Identifying and Processing Crisis Information from Social Media

Thesis

Identifying and Processing Crisis Information from Social Media

Prashant Khare

a thesis submitted

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Knowledge Media Institute
Open University
Milton Keynes, United Kingdom
February 2020

Supervisors:
Professor Harith Alani
Dr. Grégoire Burel

Thesis advisor: Professor Harith Alani Prashant Khare

Identifying and Processing Crisis Information from Social Media

Abstract

Social media platforms play a crucial role in how people communicate, particularly during crisis situations such as natural disasters. People share and disseminate information on social media platforms that relates to updates, alerts, rescue and relief requests among other crisis relevant information. Hurricane Harvey and Hurricane Sandy saw over tens of millions of posts getting generated, on Twitter, in a short span of time. The ambit of such posts spreads across a wide range such as personal and official communications, and citizen sensing, to mention a few. This makes social media platforms a source of vital information to different stakeholders in crisis situations such as impacted communities, relief agencies, and civic authorities. However, the overwhelming volume of data generated during such times, makes it impossible to manually identify information relevant to crisis. Additionally, a large portion of posts in voluminous streams is not relevant or bears minimal relevance to crisis situations.

This has steered much research towards exploring methods that can automatically identify crisis relevant information from voluminous streams of data during such scenarios. However, the problem of identifying crisis relevant information from social media platforms, such as Twitter, is not trivial given the nature of unstructured text such as short text length and syntactic variations among other challenges. A key objective, while creating automatic crisis relevancy classification systems, is to make them adaptable to a wide range of crisis types and languages. Many related approaches rely on statistical features which are quantifiable properties and linguistic properties of the text. A general approach is to train the classification model on labelled data acquired from crisis events and evaluate on other crisis events. A key aspect missing from explored literature is the validity of crisis relevancy classification models when applied to data from unseen types of crisis events and languages. For instance, how would the accuracy of a crisis relevancy classification model, trained on earthquake type of events, change when applied to flood type of events. Or, how would a model perform when trained on crisis data in English but applied to data in Italian.

This thesis investigates these problems from a *semantics* perspective, where the challenges posed by diverse types of crisis and language variations are seen as the problems that can be tackled by enriching the data semantically. The use of knowledge bases such as DBpedia, BabelNet, and Wikipedia, for semantic enrichment of data in text classification problems has often been studied. Semantic enrichment of data through entity linking and expansion of context via knowledge bases can take advantage of connections between different concepts and thus enhance contextual coherency across crisis types and languages. Several previous works have focused on similar problems and proposed approaches using statistical features and/or non-semantic features. The use of semantics extracted through knowl-

edge graphs has remained unexplored in building crisis relevancy classifiers that are adaptive to varying crisis types and multilingual data. Experiments conducted in this thesis consider data from Twitter, a micro-blogging social media platform, and analyse multiple aspects of crisis data classification. The results obtained through various analyses in this thesis demonstrate the value of semantic enrichment of text through knowledge graphs in improving the adaptability of crisis relevancy classifiers across crisis types and languages, in comparison to statistical features as often used in much of the related work.

To my parents, Sadhna & Pradeep Kumar Khare.

Acknowledgments

The last four years of my PhD have been an enriching experience, not only as a student but also at a personal level. These years have reinstated my belief in the idea that patience, diligence, honesty, and faith are still the key to pursue and accomplish goals. I realised how important it is to have a clear vision of one's goals, whether it is during setting up a scientific experiment or otherwise. A lot of it is attributed to my thesis supervisor, Professor Harith Alani, to whom I convey my sincerest gratitude. His ability to make one refine their thought process and ideas soundly, enhanced my research outcomes and scientific skills. Prof. Alani's influence lasts beyond the boundaries of research outcomes. His leadership instills mutual trust and respect among colleagues, which helps in bringing out the best. He personifies the classic definition of a mentor. I have been truly fortunate to have Prof. Alani as my PhD supervisor.

