers. All annotators were native Arabic speakers. We started with 25 coders. First, coders were trained using a short quiz with examples from each category and explanations of the correct answers. The quiz includes ten questions. To further ensure reliability, the 25 annotators were tested on 30 examples from one event, and only those scoring 70% were allowed to proceed. The final judgments were provided by 21 trusted coders. Each tweet was judged by two annotators, who were provided with annotation instructions and

[9]To comply with Twitter's policies, we explicitly avoided coding data about users' health.

a piece of news or Wikipedia article summarising the crisis. (The annotation instructions and examples from the training quiz are translated to English and presented in Appendix One).

For annotation, we considered only tweets with unique texts. We excluded duplicate messages as identified in Section 4.3 to avoid labelling messages with the same content. We did not consider propagating labels to duplicate tweets after labelling the unique messages to avoid experimental bias in classification. Including duplicates in the dataset results in an overestimated performance if there is an overlap between test and training data (Alam et al., 2021b). We also removed tweets containing less than four tokens as these are too short to convey any meaningful message. When calculating a tweet's length, we split the hashtags. As noted earlier, we did not consider user mentions and URLs as proper tokens. Each hyperlink was replaced with the Arabic word رابط (link) to inform coders of a link referring to a website, image or video. Annotators were not required to visit the hyperlinks, as tweets were judged only on their text content. We sampled a different number of tweets for annotation from each event, ensuring that samples were taken from different timeframes. About 70 – 85% of flood events data were considered, along with all unique examples from the Cairo bombing (which contains only 711 distinct tweets). In the case of Dragon storms and the Beirut explosion, about 1050 posts were sampled from each crisis. Regarding COVID-19, we considered 2005 tweets.

We used Krippendorff's alpha coefficient (Artstein and Poesio, 2008) to calculate the inter-rater agreement as it supports multi-label annotation. The average Krippendorff's alpha for the seven events was about 0.7, indicating substantial agreement and clear instructions. Most disagreements occurred because of the multi-label scheme, as annotators agreed on a subset of labels per example. If two annotators disagreed, the message was judged by a third person; if the third coder did not agree with coder 1 or 2, majority voting was applied to select the label decided by at least two coders. A tweet was discarded if all three annotators disagreed entirely with each other.

In total, we obtained 12,446 labelled examples. Table 4.8 shows the distribution of tweets by relevancy and the total number of labelled tweets, while Table 4.9 shows the distribution of information categories. The dataset is imbalanced, and the distribution of information types varies

Table 4.8: Distribution of labels (relevant vs irrelevant)

Crisis	# of relevant tweets	# of irrelevant tweets	**Total**
Jordan floods	1882	118	**2000**
Kuwait floods-18	3701	399	**4100**
Hafr Albatin floods-19	978	637	**1615**
Cairo bombing	700	6	**706**
Dragon storms	701	309	**1010**
Beirut	833	177	**1010**
COVID-19	1782	223	**2005**

Table 4.9: Distribution of information types: Flood crises, Cairo bombing, Dragon storms and Beirut explosion

Label	Jordan floods	Kuwait floods-18	Hafr Albatin floods-19	Cairo bombing	Dragon storms	Beirut explosion
Affected individuals & help	331	414	83	138	70	186
Infrastructure & utilities damage	39	271	100	17	105	64
Caution, preparations & other crisis updates	268	980	475	214	252	170
Emotional support, prayers & supplications	709	816	202	222	120	277
Opinions & criticism	604	1355	189	181	221	198
Irrelevant	118	399	637	6	309	177
Total	**2000**	**4100**	**1615**	**706**	**1010**	**1010**

across events. On average, only 4.4% of dataset instances have more than one label. Most relevant messages conveyed emotional support, opinions, cautions and crisis updates. Among COVID-19 tweets, we observed that the largest category of relevant messages relates to disease spread. The non-negligible percentage of irrelevant tweets (15% of the dataset) highlights the need for a classification step following data collection to filter out irrelevant posts. Not all irrelevant messages are spam that uses crisis-related hashtags. Some off-topic tweets were crawled due to the keyword ambiguity, as in the example shown in Table 4.7.

The annotated dataset can be leveraged for several tasks, including crisis detection and crisis type classification. Assigning messages to categories to identify informative posts can enhance situational awareness and assist emergency responders in organising effective relief efforts. The labelled dataset can also be utilised to gauge public opinion and sentiment during crises.

4.7 Conclusion

This chapter presented the accessible Twitter crisis datasets and introduced Kawarith, a large-scale Arabic Twitter corpus for 22 crises. The corpus was built by tracking relevant keywords using an iterative collection process as described in Section 4.1. Then, we reported a preliminary analysis of the corpus content, including tweet-related and user-related statistics. We investigated the main information categories of conversations posted during a range of crisis events by uncovering the hidden topics using the LDA modelling technique. Then, we analysed samples of tweets belonging to the extracted topics and identified the main information types shared on Twitter during the crisis events. We proposed a common multi-label scheme based on the specified information types. We showed how the data were pre-processed before investigating the prevalent themes and including the stop-removal step, which involves compiling 405 domain-independent multi-dialect Arabic stop words. Finally, we created and published a gold-standard multi-label dataset comprising ∼12k unique tweets from seven crises. The dataset will be used in this thesis to automatically identify the crisis-related messages from Arabic Twitter data using supervised learning techniques.

Chapter Five

Evaluation of Crisis Tweet Classification Models

This chapter presents the experiments performed to evaluate classical supervised learning approaches and DNN models on SM crisis classification tasks using our Arabic Twitter dataset: the Kawarith corpus. It introduces the models' configurations, feature representation and training settings. Finally, the chapter illustrates the results and analyses the errors of the best-performing model.

5.1 Models' Settings

5.1.1 Classical ML Classifiers

Classic ML classifiers were implemented using the scikit-learn library. For the multi-label classification, we used the One-vs-Rest strategy. It decomposes a multi-label classification problem into multiple sets of binary subproblems, one for each label. Predictions are made by the most confident model. In the following, we highlight the models' parameters.

- **NB:** We experimented with the multinomial NB classifier, where the input data are modelled as occurrence counts such as BoW. Previous studies showed that the multinomial NB outperforms the Bernoulli NB for text classification (McCallum, Nigam, et al., 1998; G. Singh

et al., 2019).

- **SVM:** We used a linear SVM (Fan et al., 2008). A linear kernel is faster than other kernels and achieves good performance when using large numbers of features (Hsu et al., 2003).

- **Decision trees (DT):** We used the default values for the parameters as specified by the scikit-learn library. For example, the maximum depth of the tree was set to none, and the minimum number of samples the algorithm requires to split an internal node was kept as the default value (i.e. 2).

5.1.2 DNN Architectures

Deep learning models were built using the Keras library[1]. The input sequences, the embedding and output layers were similar for the DNN models described below. The embedding layer was used as the first hidden layer to map words (input sequences) to dense vectors. The output layer mapped its input vectors—obtained from the last hidden layer in each model—to a probability for each class using the sigmoid activation function. Figre 5.1 shows the general architecture of DNN models. The configurations of the DNNs that we experimented with are described below.

- **CNN:** We experimented with two CNN architectures. The first one was similar to that proposed by Y. Kim (2014). We used two 1D convolutions applied in parallel to the input layer vectors, extracting local patches from sequences using convolution windows of sizes 3 and 5 with 100 feature maps each, which were empirically chosen. A sliding max-pooling operation of size two was applied over each feature map to obtain the maximum value, representing the most important feature. The output vectors of the two convolutions were concatenated, and a 0.5 dropout rate was applied for regularisation. The output was fed into a 100-dimension fully connected layer with Rectified Linear Unit (*ReLU*) activation. We call this model 'CNN-K'. The second variant was similar to Kim's CNN but applied two 1D convolutions sequentially using windows of sizes 5 and 3 with 100 and 50 feature maps, respectively.

[1] https://keras.io

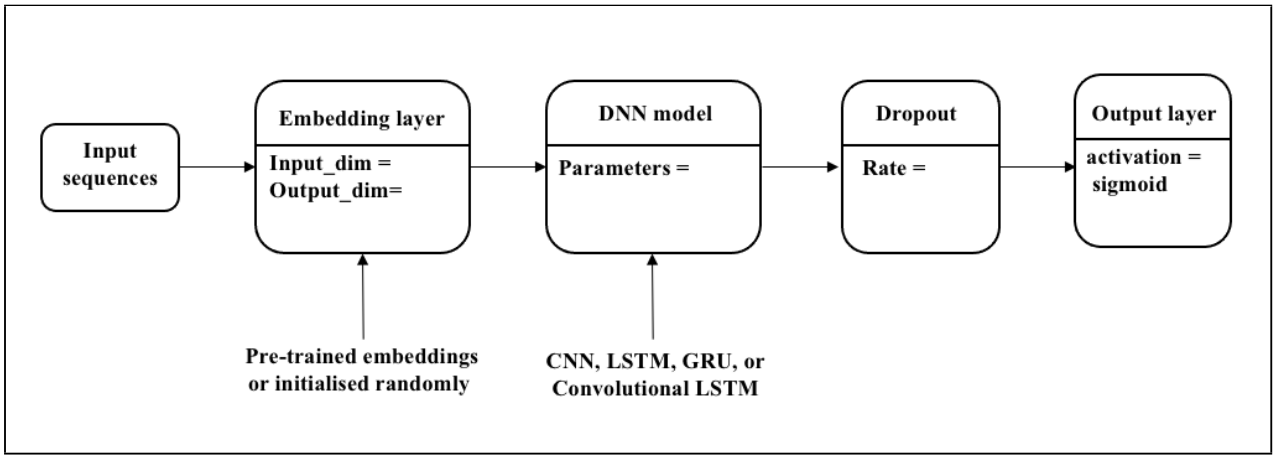

Figure 5.1: General architecture of the DNNs

- **LSTM:** This model involved one LSTM layer with 100 hidden output dimensionalities. Experimenting with different output dimensionalities (e.g. 50, 100 and 150) produced similar results (in most cases). Also, using sizeable hidden output dimensionalities for RNNs is not recommended when classifying short text (Chollet, 2017). As proposed by Gal (2016), we applied a dropout rate for input units of the LSTM layer and a dropout rate of the recurrent units for regularisation. In the experiments, both were set to 0.2 and were empirically chosen.

- **GRU:** This model included a GRU layer with hidden units and dropout rates similar to those applied to the LSTM layer to make a fair comparison. GRUs train faster than LSTMs. We experimented with this network structure to compare its performance against the LSTM and compare gated RNNs against the CNN models on the two classification tasks.

- **Convolutional LSTM** ($CLSTM$): This model is similar to the CNN-K described above, except that the fully connected dense layer is replaced by an LSTM layer identical to the one presented above. In this architecture, the CNN was used to extract features that were fed into an LSTM layer, which processed down-sampled high-level input sequences.

5.2 Text Pre-processing and Feature Representation

We conducted the following pre-processing steps:

1. Noise removal: we removed emojis, punctuation, special characters, non-Arabic characters and diacritics. Stop words were eliminated as they are uninformative in such topic-based classification tasks, and do not affect the overall meaning of a document.

2. Text normalisation: we performed a set of normalisation techniques commonly applied to Arabic text to transform words into a more uniform sequence (Albalawi et al., 2021). We normalised elongation. Words including more than two consecutive repeated letters were normalised by keeping only two occurrences of the same letter to avoid eliminating legitimate consecutive identical letters. For instance, the word لا (No), written informally as لاااا (Noooo) for emphasis, was normalised to لاا (Noo). We also performed three types of letter normalisation: the variant forms of alef to bare alef, alef maqsora to ya, wāw mahmoza to wāw, and ta marbouta to ha. This was done because people often misspell various forms of alef/wāw and do not distinguish between ta marbouta and ha when these letters occur at the ends of words. As in previous studies (D. T. Nguyen et al., 2016; Verma et al., 2011), we replaced each URL and user handle with the words رابط (hyperlink) and إشعار (mention), respectively, instead of removing them (i.e. replacing them with white space), assuming that URLs and usernames are helpful features in distinguishing crisis messages.

3. Tokenisation: we conducted whitespace-based tokenisation. We did not perform stemming to the tokens as Alabbas et al. (2017) suggested that most ML classifiers perform better on Colloquial Arabic without applying stemming.

In both types of models, we limited the vocabulary to the most common 5000 words in the training corpus. Concerning classical ML models, unigrams, bigrams and trigrams of words were extracted. In the case of NB, text was represented as BoW (without a weighting mechanism) as we experimented with a multinomial NB classifier which is suitable for classification with discrete

Table 5.1: The macro F1 scores of the ML models with un-processed and pre-processed input text

Model	Un-processed text	Pre-processed text
NB	75.03%	77.15%
SVM	74.66%	77.62%
DT	68.08%	70.65%

features. For SVM and DT models, the features were transformed into TF-IDF vectors, in which each tweet represented a document. To explore whether using character-based features produce better performance, we also experimented with character n-grams, with $2 \leq n \leq 5$, which was empirically chosen.

To investigate the impact of the performed pre-processing steps, we evaluated the performance of the three classical ML models (i.e. NB, SVM and decision trees) before and after pre-processing the text. We randomly shuffled and split the labelled Kawarith data into train and test sets (75% and 25%, respectively) and classified them into relevant or irrelevant using the unigrams, bigrams and trigrams of words as features. Table 5.1 shows that the performance of all models has been improved with pre-processing the text by values ranging from 2.12% to 2.9%.

For DNNs, texts were segmented into words. The maximum length of input sequences per tweet was set to 60 tokens, as the longest tweet in Kawarith has 60 words. Messages comprising fewer than the set maximum length were zero-padded. Each word was transformed into a vector. Concerning the BERT experiments, the maximum length was set to 100, which is the longest sequence after the segmentation. As the BERT is a contextualised representation, we kept the stop words.

Word vectors were first initialised from the Twitter CBOW AraVec (Soliman et al., 2017). The model was trained using a CBOW technique on Arabic Twitter text of 66.9M Arabic tweets and 1090M tokens. Using the Twitter API, their training Arabic Twitter dataset was collected between 2008 and 2016 from different random geographical locations. The authors set the window size to three words as tweets are short (i.e. the maximum length of a tweet is 140 characters). The

vector dimension of the model is 300. We also explored the performance of DNNs using character embeddings. We initialised the vectors from a FastText model trained on 10M Arabic tweets (A. I. Alharbi and M. Lee, 2020). The tweets were collected from different Arabic countries. The model's vector dimension was set to 200. In their model, the authors have ignored the words with a total frequency lower than three.

For the contextualised representation, we experimented with the AraBERT Base model (Antoun et al., 2020), which has 12 encoder layers, 768 hidden dimensions, 12 attention heads and 512 maximum sequence length. AraBERT was trained on Arabic news articles collected from various media in different Arabic countries. The training dataset included 70 million sentences. We experimented with the version that uses Farasa (Abdelali et al., 2016) to segment the text before training the tokeniser as we found it leads to better performance on our task. As explained in the methodology, we used the BERT embeddings in two ways: fine-tuning the embeddings during training with a linear classification layer and combining BERT with the best performing DNN. For the former, the linear layer was preceded by a dropout layer of a probability (0.2) to prevent the model from over-fitting.

5.3 Imbalance Handling

We used the Kawarith corpus. As the labelled dataset had imbalanced classes, we handled that before training. For the binary classification (relevancy detection) task, we duplicated the samples belonging to the 'irrelevant' class. We performed up-sampling rather than randomly down-sampling the majority class because the latter can result in losing many instances and thereby degrade the performance, particularly for DNNs. Besides that, the sample of the chosen majority class could be biased. Table 5.2 displays the size of the training data for the relevancy detection task before and after up-sampling the minority class.

Regarding the information classification, we also duplicated the samples in the minority class: 'infrastructure and utilities damage', as they were much fewer than examples of other classes. The

Table 5.2: Number of training examples for the relevancy detection task after the up-sampling

Target data	Training data	# of training examples before the up-sampling =[on-topic+off-topic] posts	# of training examples after the up-sampling =[on-topic+up-sample(off-topic)]
Jordan floods **(JF)**	KF+HF+CB+ CD+DS+BE	10446 [8695+1751]	12197 [8695+3502]
Kuwait floods-18 **(KF)**	JF+HF+CB+ CD+DS+BE	8346 [6876+1470]	9816 [6876+2940]
Hafr Albatin floods-19 **(HF)**	JF+KF+CB+ CD+DS+BE	10831 [9599+1232]	12063 [9599+2464]
Cairo bombing **(CB)**	JF+KF+HF+ CD+DS+BE	11740 [9877+1863]	13603 [9877+3726]
COVID-19 **(CD)**	JF+KF+HF+ CB+DS+BE	10441 [8795+1646]	12087 [8795+3292]
Dragon storms **(DS)**	JF+KF+HF+ CB+CD+BE	11436 [9876+1560]	12996 [9876+3120]
Beirut explosion **(BE)**	JF+KF+HF+ CB+CD+DS	11436 [9744+1692]	13128 [9744+3384]

Table 5.3: Number of training examples for the information classification task after the up-sampling

Target data	Training data	# of training examples after the up-sampling
Jordan floods **(JF)**	KF+HF+CB+DS+BE	6913
Kuwait floods-18 **(KF)**	JF+HF+CB+DS+BE	5094
Hafr Albatin floods-19 **(HF)**	JF+KF+CB+DS+BE	7817
Cairo bombing **(CB)**	JF+KF+HF+DS+BE	8095
Dragon storms **(DS)**	JF+KF+HF+CB+BE	8094
Beirut explosion **(BE)**	JF+KF+HF+CB+DS	7962

information categorisation task was assessed separately, i.e. we suppose that we managed to filter out all irrelevant posts and need to categorise the crisis-related tweets into pre-defined information types. Table 5.3 shows the size of the training set, after the data balance, for the information classification task. Figure 5.2 displays the distribution of labels in the training data for each target event.

5.4 Training Settings and Evaluation Metrics

The DNN models were trained for five epochs in mini-batches of 32 samples, which we chose empirically. The optimiser and loss function arguments were set to Adam and binary cross-entropy, respectively. For the BERT models, we trained for three epochs and set the Adam optimiser's learning rate to 5e-5, as recommended by the paper's authors (Devlin et al., 2019), who identified a set of values that work well across all tasks.

To evaluate the models' performance for the relevancy detection, we used the macro F1 as the off-topic class still has few instances compared with the on-topic class even after the up-sampling. For example, the training data for the HF set has 9599 on-topic examples and only 2464 off-topic examples. The difference is 7135, which is relatively big. The macro F1 metric considers

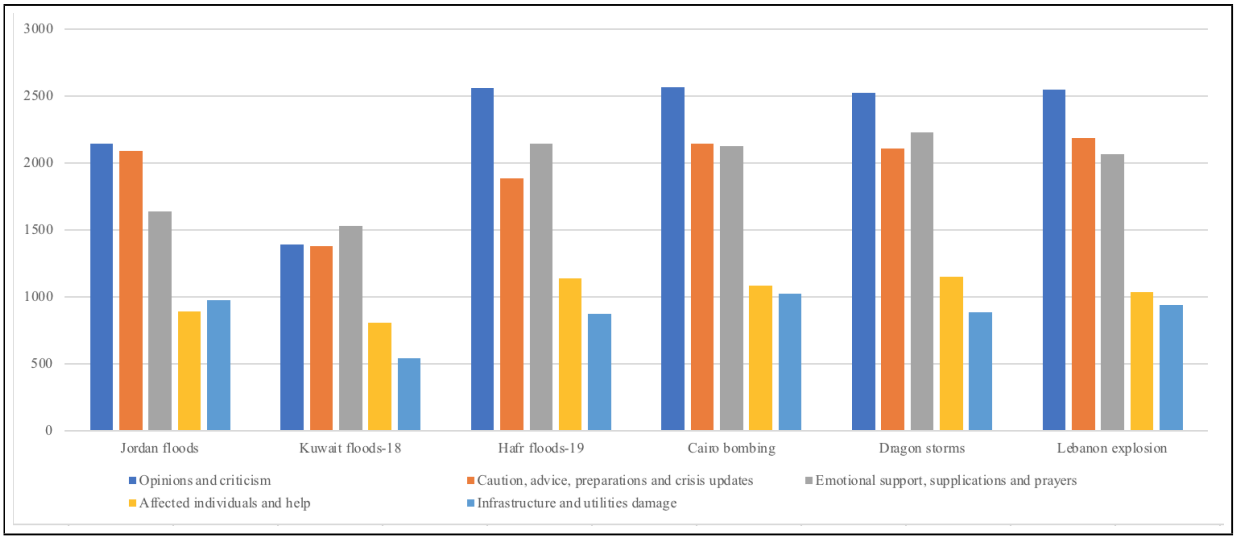

Figure 5.2: Training dataset label distribution (information types) for each target event after up-sampling the minority class

the class distribution in the corpus. The F1 score is the harmonic mean of precision and recall. Macro F1 is calculated by taking the arithmetic mean (unweighted mean) of all the per-class F1 scores. For information classification, we used the accuracy that tolerates partial errors (Godbole and Sarawagi, 2004; Sorower, 2010). Both accuracy and F1 scores produce values ranging between 0 and 1 where 1 indicates a perfect model.

5.5 Results

5.5.1 Results and Discussion

Table 5.4 presents the results of the supervised classifiers for the relevancy detection task. It is worth noting that most instances in the test set belong to the relevant class. The macro-F1 does not consider the support (weight) per class while averaging the F1 scores. That explains the low performance for the CB event, which only includes six negative examples. Table 5.5 shows the accuracy results for the information categorisation task. The best results in both tables were marked in bold. In the tables, we presented the average score for the performance of each model by

calculating the arithmetic mean of all its scores produced for the different target events. Results are discussed in the following to answer the second research question posed in Chapter 1, Section 1.3.

Table 5.4 indicates that all DNNs using AraVec (CBOW models) outperformed the SVM and DT models. Looking at the average F1 scores, the CNN-K, LSTM and CLSTM with the CBOW AraVec embeddings outperformed the $SVM_{(1,3)}$ by 6.46%, 4.13% and 6.83%, respectively. The DT showed the worst performance among the classical ML models. Decision trees do not generalise well to variations not seen in the training data (Bengio et al., 2010). They are liable to overfit the training data as they can create complicated trees that do not generalise well, mainly if the probability distributions differ between the training and test data. The NB classifier trained with the word features produced comparable results to DNNs regardless of the independence assumption. The NB predicts by finding the probability that each word in a document is positive or negative. Then, it multiplies these probabilities to obtain the final prediction, making the classifier efficient in identifying spam messages, such as advertisements, that do not include crisis-related words except the relevant hashtags, as we found when looking at the errors.

Compared to models that used only the unigrams and bigrams features, including the trigrams produced a nearly similar performance for the NB and SVM, which means capturing three-word expressions does not enhance these models' performance. On the other hand, using the trigrams with the unigrams and bigrams improved the generalisation performance for the DT model on the relevancy detection task. Although character n-grams are robust to grammatical errors (Kanaris et al., 2007) and can capture the morphological characteristics of the dialects (Eltanbouly et al., 2019), using such features resulted in low performance for the NB and DT. Regarding the SVM, using character n-grams enhanced the scores over the models that used the TF-IDF features in four cases but produced a close score on average (i.e. when considering all target events). We noticed that all classical ML models trained to classify the HF data had better results with the character n-grams than equivalent models with the BoW or TF-IDF features, indicating that character n-grams can be good discriminative features in some cases.

For the information classification problem, Table 5.5 demonstrates that all DNN models

Table 5.4: The macro F1 scores of classical ML and DNN models for the relevancy detection task

Model	JF	KF	HF	CB	CD	DS	BE	Avg.
NB_BoW$_{(1,3)}$	63.65%	66.74%	68.65%	52.32%	40.96%	69.98%	66.91%	61.32%
NB_BoW$_{(1,2)}$	65.08%	67.16%	66.61%	52.63%	40.08%	70.07%	68.13%	61.39%
NB_character	60.86%	62.27%	71.12%	45.50%	16.70%	67.29%	62.64%	55.20%
SVM_TF-IDF$_{(1,3)}$	61.42%	65.20%	63.94%	51.66%	38.11%	52.93%	65.69%	56.99%
SVM_TF-IDF$_{(1,2)}$	61.11%	66.87%	63.64%	51.78%	37.33%	53.49%	63.99%	56.89%
SVM_character	64.78%	63.73%	71.31%	49.03%	28.94%	57.97%	68.11%	57.70%
DT_TF-IDF$_{(1,3)}$	58.29%	52.43%	51.59%	50.65%	45.07%	55.90%	58.56%	53.21%
DT_TF-IDF$_{(1,2)}$	56.39%	51.58%	51.33%	48.13%	13.55%	57.99%	57.44%	48.06%
DT_character	53.74%	54.59%	53.83%	47.57%	40.16%	57.28%	52.55%	51.39%
CNN_CBOW	67.54%	66.86%	69.70%	51.36%	49.43%	64.24%	70.29%	62.77%
CNN-K_CBOW	68.43%	66.62%	69.95%	50.72%	**56.89%**	64.48%	67.10%	63.46%
LSTM_CBOW	67.94%	66.44%	72.45%	50.01%	42.72%	61.28%	67.00%	61.12%
GRU_CBOW	67.54%	66.33%	67.85%	52.83%	40.41%	60.68%	68.33%	60.57%
CLSTM_CBOW	70.35%	67.64%	68.60%	49.23%	52.28%	68.20%	70.43%	63.82%
CNN-K_random	61.49%	68.14%	68.60%	52.10%	50.57%	61.49%	58.96%	60.19%
LSTM_random	62.82%	66.34%	70.10%	49.37%	49.35%	62.33%	59.74%	60.01%
CLSTM_random	62.04%	67.90%	71.00%	50.25%	52.97%	61.57%	61.42%	61.02%
CNN-K_CBOW2	67.39%	66.71%	66.42%	49.93%	53.24%	62.70%	63.33%	61.39%
CNN-K_FastText	67.57%	66.75%	72.95%	**53.65%**	50.11%	60.21%	65.07%	62.33%
CLSTM_CBOW2	67.47%	71.73%	70.25%	49.14%	53.43%	61.07%	61.80%	62.13%
CLSTM_FastText	69.18%	69.89%	71.15%	49.79%	54.16%	62.70%	62.68%	62.79%
AraBERT	75.44%	80.50%	**76.39%**	49.79%	34.80%	**73.75%**	75.09%	**66.54%**
CNN-AraBERT	**77.69%**	**80.72%**	72.92%	49.50%	34.13%	70.65%	**79.55%**	66.45%

Table 5.5: The accuracy scores of classical ML and DNN models for the information classification task

Model	JF	JK	HF	CB	DS	BE	Avg.
NB_$\text{BoW}_{(1,3)}$	68.37%	55.43%	65.32%	60.17%	69.45%	55.84%	62.43%
NB_$\text{BoW}_{(1,2)}$	68.54%	54.91%	64.98%	61.39%	69.33%	57.88%	62.84%
NB_character	71.29%	59.05%	59.01%	55.70%	65.95%	50.36%	60.23%
SVM_$\text{TF-IDF}_{(1,3)}$	67.72%	53.90%	67.08%	64.75%	68.05%	46.60%	61.35%
SVM_$\text{TF-IDF}_{(1,2)}$	66.87%	53.77%	66.02%	64.51%	67.69%	46.65%	60.92%
SVM_character	69.43%	59.28%	55.04%	59.24%	63.86%	52.90%	59.96%
DT_$\text{TF-IDF}_{(1,3)}$	55.42%	38.78%	39.37%	49.02%	53.97%	40.18%	46.12%
DT_$\text{TF-IDF}_{(1,2)}$	55.92%	39.85%	41.98%	50.90%	57.22%	39.69%	47.59%
DT_character	57.58%	49.82%	39.15%	36.95%	49.27%	39.08%	45.31%
CNN_CBOW	77.34%	70.77%	71.37%	72.81%	76.44%	62.40%	71.86%
CNN-K_CBOW	79.01%	69.92%	72.77%	73.38%	75.30%	61.88%	72.04%
LSTM_CBOW	80.23%	71.95%	73.77%	73.81%	77.58%	65.53%	73.81%
GRU_CBOW	80.37%	71.57%	70.45%	76.38%	76.87%	65.05%	73.45%
CLSTM_CBOW	80.18%	69.79%	74.23%	73.10%	74.16%	60.66%	72.02%
CNN-K_random	77.23%	66.37%	69.38%	69.31%	70.80%	55.82%	68.15%
LSTM_random	76.99%	66.53%	67.04%	71.95%	70.73%	62.06%	69.22%
CLSTM_random	77.07%	64.44%	65.56%	70.02%	71.45%	60.14%	68.11%
LSTM_CBOW2	79.46%	72.35%	74.59%	74.02%	74.94%	62.46%	72.97%
LSTM_FastText	80.66%	70.88%	74.74%	74.38%	76.51%	66.55%	73.95%
AraBERT	**86.09%**	**80.78%**	**82.67%**	**80.82%**	80.11%	**83.00%**	**82.25%**
LSTM-AraBERT	85.20%	79.47%	80.39%	76.19%	**82.36%**	81.15%	80.79%

initialised from the CBOW model outperformed the classical ML classifiers in all cases. Again, the NB model produced the best results among the classical ML models, followed by the SVM classifier. The CNN model, which showed the lowest score among the DNNs, outperformed the best-performing ML model (i.e. NB_$BoW_{(1,2)}$) by 9% on average, confirming that DNNs outperformed classical ML classifiers on this task to a great extent. The superiority in performance of the deep models is more noticeable for the information categorisation problem. Regarding the classical models, we found that models trained on unigrams and bigrams produced results comparable to the equivalent models that used these features with the trigrams. Using the character n-gram features on this task resulted in lower performance on average. Nevertheless, they enhanced the accuracy scores over the n-grams' models for the JF and JK data, indicating that the performance of classical ML approaches with different feature selection methods varies across the datasets. For example, features that resulted in good ML performance for specific source-target pairs might produce poor performance for other cases.

We explored whether the DNNs outperformed the classical ML models due to their exploitation of the transfer learning through the pre-trained word embeddings by training the deep models from scratch without using any external knowledge (pre-trained embeddings). We experimented with one model from each architecture: CNN-K, LSTM and CLSTM. Results are presented in Tables 5.4 and 5.5. We attached the word "random" to the models' names, meaning models' weights are randomly initialised. In this case, text embeddings are learnt from scratch by the embedding layer based on the training dataset. For the relevancy detection, we found that the models' performance was still better than the SVM and DT and comparable to the NB classifier. At the same time, they surpassed the best-performing classical model (i.e. NB_$BoW_{(1,2)}$) on the information task by scores ranging from 5.2% to 6.3% on average. Results demonstrate the advantage of DNN methods as feature extractors. DNNs' generalisation ability is good as they can learn high-level abstract features through their deep structures. They are able to learn distributed representations of data using an embedding layer without external knowledge.

Results demonstrated that initialising the word vectors from pre-trained embeddings boosted

the performance of DNNs. Figures 5.3 and 5.4 depict the effect of using pre-trained word embeddings for the relevancy detection and information categorisation tasks, respectively. Using the CBOW pre-trained word embeddings learnt from Twitter data improved the average scores of the CNN-K and CLSTM models for the first task by 3.26% and 2.8%, respectively. The improvement for the LSTM might not be noticeable on the average score because of the low performance of the LSTM_CBOW on the CD data. However, it enhanced the BE, JF and HF scores by 7.26%, 5.12% and 2.35%, respectively. Concerning the second task, Table 5.5 shows that using the pre-trained CBOW enhanced the average accuracy scores by 3.9% for the CNN and CLSTM and by 4.6% for the LSTM model. The semantic meanings of words can be captured better through embeddings learnt from large corpora rather than being learnt from smaller data (Bojanowski et al., 2017) as our task-specific training set.

Looking at the DNNs initialised from the CBOW embeddings, Table 5.4 demonstrates that CNNs outperformed the RNNs on the relevancy classification. The best CNN model improved the average F1 scores by 2.3% over the best-performing RNN model (LSTM). It also trains faster. CNN models classify text by learning location invariant patterns extracted by convolutions and filters. At the same time, pooling operations can preserve the most salient information, which we can rely on to find critical features to classify relevant and irrelevant posts. However, they usually fail to detect irrelevant messages captured due to the keywords' ambiguity. Feeding the extracted CNN features into an LSTM enhanced the results over the CNN-K model of four cases, showing minor gains in performance on average.

The RNNs achieved slightly higher results while classifying tweets into informative types. As shown in Table 5.5, the RNNs produced better scores than CNNs for all events except one case (GRU for the HF event). For the BE, both RNNs outperformed the CNNs by around 3%. RNNs process sequences of data by retaining the memory of the previous state in the sequence. Thus, they can capture patterns from the whole context. Processing the entire sequence instead of depending on crucial local phrases is more suitable for this task. Regardless of the crisis task, using different architectures of CNN and RNNs resulted in a negligible difference in average scores. We

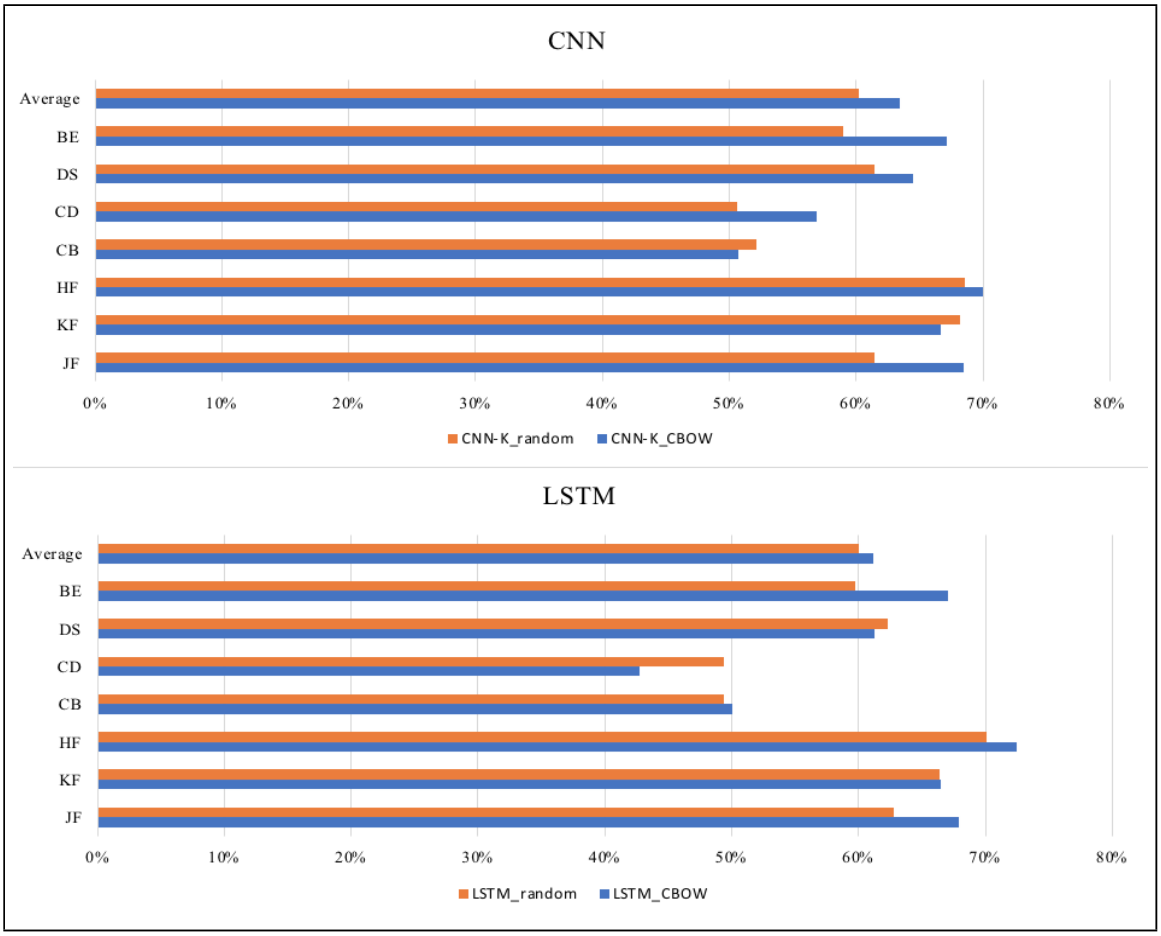

Figure 5.3: Effect of initialising the word vectors from pre-trained embeddings for the relevancy detection task