I am also sincerely grateful to my two second supervisors, Dr. Grégoire Burel and Dr. Miriam Fernandez, who took up the responsibility of enriching my scientific prowess, during two different phases of my PhD. The inputs provided by Dr. Fernandez during the earliest phase of PhD, eventually led me to expand the research scope. She, along with Prof. Alani, was very supportive during some of the testing times in the early phases. Dr. Burel helped me in enhancing the quality of scientific analysis, thereby ensuring that the research outcomes are as close to being flawless. His inputs contributed towards shaping up the most productive phases of my studies. I also thank my PhD examiners, Dr. Danica Vukadinovic Greetham and Prof. Pete Burnap, for their valuable inputs catering towards the completion of this thesis.

I would also extend my heartfelt thanks to all the staff and students at Knowledge Media Institute (KMI), who never stop inspiring by upholding the research standards to the highest level. KMI is truly representative of the world as one community. A place where diversity is perceived as a unifying force. I believe this place has made me a more humble human.

I would not have been able to pursue challenges and goals, had my parents not shown me the path of perseverance. Their values are a guiding force for me. My parents, sister, and family have been a continuous support throughout, and taught me to value success as well as failures. Lastly, I would like to thank my wife, Sanjoli, for her unconditional support and helping me build a life. Her presence has always cheered me up, which kept me going throughout my studies.

Contents

List of Tables

Listing of figures

1

Introduction

Information is best valued when conveyed and acted upon in a timely fashion. Determining which information is valuable is pivotal during crisis situations. Crisis situations, generally, refer to natural or human-induced disaster phenomena impacting safety and well being of people. Nowadays, social media platforms play a crucial role in information dissemination during crisis situations. People tend to share posts across social media platforms, related to updates, alerts, rescue in-

formation, and relief requests among other content. During Hurricane Harvey, more than 7 million tweets were posted in just over a month[*]. During Hurricane Sandy, more than 20 million tweeets were shared on Twitter[†] with the hashtags #sandy and #hurricane. The scope of social media platforms, in general, has been shown to spread across a wide range of areas, such as personal communications, citizen sensing, official communication, to name a few (Reuter et al., 2011; Reuter et al., 2012). In the course of crisis situations, social media platforms have been found to be intensely used by people to update their personal connections, such as with family or friends, to confirm their well-being or to signal that they require assistance (Olteanu et al., 2015; Vieweg, 2012). This has acted as a motivation for systems such as Facebook's *Crisis Response & Safety Check* system[‡] which is aimed towards channelising crisis response (Castillo, 2016).

Such a usage of social media platforms has turned them into a rich source of vital information in the course of crisis events. A study conducted by Rice University[§] found out that the damage maps produced by the Federal Emergency Management Agency (FEMA) during Hurricane Harvey, missed nearly 46% of the actual damage which was in fact reported on Twitter by the impacted individuals. However, given the overwhelming volume of data that gets generated on social media platforms makes it challenging to sieve through such a stream manually to identify relevant information. A vast amount of data makes these streams chaotic. Many of the messages found in such streams of messages bear minimal or no relevance to particular crisis situations, even the ones that contain crisis-specific hashtags. Many organisations and people that often deal with emergency management have highlighted that most of the messages they come across on social media during emergency situations do not appear to be related and useful (Ludwig et al., 2015). However, despite these challenges, the sig-

[*]Hurricane Harvey Twitter Dataset, digital.library.unt.edu/ark:/67531/metadc993940/

[†]Mashable: Sandy Sparks 20 Million Tweets, https://mashable.com/2012/11/02/hurricane-sandy-twitter/

[‡]https://www.facebook.com/about/crisisresponse/

[§]https://kinder.rice.edu/sites/g/files/bxs1676/f/documents/FinalTwitter%20report%20KI%202018%20Research%20Report-Lessons%20from%20Harvey%203.pdf

nificance of social media in the course of crisis events has been well identified by government and relief agencies[*].