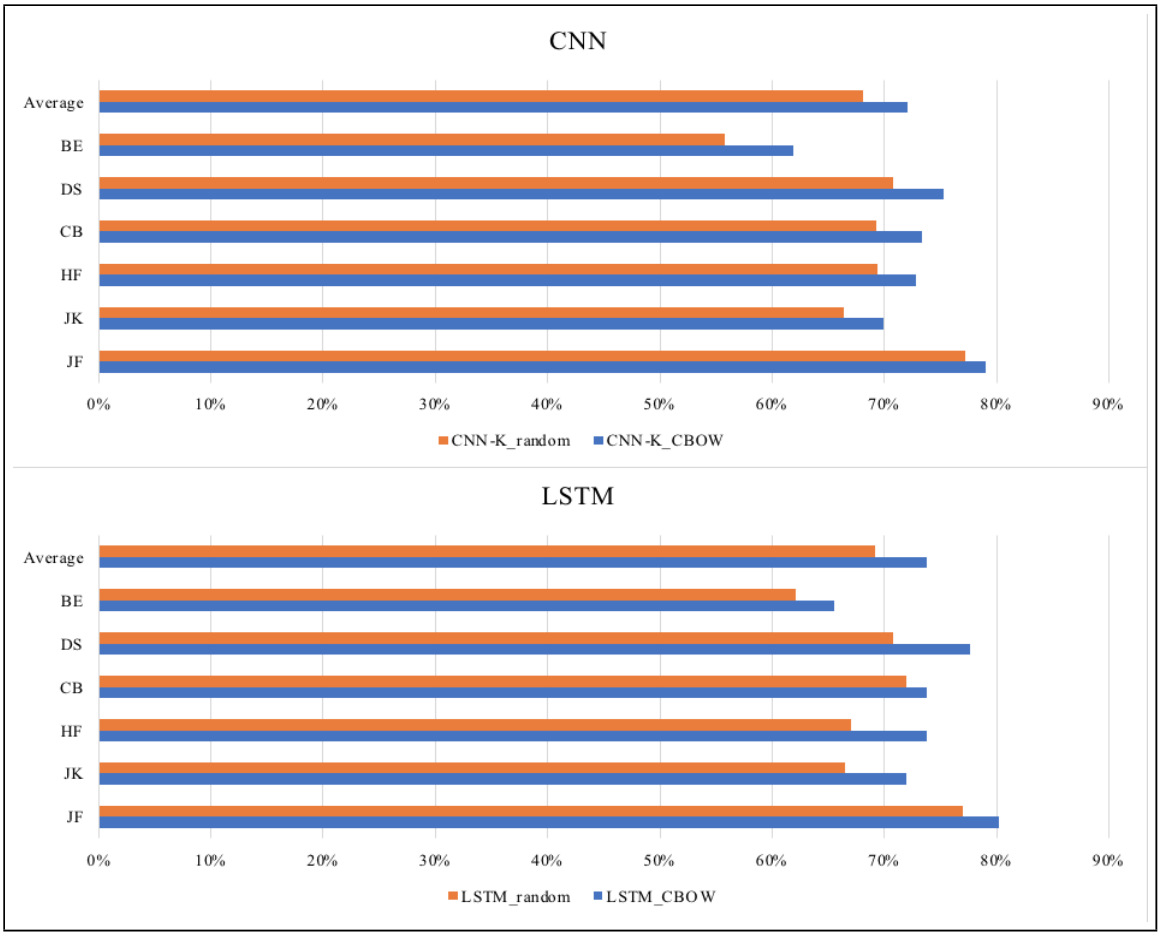

Figure 5.4: Effect of initialising the word vectors from pre-trained embeddings for the information classification task

also experimented with the Bi-LSTM and found no improvement over the LSTM model. Based on our results, LSTM produced performance slightly higher than the GRU network, which might be because the latter is a simplified variant of the former and it has a smaller number of parameters.

Given that character embeddings perform well for morphologically rich languages such as Arabic (Bojanowski et al., 2017), we evaluated the performance of DNNs with the FastText representation to explore whether using character-level embeddings instead of word embeddings improves the performance of DNNs. We trained the best-performing models for each task using the FastText embeddings and compared their performance to equivalent models with word2vec (CBOW) embeddings. To make fair comparisons, we experimented with a word2vec model generated from the exact Twitter corpus used for the FastText model. Both models have the same dimension size (i.e. 200) and have been created by A. I. Alharbi and M. Lee (2020) using the Gensim Library [2]. We called this word2vec model (CBOW2), which has been appended to DNN names in the results.

For relevancy detection, the overall scores of CNN-K_FastText and CLSTM_FastText were slightly better than the equivalent models initialised from the CBOW2. The character embeddings enhanced the performance of the CNN and CLSTM by 0.94% and 0.67%, respectively. We observed the same pattern for the second task. The LSTM_FastText yielded small gains in performance, accounting for 0.98% over the LSTM_CBOW2 model. Considering the best-performing DNN for each task, Figure 5.5 shows the models' results for each target event using FastText and CBOW2. Compared to the equivalent CBOW2 embeddings, FastText improved the performance slightly in most cases for both classification problems. We can conclude that using character-level embeddings improved the performance to a minimal extent. The main advantage of using FastText is its ability to infer the representations of OOV words. Using the word2vec model, we found that around 8% of the words in the Kawarith labelled dataset are OOV words. Most OOVs are misspelt words, such as مستشيفي (hospital) and dialectal words/phrases, such as هوصف، عميقطعوا (they're cutting, I'm describing). It can be noticed that models using the CBOW(AraVec) representations generally worked better than those with CBOW2. For example, the overall F1 of the CNN_CBOW is higher

[2] https://radimrehurek.com/gensim/index.html

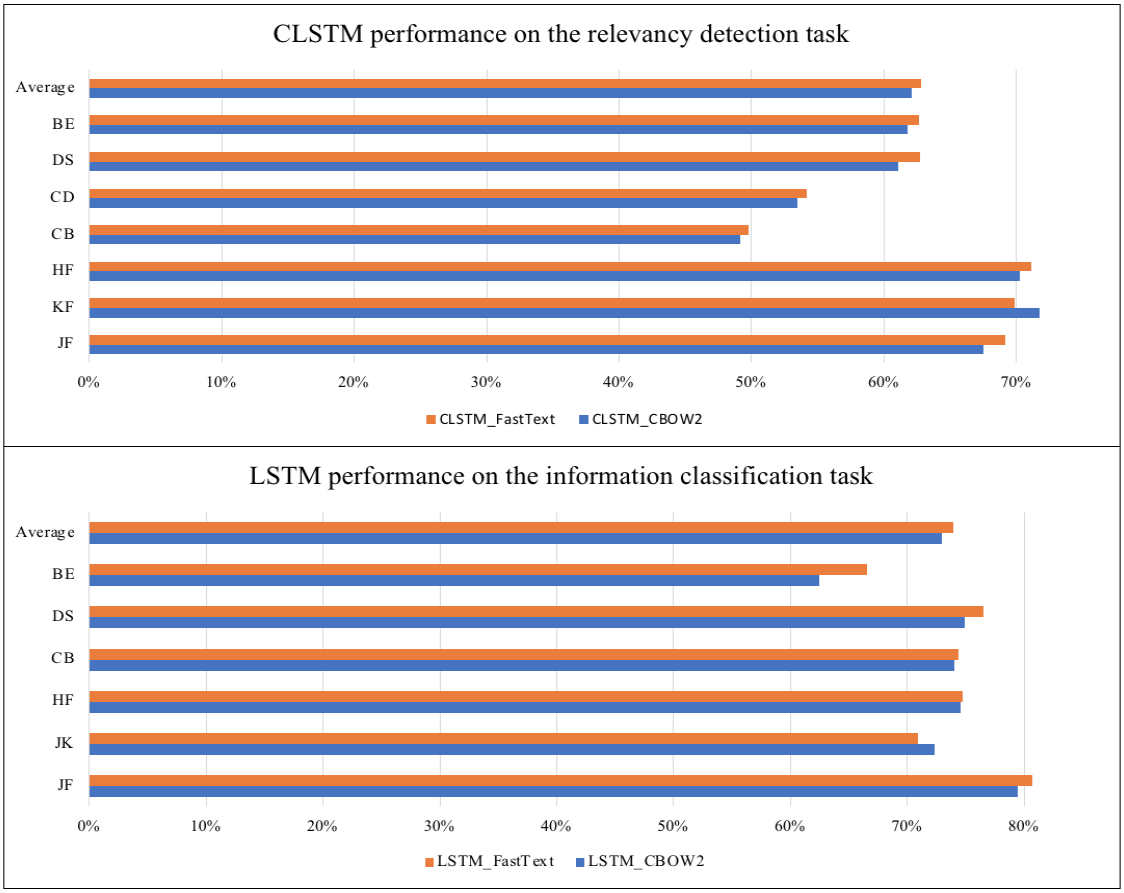

Figure 5.5: The performance of the character-level embeddings (FastText) versus the word-level embeddings (CBOW)

than that of CNN_CBOW2 by 2.07%. We relate this to the training set size for the two word2vec models, as the Twitter corpus is much more extensive for the AraVec model.

We also experimented with the contextualised BERT models to investigate whether they result in an improved performance. We found that using the contextualised BERT embeddings enhances the performance of DNNs on both tasks, whether it has been fine-tuned with a linear layer (AraBERT models) or been used as features (CNN- LSTM-AraBERT). The BERT models yielded the best performance in most cases. For the relevancy detection task, the AraBERT model improved the average macro F1 scores over the CLSTM_CBOW model by 2.72% and produced results comparable to the CNN-AraBERT. The AraBERT surpassed the CNN_CBOW by 12.86% and 7.79% for the KF and HF events, respectively. However, the BERT models perform poorly

in predicting the COVID-19 data, producing results less than other DNNs. For example, the CNN_CBOW and CLSTM_CBOW models significantly outperformed the AraBERT classifier by 22% and 17%, respectively. We will shed light on the reason later.

Regarding the information categorisation, the overall average score of the AraBERT is higher than that of the LSTM-AraBERT by 1.45%. Fine-tuning the BERT models on the data using a linear layer performed slightly better than the LSTM model that uses the BERT embeddings as features. The AraBERT model also trains faster than the LSTM-AraBERT. Averagely, the AraBERT surpassed the LSTM_FastText by 8.29%. The performance gains range from 16.45% to 3.6%. Thus, the use of BERT enhances the performance to a large extent. The BERT models have better generalisation power as their representation can capture semantic and syntactic properties of natural language in context. AraBERT uses the Farasa segmenter which enables the model to learn shared representation between words with comparable structures.

5.5.2 Error Analysis

We looked at the misclassified examples of the best-performing classifiers (i.e. BERT models) to understand the models' limitations and see how to improve their performance further. Table 5.6 presents some misclassifications for the relevancy identification task. We noticed that models mistakenly classified some general or personal supplications as relevant as in the first example in the table. In total, we found 21 misclassified examples of such messages from flood data. Muslims believe that supplications are accepted when it rains and post such tweets at the beginning and during the event that include relevant hashtags. We believe that the presented example was detected as on-topic because the training set had similar messages from the COVID-19 event, conveying prayers for COVID-19 patients. They were labelled as on-topic (emotional support) as they are related to the COVID-19 crisis. On the other hand, the misclassified instance shows a general supplication (i.e. not prayers for the crisis victims). We also found two off-topic examples, such as the second post in the Table, that were mistakenly classified as on-topic because they were about requesting help. Those tweets were captured in our data as they included some relevant hashtags. These errors

confirmed the difficulty of the crisis detection task.

Other mispredictions are opinions misclassified as irrelevant, such as the third example in Table 5.6. Such tweets are sarcastic, expressing negative views. We noticed ten examples of such errors when classifying the KF data. We also observed a few cases for tweets labelled as irrelevant but classified as related. They were somehow ambiguous comments about some topics that could be related to the crisis in some way. The fourth example shows a tweet predicted as related to the explosion but annotated as irrelevant. We think this post should have been labelled as relevant since the author likened the Beirut explosion to the Hiroshima bombing. Some misclassified tweets were written using dialectal words and included spelling mistakes, as in the fifth example, which was entirely written in the Egyptian dialect. Looking at the errors, 83% (10 out of 12) of the misclassified on-topic examples in the DS data were dialectal tweets. We also found 36% of misclassified relevant tweets in the HF were written in dialects of the Arabian Peninsula.

Examples 6 and 7 were crisis-related messages labelled as off-topic because they discuss different disasters, not the event under consideration. The former was captured during the Dragon storms. The latter is related to the KF but also crawled while collecting the Jordanian data. The existence of crisis tweets of a particular type in the irrelevant class of training data would result in false negatives when classifying new events from the same type. This is a reason behind the low performance of models trained to classify the COVID-19 data. We found many COVID-19 tweets captured in the DS set, most of which were reports about new cases. Those posts were labelled as irrelevant to the DS. Hence, the classifier learnt to identify messages about new patients as off-topic. To confirm this, we excluded the event occurring after the pandemic from the training set. We found significant performance gains for the classifiers, as shown in Table 5.7 for the AraBERT and CLSTM_CBOW models. We suggest removing any crisis-related tweets from the irrelevant class to improve the generalisation.

We also explored whether using data only from similar-type events as a source enhances the model's performance of relevancy detection classifiers over the same model that uses all historical data with multiple cross-type events. We examined the performance of the AraBERT and CLSTM

Table 5.6: Examples of misclassifications by the BERT model for the relevancy detection task

#	Example	True Label	Prediction
1	اللهم اشف مرضى السرطان اللهم اشف مرضى المسلمين. اللهم اغفر لنا وارحمنا يا ارحم الراحمين. #العاشره #امطار_الكويت "Oh God, heal cancer patients. Heal Muslim patients. Oh God, forgive us and have mercy on us. You are the most merciful of all. #Tenth #Kuwait_rain"	Off-topic	On-topic
2	#سيول_الأردن سألتكم بالله من يشوف تغريدتي هذي يعمل لها ريتويت لعلها توصل ل أهل القلوب الرحيمه. أقسم بالله وبجلال الله عندي أطفال معاقين وبنت مريضة فشل كلوي. احتاج قيمة طعامهم ومصاريفهم وكسوتهم. "#Jordan_floods I ask by God who sees this tweet, retweet it as it may reach people of compassionate hearts. I swear by God, and by his glory, I have handicapped children and a girl who is ill with kidney failure. I need money for their food, expenses, and clothing."	Off-topic	On-topic
3	كانت الابوريتات الغنائيه أهم من البنية التحتية. #الكويت_تغرق_بالفساد "The lyrical operettas were more important than the infrastructure. #KuwaitDrownInCorruption"	On-topic	Off-topic
4	رغم فداحة انفجار مرفأ بيروت، فيه أحد يذكره بهيروشيما؟ "Despite the enormity of the Beirut port explosion, does this remind anyone of Hiroshima?"	Off-topic	On-topic
5	خدمات اي حد ف وسط سيناء ومش عارف يروح، بيتي تحت امرو ودا رقمي *** #عاصفة_التنين "Service! Anyone in the middle of Sinai who can't return, my house is there and this is my number *** #Dragon_storms"	On-topic	Off-topic
6	بعد نفيها الأرصاد الجوية تؤكد هزة أرضية خفيفة في البحرين بقوة ٣.٠١ ريختر أمس. "After its denial, the Meteorological Department confirmed that a minor earthquake hit Bahrain yesterday, measuring 3.01 on the Richter scale."	Off-topic	On-topic
7	استقالة وزير الأشغال الكويتية على خلفية غرق الكويت بالكامل بمياه الأمطار. #سيول_الاردن #الكويت_الان "The resignation of the Kuwaiti Minister of Public Works as Kuwait was entirely flooded with rainwater. #Jordan_floods #Kuwait_now	Off-topic	On-topic

Table 5.7: The macro F1 scores of the AraBERT and CLSTM models trained on selected events to identify on-topic posts for the COVID-19 and flood events

Model	Training Set	Target event	Macro F1
AraBERT	JF + KF + HF + CB	CD	63.29%
AraBERT	KF + HF	JF	77.27%
AraBERT	JF + HF	KF	78.77%
AraBERT	JF + KF	HF	79.28%
CLSTM_CBOW	JF + KF + HF + CB	CD	65.35%
CLSTM_CBOW	KF + HF	JF	63.95%
CLSTM_CBOW	JF + HF	KF	65.29%
CLSTM_CBOW	JF + KF	HF	65.36%

on flood data using the leave-one-event-out strategy. Again, we up-sampled the minority class. Results are presented in Table 5.7. We found that the AraBERT's scores were improved for the JF and HF by 1.83% and 2.89%, respectively. However, using all multi-source data produced a higher F1 score for the KF. For the CLSTM, using all training data generated better results for all flood events. DNNs using conventional pre-trained word2vec still work better using larger training data. On the other hand, BERT achieves good generalisation with smaller labelled data from a domain similar to the target event. These findings highlight a possible direction of research that will be investigated in the following chapter. We will explore whether fine-tuning the AraBERT on a subset of training data that is most similar to the target event would result in performance gain compared to the equivalent models that use all training data.

As the BERT model generalises well using smaller data from same-type events, we explored whether the down-sampling technique produces results higher than the up-sampling on same-type crisis classification. Instead of up-sampling the minority class, we shuffled the training set and randomly down-sampled the majority class to have an equal class distribution. We repeated the experiment for each target flood event five times by choosing a different training split for each run. The mean and standard deviation scores are reported in Table 5.8. First, we found that training on all data with up-sampling the minority class resulted in performance gain over using the down-sampling method. For the down-sampling scores, high standard deviations indicate a variance in the F1 scores around the means. Results showed that some randomly chosen subsets of the on-topic posts produced higher scores than others even when training on data with the same crisis type, which also motivated us to find an optimal subset of training data that improves the model generalisation for the target crisis.

For the information categorisation, we noticed all BERT models had the lowest performance for the minority class: 'infrastructure damage'. We also found that most mispredictions were related to the 'cautions, preparations & other crisis updates', as their topics can vary across events (particularly for the BE event). The models managed to identify posts related to emotional support and prayers accurately. These messages usually have the same Arabic phrases, regardless of the

Table 5.8: The macro F1 scores of the AraBERT model trained on same-type events to identify on-topic posts using different re-sampling techniques: up-sampling and down-sampling

Training Set	Target Event	Macro F1 (up-sampling)	Macro F1 (down-sampling)
KF + HF	JF	77.27%	69.56% (+/- 5.3%)
JF + HF	KF	78.77%	66.38% (+/- 5.4%)
JF + KF	HF	79.28%	79.94% (+/- 2.2%)

event. Overall, 19% of errors resulted from mispredicting the multi-labelled instances. We noticed that the classifier usually made partial mistakes. It partially predicted the correct labels of 262 out of 317 cases, which accounts for 82%. For example, the model recognised the tweet below as 'affected individuals & help' but missed the warning updates (caution class).