This has driven a significant amount of research exploring methods to automatically determine the relevant information in crisis scenarios, from the voluminous streams of social media data. Automatic detection of crisis-relevant messages from social media platforms is not a trivial task, considering the nature of the unstructured social media data such as short text length, colloquialism, and syntactic variations in the text. A principal goal while creating such automatic classification methods is to make them adaptive and applicable to a wide range of crisis events across various crisis *types* and *languages*. This is a significant limitation of most existing approaches that often focus on specific crisis types and on data written in specific languages. Different *types* of crisis situations result in a wide spectrum of data which gets posted online by people who are impacted in one way or another. Geographical and demographic diversity also results in multilingual data.

Given the high volume of crises data, there is a need for automated methods to detect their relevancy, and given the diversity in crisis types and languages, such methods must be able to process crisis data regardless of such variations. In this thesis, we explore the use of semantic representation, linking, and expansion to leverage the relations between words across varying crisis types and languages. For example, the two words *floods* and *earthquake* are types of *natural disasters*, which is a common concept linking both words. Creating automatic classification methods that are agnostic to varying crisis *types* and *languages* is a key aspect for overcoming the challenges above.

In this thesis, we investigate the impact that *semantics* could have on building classification models to identify crisis *related* information across diverse crisis events and languages. In particular, we explore how such classification models, based on semantic features or statistical features, perform under discrete settings where the model is applied to a new crisis event, a *type* of crisis event, crisis event in

[*] https://www.fema.gov/news-release/2018/04/16/social-media-and-emergency-preparedness

a new *language*, or a combination of these scenarios. Given Twitter's* popularity during crisis situations, its public nature, and ease of access to its data, we used it for collecting social media data during crisis events and performing the experiments. The research conducted in this thesis, broadly, makes the following contributions:

- Hybrid classification models are developed by combining statistical and semantic features for classifying Twitter data based on relevancy to crisis situations.
- Deepened understanding of how transferable the classifiers are when applied on (a) new crisis events, (b) new types of crises, (c) crises from different languages, and (d) crises of a different type in a different languages.
- Two approaches for classifying multilingual data are evaluated: using automatic translators, and using semantic information extraction.

1.1 Motivation

In the course of crisis events there is usually an increase in content generation on social media and also in content demand. Online searches related to crisis specific terminology tend to increase during such events (Guo et al., 2013). On social media platforms, people often provide an account of their experiences, and also seek information to raise their awareness or to support their decision making. For example, internet usage in the East Coast of the United States was reportedly found to have increased by 114% when Hurricane Sandy was about to hit†. During the 2011 earthquake and tsunami in Japan, there was a 500% increase in the tweets from Japan as people reached out to family and friends ‡. As mentioned earlier, crisis events such as Hurricane Harvey and Sandy witnessed millions of tweets being posted in a months period.

*Twitter, https://twitter.com/

†https://www.zdnet.com/article/internet-usage-rocketed-on-the-east-coast-during-sandy-report/

‡https://blog.twitter.com/official/en_us/a/2011/global-pulse.html

Vieweg (Vieweg, 2012) stressed that the content generated during crisis situations does contribute to situational awareness, and there are different categories among which the users tend to post their tweets. Several other works have studied the presence of crisis related information in social media posts generated across crisis events (Bruns et al., 2011; Kanhabua & Nejdl, 2013; Metaxas & Mustafaraj, 2013; Munro & Manning, 2012; Olteanu et al., 2014; Qu et al., 2011; Starbird et al., 2010; Thomson et al., 2012; Vieweg et al., 2010). It has also been reported that in the course of crisis events, there is usually a decline in self-oriented and context-free posts, while an increase in goal-driven and information oriented posts potentially increase (Naaman et al., 2010). Also, users tend to promote retweeting (a term given to re-sharing someone else' post on Twitter) (Heverin & Zach, 2010; Hughes & Palen, 2009).

These findings provide enough evidence of the significance of social media platforms, particularly Twitter, for people to rely on in the course of crisis situations to share and subscribe to relevant information.

1.1.1 Crisis Relevancy Classification: Opportunities

There are several works that focus on building classification models for crisis relevant data (Burel et al., 2017b; Burel et al., 2017a; Imran et al., 2016b; Pedrood & Purohit, 2018; Li et al., 2015; Li et al., 2017; Li et al., 2018a; Li et al., 2012b; Imran et al., 2013b). Many of these approaches rely on using statistical features from data. Statistical features reflect statistical (e.g., length, number of words, special characters, etc.) and linguistic attributes (e.g., part of speech) of the text. Generally, related works adopt the approach of training a classification model on labelled data from crisis events and then evaluate the model on other events. A major limitation observed, while covering the appropriate literature in depth (as covered in detail in Chapter 2), was that while many of the works do undertake this problem from a perspective of model adaptability where a supervised classification model should be applicable to new unseen crisis data, they do not consider the distinctiveness in the *type* of crisis

events. These works do not highlight the limitation experienced by the classification models when they are applied to data from an entirely new *type* of crisis event. For instance, how will a classification model perform if it was trained on data from *earthquake* type of events and applied to data from *flood* type of events. The second limitation from the crisis data point of view is the language, which is a major challenge in making language agnostic classification models. As a result of crisis events occurring around the globe and demographic diversity, crisis data is multilingual in nature. In fact, sometimes within a single crisis event there can be multilingual data. Variation in languages result in variations in the vocabulary of the data. If we aim to develop crisis relevancy classification systems that are applicable to diverse crisis scenarios, it is essential to ensure that a given classifier holds its applicability across varying languages. For instance, how will the classification model perform if it was trained on crisis data in *English* and applied to crisis data in *Spanish*.

In this thesis, we identify these areas as potential opportunities. We explore these problems with a *semantics* lens, where the diverse crisis types and variations in the language are envisioned as the problems that can be tackled by enriching the data with semantics. The use of knowledge sources such as WordNet[*], Wikipedia[†], and DBpedia[‡], for enriching the text semantically for text classification problems has often been studied (Siolas & d'Alché Buc, 2000; Hu et al., 2008; Abel et al., 2011). Various works have established that knowledge bases such as WordNet or Wikipedia can be exploited for identifying semantic similarities across different words (Agirre et al., 2009; Zhang et al., 2011). In the context of crisis data, enriching the data with semantic representations such as entity linking and expansion through knowledge bases can help in leveraging the connections between different concepts across varying crisis types and languages. Thus enhancing the contextual coherency in crisis data across crisis types and languages. We investigate multiple aspects of crisis data classification, and evaluate the impact of semantics of the data on adaptability of the classification models when the new unseen data

[*]WordNet, https://wordnet.princeton.edu/
[†]Wikipedia, https://www.wikipedia.org/
[‡]DBpedia, https://wiki.dbpedia.org/

is uniquely different in its origin.

In the next section, we describe our research questions and the hypotheses.

1.2 Research Questions and Hypotheses

The principal research question investigated in this thesis is:

"To what extent could semantics improve crisis relatedness classification of Twitter data?"

As mentioned earlier, we aim to build crisis relevancy classification models that are adaptive to new crisis events which might be of new types and languages. We breakdown our research into the following research questions.

- RQ1 - *How could the addition of semantics improve the binary classification of Tweets with regards to their relevancy to crises?*

Standard statistical features based classification models can be built to develop binary classification models that classify the data as crisis *related* or *not related* (Li et al., 2012b; Karimi et al., 2013; Zhang & Vucetic, 2016; Imran et al., 2016b). In this research question, we explore how using semantic representations, in the form of labels of linked entities which are annotated via a Named Entity Recognition (NER) service, and expansion to broader concepts such as hypernyms (broader concepts of entities) from a knowledge base, as features impact the performance of the classification model.