الأردن #عاجل. وفاة طفلة بسبب السيول الجارفة اليوم الجمعة وفقدان عدد من الأشخاص في عدة مناطق. وتشهد الأردن ومناطق واسعة من #السعودية موجة من الأمطار الغزيرة وتشكلا للسيول. ويتوقع اشتداد الحالة الجوية هذا المساء وغدا السبت. #سيول_الاردن #وسم #غدق #طقس

"#Jordan #Breaking. A child has died in flash flooding this Friday and several people are missing in many areas. Jordan and wide regions of Saudi are witnessing heavy rainfall leading to floods. Severe weather conditions are expected this evening and tomorrow. #Jordan_flood #wasm #Ghadag #weather"

Interestingly, the classifier assigned two labels to the post below: 'affected individuals & help' and 'emotional support, prayers & supplications'. However, the tweet was annotated as affected individuals as two annotators missed the condolences segment.

المتوفيين في ضبعة معلمة واثنتين من بناتها. نسأل الله لهن الرحمة، والمغفرة والفردوس الأعلى. إنا لله وإنا إليه راجعون #مادبا #الاردن #سيول_الاردن #منطقة_ضبعة #عمان

"The fatalities in Dabaa are a teacher and two of her daughters. We ask God for their mercy, forgiveness and the highest paradise. Verily we belong to Allah, and to Him we return. #Madaba #Jordan_floods #DabaaRegion #Amman"

Solving the multi-labelled problem is hard. Nevertheless, the BERT models trained on a

multi-source crisis dataset generalised well for unseen target crises. In the following chapter, we will also investigate whether using a data selection adaptation approach would enhance the performance on this task.

5.6 Conclusion

This chapter presented an empirical evaluation of three widely adopted classical ML approaches and different architectures of DNNs for Arabic Twitter cross-crisis classification. Using training data from multiple historical events, models were evaluated on two tasks: relevancy detection and information categorisation. DNNs were found to have better generalisation ability. We also conducted experiments using different text representations and explored whether they enhance the models' performance on unseen crisis data. We found that BERT-based models remarkably surpassed other models, whether they were fine-tuned using a linear layer or used as features fed to a DNN model. The BERT models are more robust in classifying out-of-distribution target events and generalise well using relatively small training data. Based on the error analysis, we suggest conducting a data selection approach to adapt the model for the target crisis. In the next chapter, we will investigate whether adopting a data selection approach would enhance the BERT's results.

Chapter Six

Classifying Crisis Tweets using Data Selection Models

In the previous chapter, we evaluated several models for cross-event crisis classification using training data from multiple emergency events. We have demonstrated that DNNs generalised better than conventional ML approaches, and the BERT-based models achieved the best generalisation performance. The results showed that the BERT models performed better when trained on in-type events (i.e. floods) than those that used all training data. This outcome motivated us to explore whether we can find a subset of source data similar to the target event to be used for training the classifier and boosting its performance. Thus, we propose a selection-based domain adaptation approach. This chapter explains our unsupervised multi-source data selection approach using the K-nearest neighbours algorithm. The strategy aims at building a good model for target data by leveraging labelled data from several related source domains and unlabelled data from a target domain. The chapter also presents the experimental study and results.

6.1 Multi-source Data Selection for Crisis Classification

As mentioned in Chapter 3, we performed a selection-based domain adaptation approach. In the NLP literature, the notion of domain typically refers to a specific corpus that differs from other domains in the topic, genre, style, etc. In this work, we define a domain as a dataset that has been

collected from SM for a specific crisis, so that each crisis data represents a distinct domain. We hypothesise that fine-tuning MLMs on the most similar data can produce more accurate results on crisis classification tasks than using all multi-source training data.

As in the real scenarios, we assume that we are provided with labelled examples from past crises (multi-source datasets) and only unlabelled data from an emerging crisis (target). The goal is to find an optimal set of training data from a multi-source domain that enhances the model generalisation for the target data. Finding this set can be achieved by identifying the instances from training source data that are similar (as close as possible) to the target domain. Thus, we consider an instance-level data selection strategy using cosine similarity in the embedding space of the pre-trained MLMs. Cosine similarity is used widely in information retrieval to calculate the similarity between two documents by measuring the cosine of the angle between their vectors in the embedding space (Rahutomo et al., 2012). Similar vectors would have the same orientation. In data selection, similarity values are calculated between the query document and all documents in the source domain, and those documents with the highest similarity scores will be selected. The selected examples are expected to have similar feature distribution to the target domain of interest. In our case, the query document is a tweet from the target domain. The selection process will be done iteratively for each document in the target data, and the selected data from the source will be added to the training set. The cosine similarity between two vectors x and y is calculated as follows:

$$cos(x, y) = \frac{x.y}{\|x\|.\|y\|} \tag{6.1}$$

where $\|x\|$ represents the magnitude of the vector, and $x.y$ is the dot product between x and y.

In our approach, we selected the data only from the crisis-related messages. Due to the imbalanced dataset, we used all off-topic posts for the relevancy detection task. Such a strategy would result in relatively balanced training datasets for the relevancy detection task, as the data selection from the on-topic set will down-sample the majority class. Messages labelled irrelevant to a particular crisis but related to other emergency events were excluded from the off-topic collection, as depicted in Figure 6.1. For example, the Jordanian flood data contain two tweets about the

Kuwait flood. Those posts were labelled irrelevant to Jordan flood because the annotation was performed per event level. We filtered out such posts from the off-topic training set as they can adversely impact the performance, as explained in the previous chapter regarding the COVID-19 case. Hence, the classifier will be trained to learn whether or not a post represents a crisis, as the aim is to build a model for cross-event crisis detection.

We leveraged the contextualised text representations produced by pre-trained language models. In this work, we used Sentence-BERT (S-BERT) (Reimers and Gurevych, 2019). The authors created S-BERT by fine-tuning BERT using a Siamese network architecture to produce fixed-sized sentence embeddings that can be compared using similarity measures. As there is no monolingual Arabic S-BERT, we used the multilingual S-BERT (Reimers and Gurevych, 2020). We used the K-nearest neighbours algorithm to select the K most similar instances for each example in the target set based on the cosine similarity on the S-BERT embedding space. The selected data was used to fine-tune a BERT model for classifying target tweets as outlined in Algorithm 1. We used BERT as a classifier as it outperformed other DNNs models on cross-crisis classification. Fine-tuning a model pre-trained on large data eliminates the need for massive training examples that are required to train a DNN from scratch. We also investigated the effect of combining the data selection approach with self-training. In this work, we used a hard-labelled approach for retraining, in which most confidently classified instances are added with their predicted labels (e.g. 0 or 1) to the training set in subsequent training iterations. Figure 6.1 illustrates the domain adaptation framework that combines data selection and self-training.

6.2 Experimental Setup

For data selection, instances are encoded using the distiluse-base-multilingual-cased-v1 multilingual model[1] that supports 15 languages, including Arabic. K is the number of nearest neighbours. We experimented with different values of K (3, 5 and 10). We evaluated the binary classification model

[1] https://www.sbert.net/docs/pretrained_models.html

Algorithm 1: Multi-source instance-level data selection

Input:

S_L: { $S_1 \cup S_2 \cup \cdots \cup S_n$ } Labelled source domain examples from n historical crisis data

T_U: Unlabelled target domain data for a new crisis

SET *trainset* to [ ]

$S_R = S_L$[label=relevant]

$S_I = S_L$[label=irrelevant]

Remove duplicates in T_U

Encode data in S_R using S-BERT

FOR EACH instance in T_U **DO:**

Encode instance using S-BERT

Select the nearest k instances S_k from S_R

$trainset.append(S_k)$

END LOOP

Remove duplicates in *trainset*

$trainset.append(S_I)$

$trainset.shuffle()$

Output: *trainset* to **Fine-tune** a BERT model M for classifying T_U

on each selected set. The information type classifier was assessed on the last setting (K=10). This is because some classes can be under-represented in training set when we set the number of nearest neighbours to 3 or 5, as the target data is imbalanced. AraBERT was used as a classification model as it outperformed other DNNs models.[2] We fine-tuned AraBERT using the same parameters and text pre-processing steps introduced in the previous chapter. For reproducibility, the random seed was set to 1. For self-training, we set the confidence threshold and the number of iterations to 0.99 and 2, respectively.

Models were evaluated on the Kawarith corpus. We manually removed those messages related to other crises from the off-topic set in the source data. We found 30 such tweets in the dataset, accounting for 1.6% of the total irrelevant tweets, most of which (18 posts) were in the

[2]We found that AraBERT (trained on news corpus) slightly outperforms other Arabic BERT variants (trained on Twitter corpora) on our task.

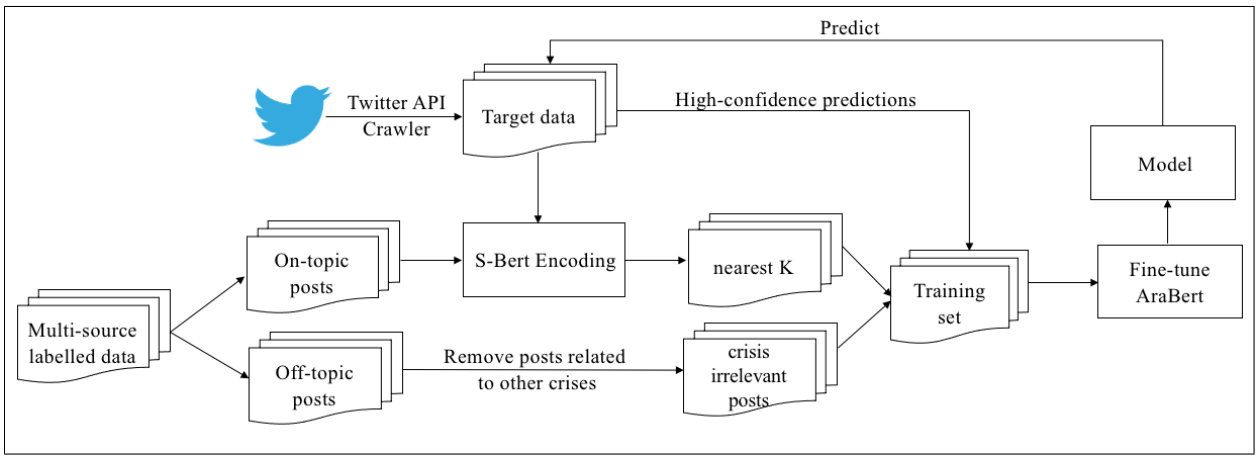

Figure 6.1: The Domain adaptation approach with data selection and self-training

Dragon storm data. We did not perform up-sampling for the minority class. Training on all data, we found that BERT is generally capable of handling imbalanced class distribution. Besides that, the random over-sampling can increase the likelihood of over-fitting the training data as it creates exact copies of existing instances.

We experimented with different source and target crisis pairs using the leave-one-event-out strategy. For self-training, the evaluation was performed using 3-folds cross-validation. The target data was split into three parts: one for testing and the rest (two-thirds) for adaptation. We report the average score. We used the weighted F1 and macro F1 to evaluate the models' performance on the relevancy detection task as they consider the imbalanced class distribution. The weighted F1 is computed by taking the mean of all per-class F1 scores, considering the proportion for each class in the dataset. For information classification, we used the accuracy and macro F1 score.

In the following section, we will present the results of our domain adaptation classifiers: the data selection method and the data selection with self-training. In this area of research, there is a lack of comparison between the existing proposed methods, including the domain adaptation approaches, due to a lack of standardised datasets and evaluation frameworks. We could not compare our adaptation approach to the ones in previous studies as we used a different dataset, and the proposed systems are not publicly available. Instead, we will compare the performance of our models with their counterparts fine-tuned utilising the entirety of the historical source data (baseline-1).

Besides that, we compare them against the self-training model (baseline-2) as self-training with BERT shows excellent results in one of the most recent works on English Twitter crisis detection (H. Li et al., 2021). It is worth emphasising that we used a hard-labelling self-training strategy, while H. Li et al. (2021) adopted soft-labelling.

6.3 Results and Discussion

We will discuss the results for each task separately. Table 6.1 presents the results of the proposed models and baselines on the relevancy detection problem. The best results were marked in bold. The number of training examples differs across events for each data selection model because our target datasets have different sizes, and we performed an instance-based data selection method. Besides that, the same tweets could be selected from the source as nearest neighbours for different target instances. Such duplicates were removed, as shown in Algorithm 1. In the following, we discuss the results in light of research questions: ***RQ3***, ***RQ4*** and ***RQ5***.

RQ3) *How do the results of the proposed data selection model compare to the results of the BERT model that learns from all source data?*
We found that fine-tuning BERT on the most similar data (BERT-DS models) improves performance in all cases over BERT(all) despite using smaller training data. On average, choosing different values of K results in slightly different performance. BERT-DS(k=10) achieved the best scores in four out of seven cases. We compared the best model with the baselines.

We found that BERT-DS(k=10) improved the average weighted F1 and macro F1 over BERT(all) by 3.67% and 4.57%, respectively. The improvement in weighted F1 scores ranges from 1% to 7.53%, whereas the macro F1 improved by values ranging from 1.38% to 9.93%. We observed that the macro F1 was enhanced by 9.93%, 6% and 5.79% when classifying KF, HF and CD, respectively. The pronounced enhanced performance for classifying CD emphasised the usefulness of domain adaptation based on data selection, as the features differ substantially between COVID-19 data and other crises.

Table 6.1: The weighted F1 and macro F1 scores for the relevancy detection task (DS and ST refer to the data selection and self-training techniques, respectively. The keyword (all) indicates training on the whole labelled data. K is the number of most similar instances.)

Target Data	Model	# of human-labelled examples	Weighted F1	Mac. F1
JF	BERT(all)	10418	93.26	73.52
	BERT-DS(K=3)	3901	94.08	73.89
	BERT-DS(K=5)	4724	94.19	76.32
	BERT-DS(K=10)	6082	94.45	76.14
	BERT-ST(all)	10418	94.85	75.95
	BERT-DS(K=10)+BERT-ST	6082	**95.06**	**77.83**
KF	BERT(all)	8317	87.40	72.39
	BERT-DS(K=3)	3552	90.95	78.34
	BERT-DS(K=5)	4227	90.50	76.83
	BERT-DS(K=10)	5329	93.69	**82.32**
	BERT-ST(all)	8317	**95.83**	77.51
	BERT-DS(K=10)+BERT-ST	5329	95.18	77.94
HF	BERT(all)	10801	76.39	78.36
	BERT-DS(K=3)	3107	**85.25**	**84.54**
	BERT-DS(K=5)	3850	80.34	81.69
	BERT-DS(K=10)	5137	82.40	83.63
	BERT-ST(all)	5137	71.07	67.99
	BERT-DS(K=10)+BERT-ST	10801	73.86	85.25
CB	BERT(all)	11710	96.51	55.43
	BERT-DS(K=3)	2783	97.63	56.58
	BERT-DS(K=5)	3169	**98.21**	**65.39**
	BERT-DS(K=10)	3966	97.94	58.01
	BERT-ST(all)	3966	96.96	61.52
	BERT-DS(K=10)+BERT-ST	11710	97.29	63.46
CD	BERT(all)	10412	64.11	49.88
	BERT-DS(K=3)	4071	69.56	52.95
	BERT-DS(K=5)	5044	69.20	53.82
	BERT-DS(K=10)	6484	71.64	55.67
	BERT-ST(all)	6484	76.29	61.21
	BERT-DS(K=10)+BERT-ST	10412	**84.17**	**66.03**
DS	BERT(all)	11424	77.71	71.24
	BERT-DS(K=3)	3115	**84.55**	**80.81**
	BERT-DS(K=5)	3809	81.36	76.31
	BERT-DS(K=10)	5092	78.70	72.62
	BERT-ST(all)	11424	77.60	71.13
	BERT-DS(K=10)+BERT-ST	5092	79.45	73.66
BE	BERT(all)	11414	86.72	78.1
	BERT-DS(K=3)	3019	87.8	79.04
	BERT-DS(K=5)	3628	**90.54**	**82.76**
	BERT-DS(K=10)	4824	89.72	81.75
	BERT-ST(all)	11414	87.77	77.49
	BERT-DS(K=10)+BERT-ST	4824	89.01	79.32

Table 6.2: The accuracy and macro F1 scores for the information classification task (DS and ST refer to the data selection and self-training techniques, respectively. The keyword (all) indicates training on the whole labelled data. K is the number of most similar instances.)

Target Data	Model	# of human-labelled examples	Accuracy	Mac. F1
JF	BERT(all)	6913	82	75.97
	BERT-DS(K=10)	3743	87.44	81.65
	BERT-ST(all)	6913	**88.78**	**82.42**
	BERT-DS(K=10)+BERT-ST	3743	87.36	79.29
KF	BERT(all)	5094	79.01	77.32
	BERT-DS(K=10)	3148	**81.23**	**78.86**
	BERT-ST(all)	5094	73.09	68.84
	BERT-DS(K=10)+BERT-ST	3148	71.45	56.55
HF	BERT(all)	7817	84.27	81.56
	BERT-DS(K=10)	3397	83.73	82.98
	BERT-ST(all)	7817	**85.18**	**82.97**
	BERT-DS(K=10)+BERT-ST	3397	83.35	79.65
CB	BERT(all)	8095	77.05	71.36
	BERT-DS(K=10)	1851	77.48	72.53
	BERT-ST(all)	8095	78.20	72.58
	BERT-DS(K=10)+BERT-ST	1851	**79.09**	**74.92**
DS	BERT(all)	8094	78.14	75.08
	BERT-DS(K=10)	3023	**81.88**	81.18
	BERT-ST(all)	8094	79.87	81.06
	BERT-DS(K=10)+BERT-ST	3023	81.83	**81.46**
BE	BERT(all)	7962	78.14	75.08
	BERT-DS(K=10)	2694	79.89	80.41
	BERT-ST(all)	7962	**83.44**	**77.72**
	BERT-DS(K=10)+BERT-ST	2694	76.70	65.01

RQ4) *How do the results of the proposed data selection model compare to the results of the self-training approach?*

First, we explored how the self-training approach compares to the BERT model that learns from all source data. We noticed that the self-training (BERT-ST(all)) model achieved higher results than BERT(all) in four cases and comparable results in two cases. The BERT-ST(all) model improved the weighted F1 by 12.18% in the COVID-19 case. On average, it enhanced the weighted F1 and macro F1 by 1.82% and 2.27%, respectively. BERT(all) worked better than BERT-ST(all) for the HF data. The reason was that many irrelevant tweets from HF were misclassified and added to the classifier in the next iteration as ground truth data, which degraded the performance.

Regarding ***RQ4***, we found that BERT-DS(K=10) outperformed the BERT-ST(all) model in three cases. The data selection model improved the macro F1 over the self-training approach on KF, HF and BE by 4.38%, 11.21% and 2.43%, respectively. BERT-ST(all) achieved comparable scores in two cases: JF and DS. However, the BERT-ST(all) model worked better for the CD data. It surpassed BERT-DS(K=10) by 5.54% and 10.36% for the weighted and macro F1 scores, respectively. Averagely, the data selection and self-training models produced comparable results. BERT-DS(K=5) outperformed BERT-ST(all) in four cases. The self-training approach surpassed the selection models when there was a significant feature distribution gap between the sources and target (as for the COVID-19 case), shifting the weights gradually towards the target data. Otherwise, the DS models generally worked better.

RQ5) *Does combining the proposed selection method with self-training result in performance gain?*

RQ5 explores whether self-training enhances the performance of the DS models. To answer this question, we combined the self-training strategy with the best DS model BERT-DS(K=10). We found that BERT-DS(K=10)+ST achieved the highest scores on the CD data. It enhanced the weighted F1 and macro F1 by 12.53% and 10.36%, respectively. However, BERT-DS(K=10) produced a higher performance for KF and HF. Otherwise, they achieved comparable results in two cases. Hence, adding the pseudo-labels to the training data does not constantly improve the performance. We recommend using self-training on the relevancy detection task when the target event

is very different from the source data, as in the case of COVID-19 and other disasters.

For the relevancy detection task, we excluded the crisis messages from the off-topic set to reduce the false negatives (crisis messages classified as not crisis). However, we still need to handle the irrelevant messages detected as relevant because they are about another crisis. For example, we found instances related to COVID-19 were classified as on-topic in the DS and BE data because examples of CD were chosen as the most similar data and were added to the selected set with their positive labels, which resulted in false positives. As we mentioned earlier in Chapter 4 that one crisis can be discussed using hashtags from another. To demonstrate this, we performed a 2D visualisation by t-SNE (Van der Maaten and Hinton, 2008) for randomly selected samples from COVID-19, Dragon storms and Beirut explosion using the S-BERT embeddings. Figure 6.2 demonstrated that many tweets collected during Dragon storms were related to COVID-19. Such messages can be considered outliers, which can be identified using outlier detection techniques such as clustering-based approaches (P. Chauhan and Shukla, 2015) after the crisis detection task. We left this for future work.

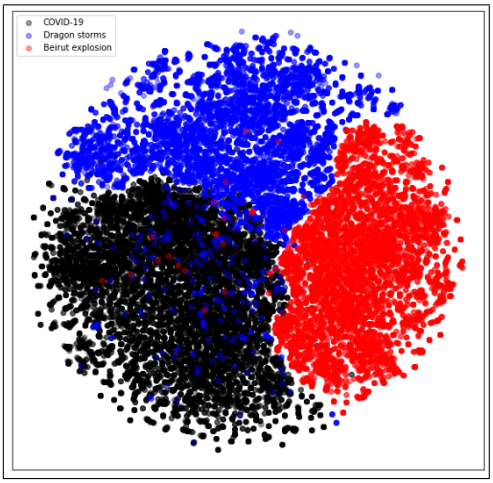

Figure 6.2: 2D visualisation of S-BERT embeddings for randomly selected samples from COVID-19, Dragon storms and Beirut explosion.

Regarding the information category classification, we set the K value to 10. This task has been assessed separately, i.e. we suppose that we managed to filter out all irrelevant posts and need to categorise the crisis-related tweets into pre-defined information types. Table 6.2 displays

the results of our experiments. When training the model on the chosen data, we obtained further improvements in macro F1 scores ranging from 1.17% to 6.73% absolute gains over BERT(all). Overall, data selection improved the performance in five out of six cases (***RQ3*** for the second task). Similarly, BERT-ST(all) generally worked better than the BERT(all) model. Our proposed approach worked better in identifying 'cautions, preparations & other crisis updates' tweets.

Comparing the performance of the data selection model with self-training on the information classification task, we found that the average of all scores showed that BERT-ST(all) and BERT-SD achieved comparable results (***RQ4***). BERT-ST(all) produced higher scores in four cases. Averagely, BERT-SD(K=10)+ST resulted in lower performance on this task compared to the two adaptation methods. For ***RQ5***, we found that combining the data selection with self-training does not improve the performance on information categorisation task. We recommend using the data selection domain adaptation for information type classification as it trains much faster and because the self-training could damage the performance as in the KF event, which failed to detect many cases related to the 'affected individuals' class. Our results on both tasks demonstrate the effectiveness of the data selection approach. Despite the smaller training set, it produced a better or comparable performance to the self-training approach. We suppose that using monolingual Arabic S-BERT for data representation may achieve better results. We will discuss this as future work presented in the next chapter.

To establish the lower limits of performance of the data selection approach, we explored how many source examples are required to learn an accurate classifier that outperforms our baseline trained on all source data. Thus, we conducted an ablation study by systematically reducing the size of selected sub-datasets and evaluating the models. For the relevancy detection task, we experimented with different amounts of source data. We randomly chose 500, 1000 and 2000 samples of the selected data when setting the most similar instances (k) to 3. We performed the experiments for each target data. Then, we fine-tuned the BERT model using the randomly chosen samples. Results are presented in Table 6.3.

We found that the performance degraded when reducing the training data size. It is worth

Table 6.3: The performance for relevancy detection in macro F1 of the data selection method with different training data sizes and the first baseline. (DS refers to the data selection and K is the number of most similar instances. The keyword (all) indicates training on the whole labelled data and n is the number of training examples.)

Target data	BERT-DS (n=500)	BERT-DS (n=1000)	BERT-DS (n=2000)	BERT-DS(K=3) (4000>n>2000)	BERT(all)
JF	65.42%	71.61%	73.66%	73.89%	73.52%
KF	71.34%	77.72%	78.65%	78.34%	72.39%
HF	80.04%	80.08%	84.25%	84.54%	76.39%
CB	53.67%	55.87%	59.99%	56.58%	55.43%
CD	51%	51.15%	52.47%	52.95%	49.88%
DS	72.68%	74.69%	77.27%	80.81%	71.24%
BE	75.82%	77.77%	80.78%	79.04%	78.10%

noting that the BERT trained on 2k examples produced better results than the BERT-DS(k=3) for the CB and BE data. We relate this to the advantage of using balanced training data as we randomly chose equal instances from each class. Training the classifier with 2K instances from the selected data noticeably outperformed the BERT model that uses all training data in all cases. For example, the BERT-DS(n=2000) improved the macro F1 over the BERT(all) model on HF and KF by 7.86% and 6.26%, respectively. Similarly, training on one thousand tweets from the selected data worked better than the baseline in five out of seven cases.

Using 500 training instances, the model produced results below the baseline. For example, the baseline worked better than the BERT(n=500) by 8.1% when classifying the JF. Because of the big difference in performance, we repeated the experiment for this case by randomly choosing another subset and found that the model produced similar results. The performance of the BERT model trained on selected data substantially degraded when using 500 tweets compared to the BERT-DS(k=3) by values ranging from 2% to 8.5%. Figure 6.3 depicts the performance of all data selection models using different k values and the models trained on random samples of selected data versus the baseline. It shows how the performance degraded by reducing the training data until the model produced results lower than the baseline in most cases when training on 500 tweets.

We also reduced the selected data for information classification and evaluated the models. We experimented with 2000 and 1000 samples. In the first experiment, we included all examples

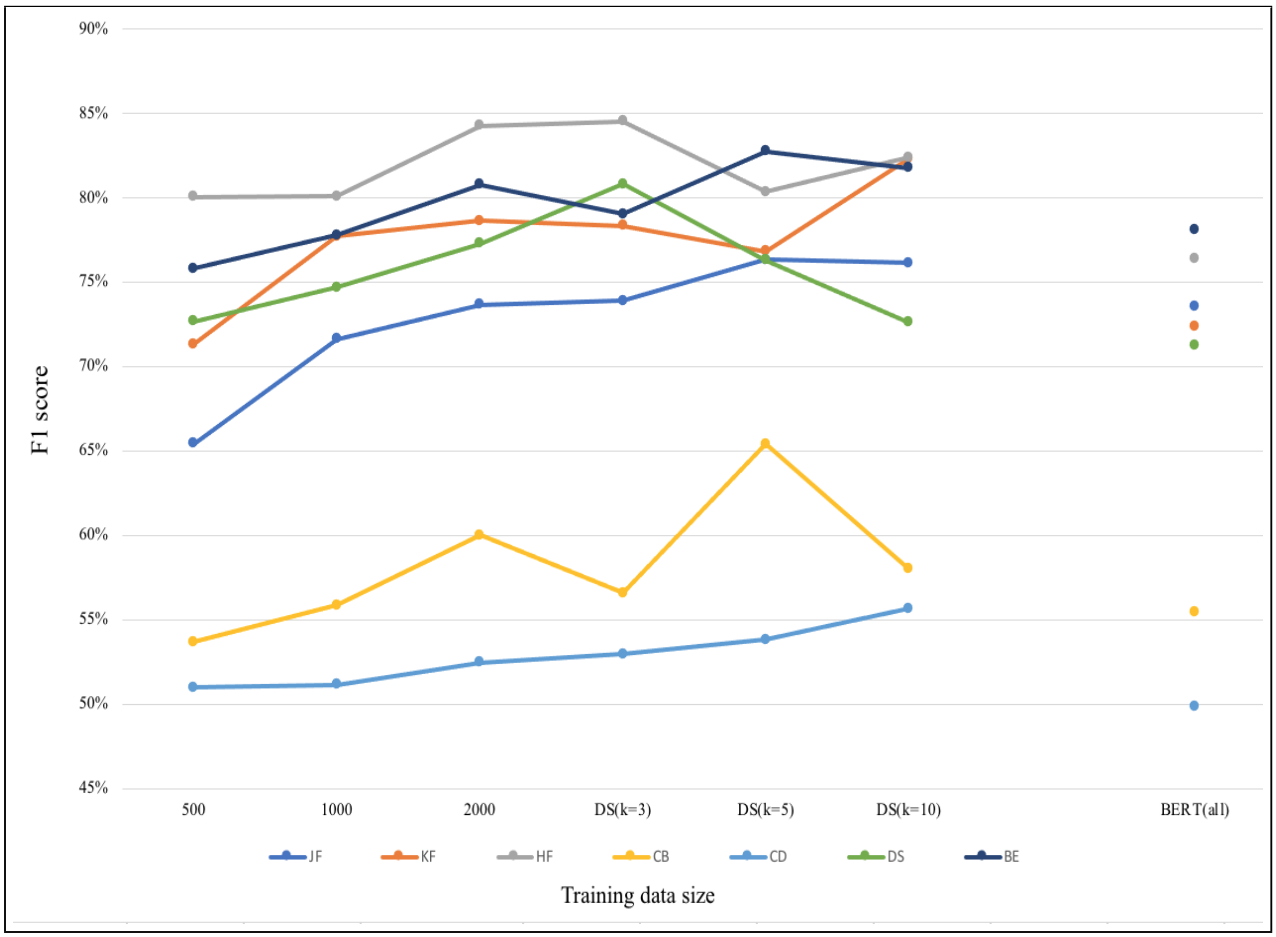

Figure 6.3: The performance for relevancy detection in macro F1 of the data selection method with different training data sizes versus the first baseline. (DS refers to the data selection and K is the number of most similar instances. The keyword (all) indicates training on the whole labelled data and n is the number of training examples.)

Table 6.4: The accuracy scores for information classification of the data selection method with different training data sizes and the first baseline. (DS refers to the data selection and K is the number of most similar instances. The keyword (all) indicates training on the whole labelled data and n is the number of training examples.)

Target data	BERT-DS (n=1000)	BERT-DS (n=2000)	BERT-DS(K=10) (4000>n>2000)	BERT(all)
JF	72.42%	82.72%	87.44%	82%
KF	76.45%	77.54%	81.23%	79.01%
HF	77.16%	80.62%	83.73%	84.27%
CB	66.10%	78%	77.48% (n=1851)	77.05%
DS	77.58%	78.22%	81.88%	7814%
BE	73.77%	78.14%	79.89%	78.14%

in the minority classes (affected individuals & help infrastructure & utility damage) and randomly took equal samples of the reset classes. As the selected data for the CB case is less than 2000, we increased the number of similar instances to include 2000 tweets in the training set. When training on one thousand samples, we down-sampled each class by randomly choosing 200 examples from each category. Results are presented in Table 6.4.

Results show the influence of the size of training data on performance. As for the first task, the accuracy scores degraded as we reduced the training data size. BERT-DS(K=10) enhanced the results over the model trained on two thousand tweets by values ranging from 3.11% to 4.74%. Training on 2k of selected instances instead of 1851 for the selection model resulted in a small performance gain for the CB data. The performance was substantially reduced when limiting the training data size to one thousand instances. For example, using only one thousand tweets generated scores lower than those achieved by the BERT-DS(K=10) model by 15.02% and 11.38% for JF and CB, respectively. The BERT-DS(n=2000) model produced slightly better results than the baseline in three out of five cases. Unlike the first task, which is a binary classification, training on 1k tweets for information classification produced performance lower than the baseline because of the small number of samples per class, as each class has only 200 tweets. Based on the results, we recommend using at least 1000 examples to fine-tune the BERT on the selected data for the relevancy detection task and more than 2000 examples to classify tweets into information types.

6.4 Conclusion

This chapter described our selection-based multi-source domain adaptation approach to identify crisis Twitter messages for new events. Data was selected using the cosine similarity metric in the embedding space generated from transformer-based models. Selecting a subset of data that is semantically similar to the target for fine-tuning BERT models showed promising results. The proposed method outperformed two baselines: training on all data for both tasks and self-training domain adaptation on relevancy detection. We also provided guidance on the training data size used in the data selection approach. We think our instance-level domain adaptation approach is helpful during the early hours of a crisis when no large unlabelled data is available from an emerging disaster. In the next chapter, we will present the conclusions of this research.

Chapter Seven

Conclusions and Future Work

The main goal of this research is to investigate the cross-event crisis detection problem from SM using supervised learning approaches. The thesis demonstrated that DNNs trained on past emergency events could predict crisis messages from current events. Such models can be utilised to build tools to support crisis-affected communities and aid emergency responders. The research focused on crisis detection from Arabic Twitter. It contributed to knowledge by creating, analysing and sharing a multi-event Arabic Twitter corpus for 22 crises. Using the LDA technique, the main topics shared on crisis tweets have been identified and used to label more than 12k messages from seven crises. It also presented an empirical evaluation of several supervised models (such as SVM, NB and DNNs) on two crisis detection tasks: relevancy detection and information classification. Finally, it proposed a domain adaptation method to enhance the models' generalisation for the target crisis.

Chapter 4 introduced our Arabic Twitter corpus for emergency events, Kawarith. It described our iterative collection process and presented a quantitative and qualitative analysis of Kawarith. The chapter showed the NLP techniques, such as LDAs used to identify the main information categories discussed on Twitter during various crises. Finally, it presented the annotation process to create our labelled dataset, which was used for cross-event classification in the following chapters.

Chapter 5 presented an empirical assessment of three supervised classical ML methods and different architectures of DNNs for crisis events classification. The chapter detailed the preprocessing steps, models and training settings. It also evaluated the performance of different pre-

trained embeddings. Results demonstrated the effectiveness of the BERT-based models as they are more robust to the covariate shift across the training (multi-source) and test (target) data, particularly for the multi-label information categorisation task. The last section of the chapter discussed the errors of the best-performing models. Error analysis suggests training BERT on a subset of training data similar to the target and excluding any crisis-related messages from the off-topic class to improve the generalisation.

Chapter 6 extended the study by proposing an unsupervised domain adaptation approach that utilises the BERT models. The chapter described our instance-level data selection method, which aims at finding an optimal set of training data from a multi-source domain that improves the model generalisation for the target data. We showed that the selection-based method achieved promising results. It worked better than a BERT model that learns from all source data, particularly for the relevancy detection task. The experiments and training settings were described in the same chapter.

7.1 Summary of Findings

Below, we briefly highlight the main research findings while revisiting the research questions. The first question aims at identifying the main information types discussed during various crises to gain insights into users' communication during emergencies, check whether the collected data would be helpful to support situational awareness and affected communities and propose a good annotation scheme.

RQ1: *What are the main information categories of conversations posted during different types of crises on Arabic Twitter data?*

We found that most of the information types used for situational awareness and coordinating humanitarian relief efforts appear in our corpus as topics with varying frequency across crises. The information categories include themes related to affected populations, rescue efforts, volunteering, infrastructure damage, interrupted services, warnings, precautions, emotional support, prayers and

opinions. We also found event-unrelated emerged topics representing spam and other events. Based on our analysis, we propose a multi-label annotation scheme of main shared types as presented in Chapter 4.

The second research question asked to evaluate the generalisation performance of different supervised models in multi-source cross-event crisis detection.

RQ2: *Using training data from multiple historical events, how does the performance of DNNs compare to that of conventional ML classifiers in identifying crisis-related posts and categorising relevant posts into different information types?*

We experimented with different features for the conventional ML and DNN models. The selected models were evaluated on two classification tasks. We found that all DNNs outperformed the SVM and DT classifiers while the NB trained using BoWs produced comparable results on the relevancy detection problem. Results of the multi-label information categorisation task showed that all DL models outperformed the classical ML approaches. Initialising word vectors from pre-trained word embeddings resulted in a noticeable performance gain.