Hypotheses *H1*- Using semantic features such as labels of annotated entities (via NER services such as Babelfy) and hypernyms can help in dealing with the variations in language expressions. For example, *train station* and *railway station* are two different terms with different word representation, but the NER services built on word sense disambiguation techniques would normally point them both to the same concept - *railway station*. This strategy helps to overcome variations in the vocabulary which often refer to the same concepts. Similarly, expanding to broader concepts such as hypernyms

can bring contextual coherency by bridging discrete words in the vocabulary that share the same hypernyms. By bringing in such semantic features, the machine learning based classification approaches could perform better when classifying crisis data from several crisis events. Statistical and linguistic features such as part of speech, text length, etc., do not capture the contextual information which semantic features can. Thus, semantic features might influence a classification model to identify crisis relevant information where the attributes such as semantic similarities might be indicative of importance or vital with respect to the situation, which otherwise could be skipped by the classification system.

- RQ2 - *To what extent could semantics improve Tweets classification for new types of crisis events?*

In this research question, we want to study the scenario when a model is trained on certain types of events (e.g., earthquakes and train crashes) and tested on types of events which were not seen in the training data (e.g., floods, typhoons, etc.). We will further analyse whether adding semantics can boost the performance of the classifier model in such scenarios.

Hypotheses *H2*- We hypothesise that adding concepts and properties of entities (e.g., type of an entity, label of an entity, category of an entity, hypernyms) improves the identification of crisis information content across crisis domains, by creating a non-specific crisis contextual semantic abstraction of crisis-related content.

- RQ3 - *To what extent could semantics improve crisis-relevancy classification of Tweets written in a new language?*

In this research question, we explore the problem scenario when a classifier is strictly trained on crisis events in a particular language (for example data in *English*), and then the classifier is applied to data from crisis events which are in a different language (for example in *Italian*).

Hypotheses *H3*- We hypothesise that semantic features such as concepts and properties of entities (e.g., type of an entity, category of an entity, hypernyms) generalise the information representation of

the crisis situations across various languages. In cross-lingual classification problems, we also hypothesise that translation of the data to the same language, using automated machine translation systems, can also help in achieving the same goal since translation can align the cross-lingual data in terms of the vocabulary. We compare both the scenarios- adding the semantics and translation of the data.

- RQ4 - *To what extent could semantics improve Tweets classification when the type of crisis event and language change?*

In this research question, we explore the problem scenario when a classifier is strictly trained on certain *types* of crisis events and in some particular languages (for example data from earthquake events in *English*), and then the classifier is applied to data from different *types* of crisis events which are also in a different language (for example flood events in *Italian*). Quite often data in a certain crisis situation can arrive in multiple languages and hence models need to be adaptive to both new *types* of crises and new *language* at the same time.

Hypotheses *H4*- Following the above two research questions, we hypothesise that for the data that is from a new crisis type and in a new language at the same time, the semantic features along with translation can enable the classification models to become more adaptable to new incoming data for classifying. The semantics could not only handle the contextual alignment between diverse crisis domains but also handle the cross-lingual alignment as mentioned above. Similarly, translating the data can also align the cross-lingual data between the training and the test data. A combination of translation and semantic feature is expected to improve the performance of the classifier.

1.3 Methodology

The principal motivation behind this thesis is to explore the impact of adding semantics in the classification of crisis data across crisis types and languages. To this end, we propose a generic methodology, which is adapted across different experiments to answer the research questions posed.