We explored which DNN architecture performs better for the two crisis-related tasks and evaluated different text representations. We found that the CNN models worked better in extracting relevant tweets. The best CNN model outperformed the best-performing RNN model by 2.3% in the average macro F1 scores. We observed that feeding the extracted CNN features into an LSTM network enhanced the results over the best-performing CNN model of several cases, showing small gains in performance on average. For the information classification task, results showed that RNNs produce slightly higher scores. Finally, we experimented with the BERT embeddings in two ways: feature-based and fine-tuned-based approaches. Results demonstrated that they remarkably improved the performance for both tasks.

Chapter 6 proposed an instance-based selection domain adaptation approach and discussed the following research questions while evaluating the model.

RQ3: *How do the results of the proposed data selection model compare to the results of the BERT model that learns from all source data?*

The data selection method outperformed the BERT model learnt from all source data on both tasks and trained faster. We found that choosing different values of K for the nearest neighbours resulted in slight differences in performance. For the binary classification task, the best data selection model improved the macro F1 by values ranging from 1.38% to 9.93%. Regarding the second problem, the improvements in macro F1 ranged from 1.17% to 6.73%.

RQ4: *How do the results of the proposed data selection model compare to the results of the self-training approach?*

For the relevancy identification problem, the best-performing selection-based model outperformed the self-training adaptation in three cases, resulting in 2.43%, 4.38% and 11.21% absolute gains. In contrast, the self-training worked better for the COVID-19 event. For the other two cases, they produced relatively similar scores. The approaches have comparable performance when classifying informative tweets into pre-defined categories.

RQ5: *Does combining the proposed selection method with self-training result in performance gain?*

Following the data selection by self-training steps does not consistently improve the performance. We recommend exploiting the self-training when the target event substantially differs in features from the source data (as in the COVID-19 case).

7.2 Future Work

For future work, we seek to investigate several research directions on SM crisis classification. Future studies should consider standardised datasets and evaluation frameworks (e.g. metrics and settings) to conduct fair comparisons between crisis detection models and domain adaptation techniques.

As shown in the last chapter, our proposed data selection approach showed promising performance. Nevertheless, messages related to other concurrent crises still need to be filtered out. We will explore outlier detection methods such as clustering to solve this problem. Besides that, we think the performance of cross-event data selection models can be improved using a monolingual S-BERT model and pre-trained embeddings trained on in-domain crisis data. Instead of using

general-purpose MLMs, we aim to pre-train/fine-tune MLMs on a crisis-related corpus, which includes news articles and tweets, and explore their performance on our downstream tasks. We also aim to investigate how BERT-based models and our proposed domain adaptation approaches generalise to different types of crises, such as social unrest, which are common in the Arab World and differ in feature distribution from floods and explosions. Another area to explore is cross-lingual transfer learning techniques that exploit crisis datasets written in other languages. We want to investigate how models learnt from in-type non-Arabic crisis data compare to the data selection domain adaptation technique that learns from multi-source cross-type datasets.

For the information classification problem, it is helpful to consider fine-grained categories such as fatalities, missing, injured, trapped, etc. We observed misinformation, rumours and false news posted during some events, including COVID-19 and Kuwait floods, that should be filtered out. Several studies addressed the task of misinformation and rumour detection during a specific event/crisis, such as COVID-19 (Hossain et al., 2020; Al-Rakhami and Al-Amri, 2020; Haouari et al., 2021b). Still, further research is required to evaluate the proposed detection models in the cross-crisis setting and to explore whether their extracted features can be generalised to other emergency events.

As mentioned in Chapter 2, little work considered cross-lingual and multi-lingual crisis detection. In this study, we focussed our data collection on Arabic tweets. Messages written in Arabizi or other languages such as English and French were excluded. In future, we would like to extend this work to detect crises in the multi-lingual setting as it represents the actual scenario. It is beneficial to extract messages posted by eyewitnesses as they should be prioritised over those composed by other people or organisations. A recent study on eyewitness post detection demonstrated that using textual features combined with domain-expert features, extracted based on analysing English tweets, achieved the best classification performance (Zahra et al., 2020). As many of such features are language-dependent, it is worth investigating this problem in the multi-lingual setting.

Finally, an impactful future research direction lies in developing and evaluating an end-to-end crisis detection framework that automatically detects the events, identifies crisis posts and

classifies them into information types. This thesis addressed the two classification tasks: relevancy detection and information categorisation. In this work, we collected the dataset by tracking specific crisis-related keywords. The collection process can be automated by performing an open-domain event detection that utilises topic detection and tracking techniques (Fiscus and Doddington, 2002) such as first story detection (Petrović et al., 2010), burst identification and clustering (C. Li et al., 2012) or wavelet-based methods (Weng and B.-S. Lee, 2011; Cordeiro, 2012; Litvak et al., 2016).

Appendix One

Data Annotation Task

Below is a description of the annotation task designed to annotate the Kawarith data. We showed the annotation instructions and examples from the quiz used to train the annotators. The quiz presents the instructions, example tweets from the dataset and explanations of the correct answer.

Task Description

The purpose of this task is to assign a set of short Arabic social media messages (tweets) to one or more of the categories below. These tweets were collected during *[crisis name]* by tracking specific event-related Twitter hashtags and keywords.

Event Description

Crisis Name: Date: Location:

Instructions:

- Please read each tweet and perform the following actions.

  1. Decide whether the tweet is related to the crisis or not.

  2. If the post is relevant, select the most suitable options/classes to describe its content, based on the displayed text only. (All attached media and hyperlinks have been deliberately hidden.)

- If you find it hard to judge the tweet for some reason (e.g. dialect variation that makes its meaning difficult to understand), you may skip the question.

- Each option/category is followed by a description, along with some examples.

Display *{Tweet Text}*

❍ Not related to the crisis

- discusses a different topic, event or crisis, regardless of the inclusion of some relevant hashtags.

- includes incomprehensible messages that require further context to be understood.

❍ Related to the crisis

❑ Affected individuals and help

Includes the following:

1. Reports or questions about fatalities, missing, trapped, injured, survivors or found, evacuated or displaced persons.

2. Mention of damage to personal property.

3. Request or offer of help, donation, volunteering or rescue effort.

❑ Infrastructure and utilities damage

Includes the following:

1. Reports of infrastructure damage (roads, bridges, buildings etc.).

2. Reports of interrupted or restored services.

Exception:

If the post includes criticisms or questions about the built environment without providing any information about damage, it is categorised as opinion and criticism.

❑ Caution, advice, preparations and other crisis updates

Includes the following:

1. Warnings, best practices and early preparations.

2. Crisis updates are not covered by any of the above categories, such as weather forecasts, flood level, and emergency location.

3. Relevant events/news or event consequences.

❑ Emotional support, supplications and prayers

Including condolences and expressions of gratitude.

❑ Opinions and criticism

Including sarcasm and personal views.

Examples from the training quiz

The following tweets were collected during the Kuwait floods by tracking certain event-related Twitter hashtags and keywords.

Instructions:

- Please read each tweet and perform the following actions.
  1. Decide whether the tweet is related to the crisis or not.
  2. If the post is relevant, select the most suitable options/classes to describe its content, based on the displayed text only.

Example 1:

لا يفوتكم الخصم القوي من باث اند بودي، خصم ٢٠ % علي جميع المنتجات #خصم #سيول_الكويت (Don't miss the big discount from Bath and Body, 20% off all products #discount #Kuwait_floods)

The tweet is: not related to the crisis.

Explanation: the tweet is an advertisement and does not provide any information about the crisis. The tweet was collected because of the event-related hashtag (#Kuwait_floods).

Example 2:

وزارة الداخلية تدعو قائدي المركبات لتوخي الحذر لتقلب حالة الطقس #امطار_الكويت (The Ministry of Interior urges vehicle drivers to be cautious due to the changing weather conditions #Kuwait_rains.)

The tweet is: related to the crisis.

The information category is: caution, advice, preparations and other crisis updates.

Explanation: the purpose of the tweet is to advise drivers to take caution during rainy weather. Hence, it is categorised as caution and advice.

References

Abdelali, Ahmed, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak (June 2016). "Farasa: A Fast and Furious Segmenter for Arabic". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. San Diego, California, pp. 11–16. DOI: 10.18653/v1/N16-3003. URL: https://aclanthology.org/N16-3003.

Addawood, Aseel (2020). *Coronavirus: Public Arabic Twitter Data Set*. URL: https://openreview.net/forum?id=ZxjFAfD0pSy.

Adel, Ghadah and Yuping Wang (2020a). "Arabic Twitter Corpus for Crisis Response Messages Classification". In: *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*. ACAI '19. Sanya, China: Association for Computing Machinery, pp. 498–503. ISBN: 9781450372619. DOI: 10.1145/3377713.3377799. URL: https://doi.org/10.1145/3377713.3377799.

— (2020b). "Detecting and Classifying Humanitarian Crisis in Arabic Tweets". In: *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE, pp. 269–274. DOI: 10.1109/ICAIBD49809.2020.9137480.

Aharoni, Roee and Yoav Goldberg (July 2020). "Unsupervised Domain Clusters in Pretrained Language Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online, pp. 7747–7763. DOI: 10.18653/v1/2020.acl-main.692. URL: https://aclanthology.org/2020.acl-main.692.

Ahmed, Wasim, Peter A Bath, Laura Sbaffi, and Gianluca Demartini (2019). "Novel insights into views towards H1N1 during the 2009 Pandemic: a thematic analysis of Twitter data". In: *Health Information & Libraries Journal* 36.1, pp. 60–72.

Aipe, Alan, Asif Ekbal, Mukuntha NS, and Sadao Kurohashi (2018). "Linguistic Feature Assisted Deep Learning Approach towards Multi-label Classification of Crisis Related Tweets". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. ISCRAM Association. ISBN: 978-0-692-12760-5. URL: http://idl.iscram.org/files/alanaipe/2018/1592_AlanAipe_etal2018.pdf.

Alabbas, Waleed, Haider M al-Khateeb, Ali Mansour, Gregory Epiphaniou, and Ingo Frommholz (2017). "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, pp. 1–8. DOI: 10.1109/SOCIALMEDIA.2017.8057358.

Alam, Firoj, Shafiq Joty, and Muhammad Imran (July 2018a). "Domain Adaptation with Adversarial Training and Graph Embeddings". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia, pp. 1077–1087. DOI: 10.18653/v1/P18-1099. URL: https://aclanthology.org/P18-1099.

Alam, Firoj, Ferda Ofli, and Muhammad Imran (2018b). "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters". In: *Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018*. AAAI Press, pp. 465–473. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17816.

— (2019). "CrisisDPS: Crisis Data Processing Services". In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain, May 19-22, 2019*. Ed. by Zeno Franco, José J. González, and José H.

Canós. ISCRAM Association. URL: http://idl.iscram.org/files/firojalam/2019/1891%5C_FirojAlam%5C_etal2019.pdf.

Alam, Firoj, Ferda Ofli, and Muhammad Imran (2020). "Descriptive and visual summaries of disaster events using artificial intelligence techniques: case studies of Hurricanes Harvey, Irma, and Maria". In: *Behav. Inf. Technol.* 39.3, pp. 288–318. DOI: 10.1080/0144929X.2019.1610908. URL: https://doi.org/10.1080/0144929X.2019.1610908.

Alam, Firoj, Ferda Ofli, Muhammad Imran, and Michaël Aupetit (2018c). "A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria". In: *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management, Rochester, NY, USA, May 20-23, 2018*. Ed. by Kees Boersma and Brian M. Tomaszewski. ISCRAM Association. URL: http://idl.iscram.org/files/firojalam/2018/1579%5C_FirojAlam%5C_etal2018.pdf.

Alam, Firoj, Umair Qazi, Muhammad Imran, and Ferda Ofli (2021a). "HumAID: Human-Annotated Disaster Incidents Data from Twitter with Deep Learning Benchmarks". In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM. AAAI Press*, pp. 933–942.

Alam, Firoj, Hassan Sajjad, Muhammad Imran, and Ferda Ofli (2021b). "CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing". In: *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*. Ed. by Ceren Budak, Meeyoung Cha, Daniele Quercia, and Lexing Xie. AAAI Press, pp. 923–932. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/18115.

Albalawi, Yahya, Jim Buckley, and Nikola S Nikolov (2021). "Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting Arabic health information on social media". In: *Journal of big Data* 8.1, p. 95.

Aldrsoni, Sulaiman (2012). معجم اللهجات المحكية *[Dictionary of Spoken Dialects]*. Riyadh, KSA: King Fahd National Library.

Alharbi, Abdullah I. and Mark Lee (2020). "Combining Character and Word Embeddings for Affect in Arabic Informal Social Media Microblogs". In: *Natural Language Processing and Information Systems*. Ed. by Elisabeth Métais, Farid Meziane, Helmut Horacek, and Philipp Cimiano. Cham: Springer International Publishing, pp. 213–224. ISBN: 978-3-030-51310-8.

Alharbi, Alaa and Mark Lee (July 2019). "Crisis Detection from Arabic Tweets". In: *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*. Cardiff, United Kingdom: Corpus Linguistics Conference, pp. 72–79. URL: https://aclanthology.org/W19-5609.

— (Apr. 2021). "Kawarith: an Arabic Twitter Corpus for Crisis Events". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): European Association for Computational Linguistics (EACL), pp. 42–52. URL: https://aclanthology.org/2021.wanlp-1.5.

— (June 2022). "Classifying Arabic Crisis Tweets using Data Selection and Pre-trained Language Models". In: *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*. Marseille, France: Language Resources and Evaluation Conference (LREC), pp. 71–78. URL: https://aclanthology.org/2022.osact-1.8.

Ali, Haseeb, MN Mohd Salleh, Rohmat Saedudin, Kashif Hussain, and Muhammad Faheem Mushtaq (2019). "Imbalance class problems in data mining: a review". In: *Indonesian Journal of Electrical Engineering and Computer Science* 14.3, pp. 1560–1571.

Alqurashi, Sarah, Ahmad Alhindi, and Eisa Alanazi (2020). "Large Arabic Twitter Dataset on COVID-19". In: *CoRR* abs/2004.04315. arXiv: 2004.04315. URL: https://arxiv.org/abs/2004.04315.

ALRashdi, Reem and Simon O'Keefe (2018). "Deep Learning and Word Embeddings for Tweet Classification for Crisis Response". In: *The 3rd National Computing Colleges Conference*. York.

ALRashdi, Reem and Simon O'Keefe (2019). "Robust Domain Adaptation Approach for Tweet Classification for Crisis Response". In: *International Conference Europe Middle East & North Africa Information Systems and Technologies to Support Learning*. Springer, pp. 124–134.

Alsarsour, Israa, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed (May 2018). "DART: A Large Dataset of Dialectal Arabic Tweets". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan. URL: https://aclanthology.org/L18-1579.

Alshaabi, Thayer, Michael V Arnold, Joshua R Minot, Jane Lydia Adams, David Rushing Dewhurst, Andrew J Reagan, Roby Muhamad, Christopher M Danforth, and Peter Sheridan Dodds (2021). "How the world's collective attention is being paid to a pandemic: COVID-19 related n-gram time series for 24 languages on Twitter". In: *Plos one* 16.1, e0244476.

Alsudias, Lama and Paul Rayson (July 2019). "Classifying Information Sources in Arabic Twitter to Support Online Monitoring of Infectious Diseases". In: *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*. Cardiff, United Kingdom: Corpus Linguistics Conference, pp. 22–30. URL: https://aclanthology.org/W19-5604.

— (July 2020a). "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?" In: *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online. URL: https://aclanthology.org/2020.nlpcovid19-acl.16.

— (May 2020b). "Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France, pp. 4842–4850. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.596.

Alsudias, Lama and Paul Rayson (2021). "Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study". In: *JMIR medical informatics* 9.9, e27670.

Althobaiti, Maha J (2020). "Automatic Arabic dialect identification systems for written texts: A survey". In: *CoRR* abs/2009.12622. arXiv: 2009.12622. URL: https://arxiv.org/abs/2009.12622.

Antoun, Wissam, Fady Baly, and Hazem M. Hajj (2020). "AraBERT: Transformer-based Model for Arabic Language Understanding". In: *CoRR* abs/2003.00104. arXiv: 2003.00104. URL: https://arxiv.org/abs/2003.00104.

Artstein, Ron and Massimo Poesio (2008). "Survey Article: Inter-Coder Agreement for Computational Linguistics". In: *Computational Linguistics* 34.4, pp. 555–596. DOI: 10.1162/coli.07-034-R2. URL: https://aclanthology.org/J08-4004.

Ashktorab, Zahra, Christopher Brown, Manojit Nandi, and Aron Culotta (2014). "Tweedr: Mining twitter to inform disaster response". In: *11th Proceedings of the International Conference on Information Systems for Crisis Response and Management, University Park, Pennsylvania, USA, May 18-21, 2014*. Ed. by Starr Roxanne Hiltz, Linda Plotnick, Mark Pfaf, and Patrick C. Shih. ISCRAM Association. URL: http://idl.iscram.org/files/ashktorab/2014/275%5C_Ashktorab%5C_etal2014.pdf.

Baali, Massa and Nada Ghneim (2019). "Emotion analysis of Arabic tweets using deep learning approach". In: *Journal of Big Data* 6.1, pp. 1–12.

Banda, Juan M., Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, Ekaterina Artemova, Elena Tutubalina, and Gerardo Chowell (2021). "A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research–An International Collaboration". In: *Epidemiologia* 2.3, pp. 315–324. ISSN: 2673-3986. DOI: 10.3390/epidemiologia2030024. URL: http://dx.doi.org/10.3390/epidemiologia2030024.

Benevenuto, Fabricio, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida (July 2010). "Detecting Spammers on Twitter". In: *Proceedings of the Seventh Annual Collabora-*

tion, Electronic messaging, Anti-Abuse and Spam Conference (CEAS). Washington, DC, USA.

Bengio, Yoshua, Olivier Delalleau, and Clarence Simard (2010). "Decision trees do not generalize to new variations". In: *Computational Intelligence* 26.4, pp. 449–467.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning Long-Term Dependencies with Gradient Descent is Difficult". In: *IEEE Transactions on Neural Networks* 5.2, pp. 157–166.

Bird, Steven (July 2006). "NLTK: The Natural Language Toolkit". In: *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*. Sydney, Australia, pp. 69–72. DOI: 10.3115/1225403.1225421. URL: https://aclanthology.org/P06-4018.