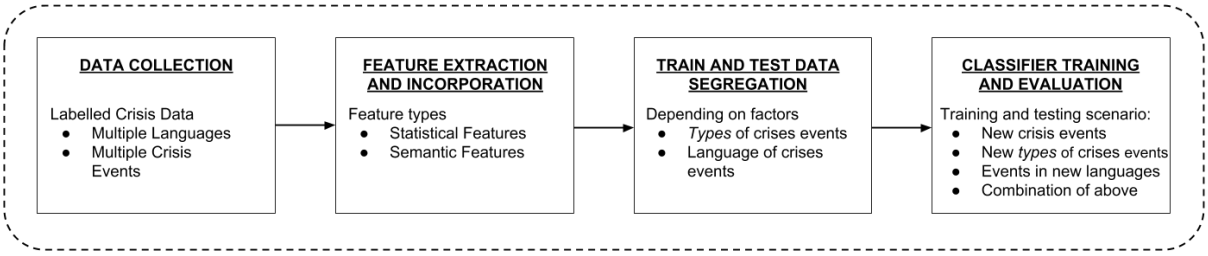

Figure 1.1: Methodology Pipeline

In general we have the following phases:

- **Data Collection**: We collect a dataset which comprises of multiple crisis events, and contains manually annotated labels reflecting the relatedness of documents (in this work we use tweets) with the crisis.

- **Feature Extraction and incorporation**: We extract two types of features - *Statistical Features* and *Semantic Features*. The *Semantic Features* are incorporated by concatenating with the original text and are represented as unigrams in the vector space, while the *Statistical Features* are used as unique features in the vector space. In some experiments we also employ translation services in order to analyse the impact of translation in multi-lingual crisis data. In Chapter 2, in section 2.1 and 2.1.1 we provide background knowledge about the type of semantics that we derive from text.

- **Train and Test Data Segregation**: We select the training data and test data of crisis events based on the scope of the particular research questions being addressed in a given experiment.

- **Classifier Training and Evaluation**: We train the classifier on the selected training data and evaluate the model across various combinations of test data.

As mentioned, we have different scopes defined for each experiment and accordingly we make variations in the adapted methodology. In some experiments we focus on the *types* of the crises, whereas