Blei, David M. (Apr. 2012). "Probabilistic Topic Models". In: *Commun. ACM* 55.4, pp. 77–84. ISSN: 0001-0782. DOI: 10.1145/2133806.2133826. URL: http://doi.acm.org/10.1145/2133806.2133826.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (Mar. 2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. DOI: 10.1162/tacl_a_00051. URL: https://aclanthology.org/Q17-1010.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik (1992). "A Training Algorithm for Optimal Margin Classifiers". In: *Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT '92), July 27-29, 1992, Pittsburgh, PA, USA*. ACM Press, New York, NY, USA, pp. 144–152. URL: http://doi.acm.org/10.1145/130385.130401.

Botha, Jan A. and Phil Blunsom (2014). "Compositional Morphology for Word Representations and Language Modelling". In: *ICML*. Vol. 32. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 1899–1907.

Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone (1984). *Classification and Regression Trees*. Wadsworth. ISBN: 0-534-98053-8.

Bukar, Umar Ali, Marzanah A Jabar, Fatimah Sidi, Rozi Nor Haizan Binti Nor, Salfarina Abdullah, and Mohamed Othman (2020). "Crisis Informatics in the Context of Social Media Crisis Communication: Theoretical Models, Taxonomy, and Open Issues". In: *IEEE Access* 8, pp. 185842–185869.

Burel, Grégoire and Harith Alani (2018). "Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media". In: *ISCRAM*. ISCRAM Association. ISBN: 978-0-692-12760-5.

Burel, Grégoire, Hassan Saif, and Harith Alani (2017a). "Semantic Wide and Deep Learning for Detecting Crisis-Information Categories on Social Media". In: *International Semantic Web Conference (1)*. Vol. 10587. Lecture Notes in Computer Science. Springer, pp. 138–155. ISBN: 978-3-319-68288-4.

Burel, Grégoire, Hassan Saif, Miriam Fernandez, and Harith Alani (May 2017b). "On Semantics and Deep Learning for Event Detection in Crisis Situations". In: *Workshop on Semantic Deep Learning (SemDeep)*. Portoroz, Slovenia.

Caragea, Cornelia, Adrian Silvescu, and Andrea H Tapia (2016). "Identifying Informative Messages in Disaster Events using Convolutional Neural Networks". In: *ISCRAM 2016 Conference Proceedings - 13th International Conference on Information Systems for Crisis Response and Management*. Information Systems for Crisis Response and Management, ISCRAM, pp. 137–147.

Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete (2011). "Information credibility on twitter". In: *Proceedings of the 20th international conference on World wide web*.

WWW '11. ACM. New York, NY, USA, pp. 675–684. URL: http://doi.acm.org/10.1145/1963405.1963500.

Chauhan, Nitin Kumar and Krishna Singh (2018). "A Review on Conventional Machine Learning vs Deep Learning". In: *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE. Greater Noida, India, pp. 347–352. DOI: 10.1109/GUCON.2018.8675097.

Chauhan, Prashant and Madhu Shukla (2015). "A review on outlier detection techniques on data stream by using different approaches of K-Means algorithm". In: *2015 International Conference on Advances in Computer Engineering and Applications*. IEEE. Ghaziabad, India, pp. 580–585. DOI: 10.1109/ICACEA.2015.7164758.

Chen, Emily, Kristina Lerman, and Emilio Ferrara (2020). "COVID-19: The First Public Coronavirus Twitter Dataset". In: *CoRR* abs/2003.07372. arXiv: 2003.07372. URL: https://arxiv.org/abs/2003.07372.

Chen, Qi, Wei Wang, Kaizhu Huang, Suparna De, and Frans Coenen (2020). "Adversarial Domain Adaptation for Crisis Data Classification on Social Media". In: *2020 International Conferences on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*. IEEE, pp. 282–287. DOI: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics50389.2020.00061.

Chiang, David, Mona Diab, Nizar Habash, Owen Rambow, and Safiullah Shareef (Apr. 2006). "Parsing Arabic Dialects". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy, pp. 369–376. URL: https://aclanthology.org/E06-1047.

Cho, Kyunghyun, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio (Oct. 2014). "On the Properties of Neural Machine Translation: Encoder–Decoder Approaches". In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and*

Structure in Statistical Translation. Doha, Qatar, pp. 103–111. DOI: 10.3115/v1/W14-4012. URL: https://aclanthology.org/W14-4012.

Chollet, François (2017). *Deep Learning with Python*. Manning. ISBN: 9781617294433.

Cobo, Alfredo, Denis Parra, and Jaime Navón (2015). "Identifying Relevant Messages in a Twitter-based Citizen Channel for Natural Disaster Situations". In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 1189–1194.

Cordeiro, Mário (2012). "Twitter event detection: Combining wavelet analysis and topic inference summarization". In: *Doctoral Symposium on Informatics Engineering, DSIE*, pp. 11–16.

Cresci, Stefano, Maurizio Tesconi, Andrea Cimino, and Felice Dell'Orletta (2015). "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*. ACM, pp. 1195–1200.

Cristianini, Nello and Bernhard Scholkopf (2002). "Support Vector Machines and Kernel Methods: The New Generation of Learning Machines". In: *AI Magazine* 23.3, pp. 31–31. DOI: 10.1609/aimag.v23i3.1655. URL: https://ojs.aaai.org/index.php/aimagazine/article/view/1655.

Cunningham, Pádraig, Matthieu Cord, and Sarah Jane Delany (2008). "Supervised Learning". In: *Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 21–49. ISBN: 978-3-540-75171-7. DOI: 10.1007/978-3-540-75171-7_2. URL: https://doi.org/10.1007/978-3-540-75171-7_2.

Derczynski, Leon, Kenny Meesters, Kalina Bontcheva, and Diana Maynard (2018). "Helping Crisis Responders Find the Informative Needle in the Tweet Haystack". In: *ISCRAM*. URL: http://idl.iscram.org/files/leonderczynski/2018/1587_LeonDerczynski_etal2018.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Diriöz, Ali Oğuz (2013). "Twitter & The Middle East". In: *Center for Middle Eastern Strategic Studies* 5.1, pp. 67–75.

Eberhard, David M, Gary Simons, and Charles Fennig (2019). *Ethnologue: Languages of the World twenty-second edition*. Dallas, Texas: SIL International. URL: https://www.ethnologue.com/.

Eltanbouly, Sohaila, May Bashendy, and Tamer Elsayed (2019). "Simple But Not Naive: Fine-Grained Arabic Dialect Identification Using Only N-Grams". In: *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pp. 214–218.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin (2008). "LIBLINEAR: A library for large linear classification". In: *the Journal of machine Learning research* 9, pp. 1871–1874.

Farha, Ibrahim Abu and Walid Magdy (2022). "The Effect of Arabic Dialect Familiarity on Data Annotation". In: *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*. Abu Dhabi, United Arab Emirates (Hybrid), pp. 399–408. DOI: 10.18653/v1/2022.wanlp-1.39. URL: https://aclanthology.org/2022.wanlp-1.39.

Fiscus, Jonathan G. and George R. Doddington (2002). "Topic Detection and Tracking Evaluation Overview". In: *Topic Detection and Tracking: Event-based Information Organization*. Ed. by James Allan. Boston, MA: Springer US, pp. 17–31. ISBN: 978-1-4615-0933-2. DOI: 10.1007/978-1-4615-0933-2_2. URL: https://doi.org/10.1007/978-1-4615-0933-2_2.

Gal, Yarin (2016). "Uncertainty in Deep Learning". PhD thesis. University of Cambridge.

Gers, Felix A, Jürgen Schmidhuber, and Fred Cummins (2000). "Learning to Forget: Continual Prediction with LSTM". In: *Neural computation* 12.10, pp. 2451–2471.

Godbole, Shantanu and Sunita Sarawagi (2004). "Discriminative Methods for Multi-labeled Classification". In: *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 22–30. ISBN: 978-3-540-24775-3.

Graf, David, Werner Retschitzegger, Wieland Schwinger, Birgit Pröll, and Elisabeth Kapsammer (2020). "Exploiting Twitter for Informativeness Classification in Disaster Situations". In: *Transactions on Large-Scale Data-and Knowledge-Centered Systems XLV*. Springer, pp. 27–55.

Guo, Han, Ramakanth Pasunuru, and Mohit Bansal (2020). "Multi-Source Domain Adaptation for Text Classification via DistanceNet-Bandits". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05, pp. 7830–7838. DOI: 10.1609/aaai.v34i05.6288. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6288.

Gupta, Raj Kumar, Ajay Vishwanath, and Yinping Yang (2020). "COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes". In: *CoRR* abs/2007.06954. arXiv: 2007.06954. URL: https://arxiv.org/abs/2007.06954.

Habash, Nizar, Abdelhadi Soudi, and Timothy Buckwalter (2007). "On Arabic Transliteration". In: *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Dordrecht: Springer Netherlands, pp. 15–22. ISBN: 978-1-4020-6046-5. DOI: 10.1007/978-1-4020-6046-5_2. URL: https://doi.org/10.1007/978-1-4020-6046-5_2.

Hamoui, Btool, Mourad Mars, and Khaled Almotairi (May 2020). "FloDusTA: Saudi Tweets Dataset for Flood, Dust Storm, and Traffic Accident Events". English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France, pp. 1391–1396. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.174.

Haouari, Fatima, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed (Apr. 2021a). "ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): European Association for Computational Linguistics (EACL), pp. 82–91. URL: https://aclanthology.org/2021.wanlp-1.9.

— (Apr. 2021b). "ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection". In: *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. Kyiv, Ukraine (Virtual): European Association for Computational Linguistics (EACL), pp. 72–81. URL: https://aclanthology.org/2021.wanlp-1.8.

Harris, Zellig S. (1954). "Distributional Structure". In: *WORD* 10.2-3, pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

Al-Hassan, Areej and Hmood Al-Dossari (2022). "Detection of hate speech in Arabic tweets using deep learning". In: *Multimedia systems* 28.6, pp. 1963–1974.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural computation* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

Hossain, Tamanna, Robert L. Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh (Dec. 2020). "COVIDLies: Detecting COVID-19 Misinformation on Social Media". In: *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Online. DOI: 10.18653/v1/2020.nlpcovid19-2.11. URL: https://aclanthology.org/2020.nlpcovid19-2.11.

Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin (2003). *A Practical Guide to Support Vector Classification*. Tech. rep. Department of Computer Science, National Taiwan University.

Imran, Muhammad, Carlos Castillo, Fernando Diaz, and Sarah Vieweg (2018). "Processing Social Media Messages in Mass Emergency: Survey Summary". In: *Companion of the*

The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018. ACM, pp. 507–511. DOI: 10.1145/3184558.3186242.

Imran, Muhammad, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg (2014). "AIDR: Artificial Intelligence for Disaster Response". In: WWW '14 Companion. Seoul, Korea: Association for Computing Machinery, pp. 159–162. ISBN: 9781450327459. DOI: 10.1145/2567948.2577034. URL: https://doi.org/10.1145/2567948.2577034.

Imran, Muhammad, Shady Elbassuoni, Carlos Castillo, Fernando Diaz, and Patrick Meier (2013a). "Extracting Information Nuggets from Disaster- Related Messages in Social Media". In: *10th Proceedings of the International Conference on Information Systems for Crisis Response and Management, Baden-Baden, Germany, May 12-15, 2013*. ISCRAM Association.

— (2013b). "Practical Extraction of Disaster-Relevant Information from Social Media". In: *Proceedings of the 22nd International Conference on World Wide Web*. WWW '13 Companion. Rio de Janeiro, Brazil: Association for Computing Machinery, pp. 1021–1024. ISBN: 9781450320382. DOI: 10.1145/2487788.2488109. URL: https://doi.org/10.1145/2487788.2488109.

Imran, Muhammad, Prasenjit Mitra, and Carlos Castillo (May 2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia, pp. 1638–1643. URL: https://aclanthology.org/L16-1259.

Janiesch, Christian, Patrick Zschech, and Kai Heinrich (2021). "Machine learning and deep learning". In: *Electronic Markets* 31.3, pp. 685–695. DOI: 10.1007/s12525-021-00475-2. URL: https://doi.org/10.1007/s12525-021-00475-2.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng (2007). "Why We Twitter: Understanding Microblogging Usage and Communities". In: WebKDD/SNA-KDD '07.

ACM. San Jose, California: Association for Computing Machinery, pp. 56–65. ISBN: 9781595938480. DOI: 10.1145/1348549.1348556. URL: https://doi.org/10.1145/1348549.1348556.

Kabakus, Abdullah Talha and Resul Kara (2017). "A Survey of Spam Detection Methods on Twitter". In: *International Journal of Advanced Computer Science and Applications* 8.3, pp. 29–38.

Kanaris, Ioannis, Konstantinos Kanaris, Ioannis Houvardas, and Efstathios Stamatatos (2007). "Words versus Character n-Grams for Anti-Spam Filtering". In: *International Journal on Artificial Intelligence Tools* 16.06, pp. 1047–1067.

Karami, Amir, Morgan Lundy, Frank Webb, and Yogesh K. Dwivedi (2020a). "Twitter and Research: A Systematic Literature Review Through Text Mining". In: *IEEE Access* 8, pp. 67698–67717. DOI: 10.1109/ACCESS.2020.2983656.

Karami, Amir, Vishal Shah, Reza Vaezi, and Amit Bansal (2020b). "Twitter speaks: A case of national disaster situational awareness". In: *Journal of Information Science* 46.3, pp. 313–324. DOI: 10.1177/0165551519828620. URL: https://doi.org/10.1177/0165551519828620.

Kersten, Jens, Anna Kruspe, Matti Wiegmann, and Friederike Klan (2019). "Robust Filtering of Crisis-related Tweets". In: *Proceedings of the 16th International Conference on Information Systems for Crisis Response and Management, València, Spain*. ISCRAM Association.

Khare, Prashant, Grégoire Burel, Diana Maynard, and Harith Alani (2018). "Cross-Lingual Classification of Crisis Data". In: *The Semantic Web – ISWC 2018*. Cham: Springer International Publishing, pp. 617–633. ISBN: 978-3-030-00671-6.

Khondker, Habibul Haque (2011). "Role of the New Media in the Arab Spring". In: *Globalizations* 8.5, pp. 675–679. DOI: 10.1080/14747731.2011.621287. URL: https://doi.org/10.1080/14747731.2011.621287.

Kim, Jooho, Juhee Bae, and Makarand Hastak (2018). "Emergency information diffusion on online social media during storm Cindy in U.S." In: *International Journal of Information Management* 40, pp. 153–165. ISSN: 0268-4012. DOI: https://doi.org/10.1016/j.ijinfomgt.2018.02.003. URL: https://www.sciencedirect.com/science/article/pii/S0268401218300562.

Kim, Yoon (Oct. 2014). "Convolutional Neural Networks for Sentence Classification". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 1746–1751. DOI: 10.3115/v1/D14-1181. URL: https://aclanthology.org/D14-1181.

Kireyev, Kirill, Leysia Palen, and Kenneth Anderson (2009). "Applications of Topics Models to Analysis of Disaster-Related Twitter Data". In: *NIPS workshop on applications for topic models: text and beyond*. Vol. 1.

Kozlowski, Diego, Elisa Lannelongue, Frédéric Saudemont, Farah Benamara, Alda Mari, Véronique Moriceau, and Abdelmoumene Boumadane (2020). "A three-level classification of French tweets in ecological crises". In: *Information Processing & Management* 57.5, p. 102284. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2020.102284. URL: https://www.sciencedirect.com/science/article/pii/S0306457320300650.

Krishnan, Jitin, Hemant Purohit, and Huzefa Rangwala (2020). "Unsupervised and Interpretable Domain Adaptation to Rapidly Filter Tweets for Emergency Services". In: *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, pp. 409–416.

Lamsal, Rabindra (2020). *Coronavirus (COVID-19) Tweets Dataset*. DOI: 10.21227/781w-ef42. URL: https://dx.doi.org/10.21227/781w-ef42.

Landwehr, Peter M and Kathleen M Carley (2014). "Social Media in Disaster Relief". In: pp. 225–257. DOI: 10.1007/978-3-642-40837-3_7. URL: https://doi.org/10.1007/978-3-642-40837-3_7.

Lazaridou, Angeliki, Marco Marelli, Roberto Zamparelli, and Marco Baroni (Aug. 2013). "Compositional-ly Derived Representations of Morphologically Complex Words in Distributional Semantics". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria, pp. 1517–1526. URL: https://aclanthology.org/P13-1149.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324. DOI: 10.1109/5.726791.

Li, Chenliang, Aixin Sun, and Anwitaman Datta (2012). "Twevent: Segment-Based Event Detection from Tweets". In: *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. CIKM '12. Maui, Hawaii, USA: Association for Computing Machinery, pp. 155–164. ISBN: 9781450311564. DOI: 10.1145/2396761.2396785. URL: https://doi.org/10.1145/2396761.2396785.

Li, Hongmin, Doina Caragea, and Cornelia Caragea (2017). "Towards Practical Usage of a Domain Adaptation Algorithm in the Early Hours of a Disaster". In: *Proceedings of the 14th International Conference on Information Systems for Crisis Response And Management*. Vol. 2017-May. Albi, France, pp. 692–704.

— (2021). "Combining Self-training with Deep Learning for Disaster Tweet Classification". In: *Proceedings of the 18th International Conference on Information Systems for Crisis Response and Management*. VA, USA, pp. 719–730.

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

Li, Hongmin, Nicolais Guevara, Nic Herndon, Doina Caragea, Kishore Neppalli, Cornelia Caragea, Anna Cinzia Squicciarini, and Andrea H Tapia (2015). "Twitter Mining for Disaster Response: A Domain Adaptation Approach". In: *ISCRAM 2015 Conference*

Proceedings - 12th International Conference on Information Systems for Crisis Response and Management. Norway.

Li, Hongmin, Xukun Li, Doina Caragea, and Cornelia Caragea (2018b). "Comparison of Word Embeddings and Sentence Encodings as Generalized Representations for Crisis Tweet Classification Tasks". In: *Proceedings of the ISCRAM Asian Pacific 2018 Conference*. Wellington, New Zealand.

Li, Hongmin, Oleksandra Sopova, Doina Caragea, and Cornelia Caragea (2018c). "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, pp. 1–20.

Li, Xukun and Doina Caragea (2020). "Domain Adaptation with Reconstruction for Disaster Tweet Classification". In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, pp. 1561–1564. ISBN: 9781450380164.

Littman, Justin (2017). *Hurricanes Harvey and Irma Tweet ids*. Version V1. DOI: 10.7910/DVN/QRKIBW. URL: https://doi.org/10.7910/DVN/QRKIBW.

Litvak, Marina, Natalia Vanetik, Efi Levi, and Michael Roistacher (2016). "What's up on Twitter? Catch up with TWIST!" In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 213–217.

Liu, Junhua, Trisha Singhal, Luciënne T. M. Blessing, Kristin L. Wood, and Kwan Hui Lim (2020). "EPIC30M: An Epidemics Corpus Of Over 30 Million Relevant Tweets". In: *CoRR* abs/2006.08369. arXiv: 2006.08369. URL: https://arxiv.org/abs/2006.08369.

Liu, Junhua, Trisha Singhal, Lucienne TM Blessing, Kristin L Wood, and Kwan Hui Lim (2021). "CrisisBERT: A Robust Transformer for Crisis Classification and Contextual Crisis Embedding". In: *Proceedings of the 32nd ACM Conference on Hypertext and*

Social Media, Virtual Event, Ireland. ACM, pp. 133–141. DOI: 10.1145/3465336.3475117. URL: https://doi.org/10.1145/3465336.3475117.

Lorini, Valerio, Carlos Castillo, Francesco Dottori, Milan Kalas, Domenico Nappo, and Peter Salamon (2019). "Integrating Social Media into a Pan-European Flood Awareness System: A Multilingual Approach". In: *CoRR* abs/1904.10876. arXiv: 1904.10876. URL: http://arxiv.org/abs/1904.10876.

Ma, Guoqin (2019). "Tweets Classification with BERT in the Field of Disaster Management". In: *StudentReport@ Stanford. edu*.

Ma, Xiaofei, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang (2019). "Domain adaptation with BERT-based domain classification and data selection". In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 76–83.

Madichetty, Sreenivasulu and Sridevi Muthukumarasamy (2020). "Detection of situational information from Twitter during disaster using deep learning models". In: *Sādhanā* 45.1, pp. 1–13. DOI: https://doi.org/10.1007/s12046-020-01504-0.

Madichetty, Sreenivasulu and M Sridevi (2019a). "Detecting Informative Tweets during Disaster using Deep Neural Networks". In: *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*. IEEE, pp. 709–713. DOI: 10.1109/COMSNETS.2019.8711095.

— (2019b). "Disaster damage assessment from the tweets using the combination of statistical features and informative words". In: *Social Network Analysis and Mining* 9.1, pp. 1–11. DOI: 10.1007/s13278-019-0579-5.

— (2020). "Improved Classification of Crisis-Related Data on Twitter using Contextual Representations". In: *Procedia Computer Science* 167, pp. 962–968. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.03.395. URL: https://www.sciencedirect.com/science/article/pii/S1877050920308619.

El-Mawass, Nour and Saad Alaboodi (2016). "Detecting Arabic spammers and content polluters on Twitter". In: *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*. IEEE, pp. 53–58.

Mazloom, Reza, HingMin Li, Doina Caragea, Cornelia Caragea, and Muhammad Imran (2018). "Classification of Twitter Disaster Data Using a Hybrid Feature-Instance Adaptation Approach". In: *Proceedings of the 15th Annual Conference for Information Systems for Crisis Response and Management (ISCRAM)*.

Mazloom, Reza, Hongmin Li, Doina Caragea, Cornelia Caragea, and Muhammad Imran (2019). "A hybrid domain adaptation approach for identifying crisis-relevant tweets". In: *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)* 11.2, pp. 1–19.

McCallum, Andrew, Kamal Nigam, et al. (1998). "A comparison of event models for naive bayes text classification". In: *AAAI-98 workshop on learning for text categorization*. Vol. 752. 1. Madison, WI, pp. 41–48.

McCreadie, Richard, Cody Buntain, and Ian Soboroff (2020). "Incident Streams 2019: Actionable Insights and How to Find Them". In: *Proceedings of the 17th ISCRAM Conference*. Blacksburg, VA, USA.

Mihunov, Volodymyr V, Nina SN Lam, Lei Zou, Zheye Wang, and Kejin Wang (2020). "Use of Twitter in disaster rescue: lessons learned from Hurricane Harvey". In: *International Journal of Digital Earth* 13.12, pp. 1454–1466. DOI: 10.1080/17538947.2020.1729879.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781. URL: https://arxiv.org/abs/1301.3781.

Mubarak, Hamdy (May 2018). "Build Fast and Accurate Lemmatization for Arabic". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan. URL: https://aclanthology.org/L18-1181.

Mubarak, Hamdy, Ahmed Abdelali, Sabit Hassan, and Kareem Darwish (2020). "Spam detection on arabic twitter". In: *Social Informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings 12*. Springer, pp. 237–251. URL: https://link.springer.com/chapter/10.1007/978-3-030-60975-7_18.

Nalluru, Ganesh, Rahul Pandey, and Hemant Purohit (2019). "Classifying Relevant Social Media Posts During Disasters Using Ensemble of Domain-agnostic and Domain-specific Word Embeddings". In: *AAAI FSS-19: Artificial Intelligence for Social Good*. URL: https://par.nsf.gov/biblio/10176484.

Nassif, Ali Bou, Abdollah Masoud Darya, and Ashraf Elnagar (2021). "Empirical evaluation of shallow and deep learning classifiers for Arabic sentiment analysis". In: *Transactions on Asian and Low-Resource Language Information Processing* 21.1, pp. 1–25.

Neppalli, Venkata Kishore, Cornelia Caragea, and Doina Caragea (2018). "Deep Neural Networks versus Naïve Bayes Classifiers for Identifying Informative Tweets during Disasters". In: *Proceedings of the 15th ISCRAM Conference*. Rochester, NY, USA.

Nguyen, Dat, Kamela Ali Al Mannai, Shafiq Joty, Hassan Sajjad, Muhammad Imran, and Prasenjit Mitra (2017). "Robust Classification of Crisis-Related Data on Social Networks Using Convolutional Neural Networks". In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1, pp. 632–635. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14950.

Nguyen, Dat Tien, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra (2016). "Applications of Online Deep Learning for Crisis Response Using Social Media Information". In: *The Fourth International Workshop on Social Web for Disaster Management (SWDM 2016)*. USA: Association for Computing Machinery. ISBN: 9781450340731.

Ning, Xiaodong, Lina Yao, Boualem Benatallah, Yihong Zhang, Quan Z Sheng, and Salil S Kanhere (2019). "Source-Aware Crisis-Relevant Tweet Identification and Key Information Summarization". In: *ACM Transactions on Internet Technology (TOIT)* 19.3,

pp. 1–20. ISSN: 1533-5399. DOI: 10.1145/3300229. URL: https://doi.org/10.1145/3300229.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Sarah Vieweg (2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *Proceedings of the Eighth International AAAI Conference on Web and Social Media* 8.1, pp. 376–385. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14538.

Olteanu, Alexandra, Sarah Vieweg, and Carlos Castillo (2015). "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*. New York, NY, USA: Association for Computing Machinery, pp. 994–1009. ISBN: 9781450329224. DOI: 10.1145/2675133.2675242. URL: https://doi.org/10.1145/2675133.2675242.

Parilla-Ferrer, Beverly Estephany, Proceso L Fernandez, and Jaime T Ballena (2014). "Automatic Classification of Disaster-Related Tweets". In: *International conference on Innovative Engineering Technologies (ICIET)*. Vol. 62. Bangkok, Thailand, pp. 62–69.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://aclanthology.org/D14-1162.

Petrovic, Sasa, Miles Osborne, Richard McCreadie, Craig Macdonald, Iadh Ounis, and Luke Shrimpton (2013). "Can Twitter Replace Newswire for Breaking News?" In: *Pro-*

ceedings of the International AAAI Conference on Web and Social Media. Vol. 7. 1, pp. 713–716. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14450.

Petrović, Saša, Miles Osborne, and Victor Lavrenko (2010). "Streaming first story detection with application to twitter". In: *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*. Association for Computational Linguistics, pp. 181–189.

Phillips, Mark Edward (2017). *Hurricane Harvey Twitter Dataset*. URL: https://digital.library.unt.edu/ark:/67531/metadc993940/.

Pilehvar, Mohammad Taher and Jose Camacho-Collados (2021). *Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning*. Springer Cham. ISBN: 978-3-031-02177-0. DOI: https://doi.org/10.1007/978-3-031-02177-0.

Qazi, Umair, Muhammad Imran, and Ferda Ofli (2020). "GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information". In: *SIGSPATIAL Special* 12.1, pp. 6–15. DOI: 10.1145/3404820.3404823. URL: https://doi.org/10.1145/3404820.3404823.

Rachunok, Benjamin, Chao Fan, Ronald Lee, Roshanak Nateghi, and Ali Mostafavi (2022). "Is the data suitable? The comparison of keyword versus location filters in crisis informatics using Twitter data". In: *International Journal of Information Management Data Insights* 2.1, pp. 1–11. ISSN: 2667-0968. DOI: https://doi.org/10.1016/j.jjimei.2022.100063. URL: https://www.sciencedirect.com/science/article/pii/S2667096822000076.

Rahutomo, Faisal, Teruaki Kitasuka, and Masayoshi Aritsugi (2012). "Semantic cosine similarity". In: *The 7th international student conference on advanced science and technology ICAST*. Vol. 4. 1, p. 1.

Al-Rakhami, Mabrook S. and Atif M. Al-Amri (2020). "Lies Kill, Facts Save: Detecting COVID-19 Misinformation in Twitter". In: *IEEE Access* 8, pp. 155961–155970. DOI: 10.1109/ACCESS.2020.3019600.

Ramponi, Alan and Barbara Plank (Dec. 2020). "Neural Unsupervised Domain Adaptation in NLP—A Survey". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 6838–6855. DOI: 10.18653/v1/2020.coling-main.603. URL: https://aclanthology.org/2020.coling-main.603.

Ray Chowdhury, Jishnu, Cornelia Caragea, and Doina Caragea (July 2020). "Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Online, pp. 292–298. DOI: 10.18653/v1/2020.acl-srw.39. URL: https://aclanthology.org/2020.acl-srw.39.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://aclanthology.org/D19-1410.

— (Nov. 2020). "Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online, pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365. URL: https://aclanthology.org/2020.emnlp-main.365.

Rish, Irina (2001). "An Empirical Study of the Naive Bayes Classifier". In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. Vol. 3. 22. IBM New York, pp. 41–46.

Ritter, Alan, Mausam, Oren Etzioni, and Sam Clark (2012). "Open domain event extraction from twitter". In: *The 18th ACM SIGKDD International Conference on Knowledge*

Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012. ACM, pp. 1104–1112. DOI: 10.1145/2339530.2339704.

Röder, Michael, Andreas Both, and Alexander Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". In: WSDM '15. Shanghai, China: Association for Computing Machinery, pp. 399–408. ISBN: 9781450333177. DOI: 10.1145/2684822.2685324. URL: https://doi.org/10.1145/2684822.2685324.

Ruder, Sebastian (2019). "Neural Transfer Learning for Natural Language Processing". PhD thesis. NUI Galway. URL: http://ruder.io/thesis/neural_transfer_learning_for_nlp.pdf.

Ruder, Sebastian, Parsa Ghaffari, and John G Breslin (2017). "Data Selection Strategies for Multi-Domain Sentiment Analysis". In: *CoRR* abs/1702.02426. arXiv: 1702.02426. URL: http://arxiv.org/abs/1702.02426.

Rudra, Koustav, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh (2018). "Extracting and Summarizing Situational Information from the Twitter Social Media during Disasters". In: *ACM Trans. Web* 12.3, pp. 1–35. ISSN: 1559-1131. DOI: 10.1145/3178541. URL: https://doi.org/10.1145/3178541.

Rudra, Koustav, Subham Ghosh, Niloy Ganguly, Pawan Goyal, and Saptarshi Ghosh (2015). "Extracting Situational Information from Microblogs during Disaster Events: A Classification-Summarization Approach". In: CIKM '15. Melbourne, Australia: Association for Computing Machinery, pp. 583–592. ISBN: 9781450337946. DOI: 10.1145/2806416.2806485. URL: https://doi.org/10.1145/2806416.2806485.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo (2010). "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors". In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: Association for Computing Machinery, pp. 851–860. ISBN: 9781605587998. DOI: 10.1145/1772690.1772777. URL: https://doi.org/10.1145/1772690.1772777.

Salton, Gerard, Anita Wong, and Chung-Shu Yang (1975). "A Vector Space Model for Automatic Indexing". In: *Commun. ACM* 18.11, pp. 613–620. ISSN: 0001-0782. DOI: 10.1145/361219.361220. URL: https://doi.org/10.1145/361219.361220.

Shoufan, Abdulhadi and Sumaya Alameri (2015). "Natural language processing for dialectical Arabic: A survey". In: *Proceedings of the second workshop on Arabic natural language processing*, pp. 36–48.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu (2017). "Fake News Detection on Social Media: A Data Mining Perspective". In: *ACM SIGKDD Explorations Newsletter* 19.1, pp. 22–36. ISSN: 1931-0145. DOI: 10.1145/3137597.3137600. URL: https://doi.org/10.1145/3137597.3137600.

Singh, Gurinder, Bhawna Kumar, Loveleen Gaur, and Akriti Tyagi (2019). "Comparison between multinomial and Bernoulli Naïve Bayes for text classification". In: *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*. IEEE, pp. 593–596.

Singh, Jyoti Prakash, Yogesh K Dwivedi, Nripendra P Rana, Abhinav Kumar, and Kawaljeet Kaur Kapoor (2017). "Event classification and location prediction from tweets during disasters". In: *Annals of Operations Research*, pp. 1–21.

Singh, Lisa, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily K. Vraga, and Yanchen Wang (2020). "A first look at COVID-19 information and misinformation sharing on Twitter". In: *CoRR* abs/2003.13907. arXiv: 2003.13907. URL: https://arxiv.org/abs/2003.13907.

Sit, Muhammed Ali, Caglar Koylu, and Ibrahim Demir (2019). "Identifying disaster-related tweets and their semantic, spatial and temporal context using deep learning, natural language processing and spatial analysis: a case study of Hurricane Irma". In: *International Journal of Digital Earth* 12.11, pp. 1205–1229. DOI: 10.1080/17538947.2018.1563219.

Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy (2017). "AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP". In: *Procedia Computer Science* 117, pp. 256–265. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2017.10.117. URL: https://www.sciencedirect.com/science/article/pii/S1877050917321749.

Sorower, Mohammad S (2010). "A Literature Survey on Algorithms for Multi-label Learning". In: *Oregon State University, Corvallis* 18, pp. 1–25.

Steiger, Enrico, João Porto Albuquerque, and Alexander Zipf (2015). "An Advanced Systematic Literature Review on Spatiotemporal Analyses of Twitter Data". In: *Transactions in GIS* 19.6, pp. 809–834.

Sutton, Jeannette N, Leysia Palen, and Irina Shklovski (2008). "Backchannels on the Front Lines: Emergent Uses of Social Media in the 2007 Southern California Wildfires". In: *Proceedings of the 5th International ISCRAM Conference*, pp. 624–631.

Syed, Shaheen and Marco Spruit (2017). "Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation". In: *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 165–174. DOI: 10.1109/DSAA.2017.61.

Takahashi, Bruno, Edson C Tandoc Jr, and Christine Carmichael (2015). "Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines". In: *Computers in human behavior* 50, pp. 392–398. DOI: https://doi.org/10.1016/j.chb.2015.04.020.

To, Hien, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi (2017). "On Identifying Disaster-Related Tweets: Matching-Based or Learning-Based?" In: *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*. IEEE, pp. 330–337. DOI: 10.1109/BigMM.2017.82.

Torres Carmen Vaca, Johnny (2019). "Cross-Lingual Perspectives about Crisis-Related Conversations on Twitter". In: *Companion Proceedings of The 2019 World Wide Web Conference*. WWW '19. New York, NY, USA: Association for Computing Machin-

ery, pp. 255–261. ISBN: 9781450366755. DOI: 10.1145/3308560.3316799. URL: https://doi.org/10.1145/3308560.3316799.

Tufekci, Zeynep and Christopher Wilson (2012). "Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square". In: *Journal of Communication* 62.2, pp. 363–379.

Turney, Peter D and Patrick Pantel (2010). "From Frequency to Meaning: Vector Space Models of Semantics". In: *Journal of Artificial Intelligence Research* 37.1, pp. 141–188. ISSN: 1076-9757.

Uniyal, Deepak and Amit Agarwal (2021). "IRLCov19: A Large COVID-19 Multilingual Twitter Dataset of Indian Regional Languages". In: Cham: Springer International Publishing, pp. 309–324. ISBN: 978-3-030-93733-1.

Van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.11, pp. 2579–2605.

Verma, Sudha, Sarah Vieweg, William Corvey, Leysia Palen, James Martin, Martha Palmer, Aaron Schram, and Kenneth Anderson (2011). "Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency". In: *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Vol. 5. 1, pp. 385–392.

Vieweg, Sarah, Amanda L. Hughes, Kate Starbird, and Leysia Palen (2010). "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*. CHI '10. Atlanta, Georgia, USA: Association for Computing Machinery, pp. 1079–1088. ISBN: 9781605589299. DOI: 10.1145/1753326.1753486. URL: https://doi.org/10.1145/1753326.1753486.

Wang, Haoyu, Eduard Hovy, and Mark Dredze (2015). "The Hurricane Sandy Twitter Corpus". In: *AAAI Workshop: WWW and Public Health Intelligence*.

Weng, Jianshu and Bu-Sung Lee (2011). "Event Detection in Twitter". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 5. 1, pp. 401–408.

Wiegmann, Matti, Jens Kersten, Friederike Klan, Martin Potthast, and Benno Stein (2020). "Analysis of Detection Models for Disaster-Related Tweets". In: *Proceedings of the 17th ISCRAM Conference*. Blacksburg, VA, USA, pp. 872–880.

Zahra, Kiran, Muhammad Imran, and Frank O Ostermann (2020). "Automatic identification of eyewitness messages on twitter during disasters". In: *Information Processing & Management* 57.1, p. 102107. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2019.102107.

Zaidan, Omar F. and Chris Callison-Burch (Mar. 2014). "Arabic Dialect Identification". In: *Computational Linguistics* 40.1, pp. 171–202. DOI: 10.1162/COLI_a_00169. URL: https://aclanthology.org/J14-1006.