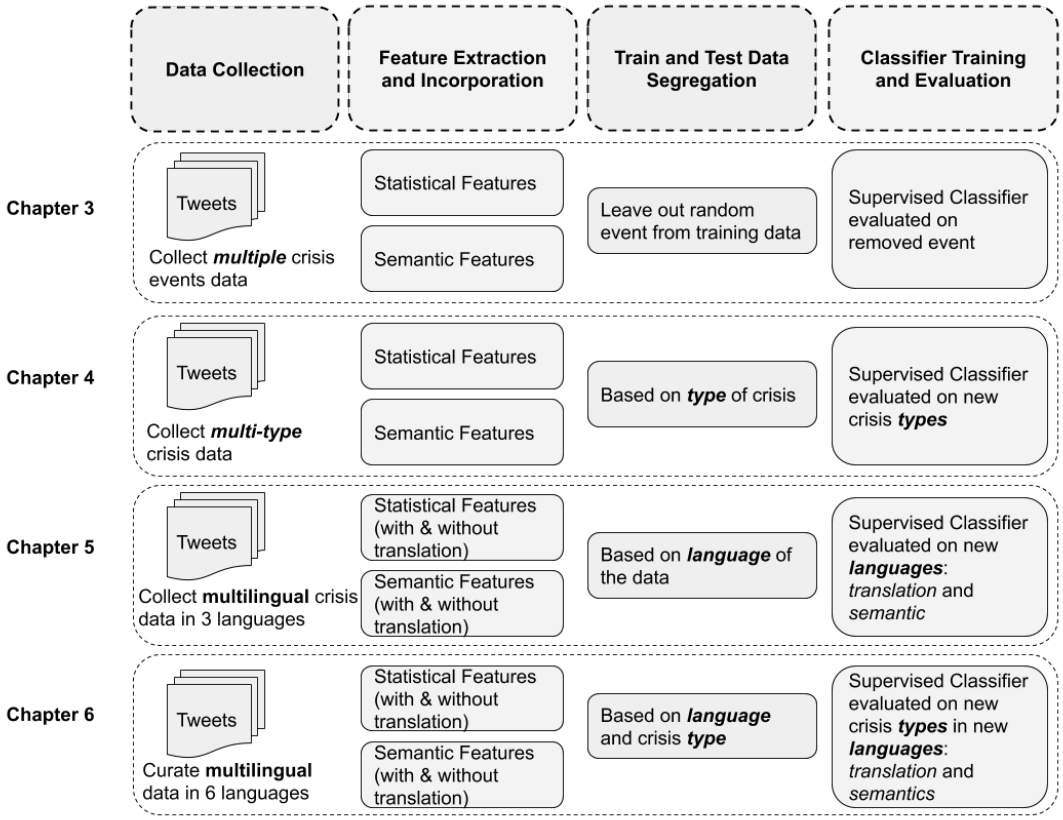

Figure 1.2: Methodology pipeline across each chapter, and thesis contribution outline

in others we focus on the language, and in other cases we focus on both. These factors drive the ways in which we select the data sets. Figure 1.1 shows the phases as a pipeline of the general methodology.

We use Twitter labelled data for all our experiments. There are some prominent data sets publicly available, which we have used in our experiments. In the evaluation phase, we compare the methodology with a particular baseline model which is developed in each set of experiment.

The thesis is structured as individual chapters, as follows:

- **Chapter 2** - '*Background and Literature Review*', we provide a background knowledge of text classification approaches, machine learning methods, machine learning features, natural

language processing, and semantics. Further, we describe the relevant literature focusing on processing and classifying social media data in general and during crises situations.

- **Chapter 3** - we present our work on using the semantics to classify the crisis related data across different crisis events. We train the classifier on some crisis events and create a model which is then validated on a new crisis event. We explore the impact of semantic features in comparison to statistical features. In this chapter, we address the first research question.

- **Chapter 4** - we present our work on using semantics to classify crisis related data across different *types* of crisis events. We train the classifier on some *types* of crisis events and create a model which is then validated on a new *types* of crisis events. We explore the impact of semantic features when the tested crisis *type* is not seen in the training data. In this chapter, we address the second research question.

- **Chapter 5** - we present our work on using the semantics, translation, and combination of adding semantics and translation, to classify crisis related data across different *languages*. We train the classifier on crisis events in a certain *language* and create a model which is then validated on new crisis events in a different *language*. We explore the impact of semantics, translation, and combination of adding semantics and translation when the tested crisis *language* is not seen in the training data. In this chapter, we address the third research question.

- **Chapter 6** - we present our work on combining the problem cases explored individually in Chapter 4 and Chapter 5, i.e., how classification models perform when applied to crisis events which are not only of a new *type* but also in a new *language*. We train the classifier on certain *types* of crisis events in a certain *language* and create a model which is then validated on new *types* of crisis events in a different *language*. We explore the impact of the semantics, translation, and combination of adding semantics and translation, when the tested crisis is of

a different *type* and *language* than what is seen in the training data. In this chapter, we address the fourth research question.

- **Chapter 7** - *Discussion*, we discuss the research work presented in the thesis, highlight the scientific outcome, and point towards future work.
- **Chapter 8** - *Conclusion*, we present the main conclusions of the thesis.

1.4 Publications

Each individual chapter which addresses a research question in this thesis is reflected in a peer reviewed workshop/conference/journal paper. Here, we highlight the publications which are based on individual chapters of this thesis:

- *Chapter 3*
  - Khare, P., Fernandez, M., & Alani, H. (2017). Statistical semantic classification of crisis information. In 1st workshop *HSSUES* at *International Semantic Web Conference*, Vienna, Austria.
- *Chapter 4*
  - Khare, P., Burel, G., & Alani, H. (2018). Classifying crises-information relevancy with semantics. In *European Semantic Web Conference* (pp. 367–383), Heraklion, Crete.: Springer.
- *Chapter 5*
  - Khare, P., Burel, G., Maynard, D., & Alani, H. (2018). Cross-lingual classification of crisis data. In *International Semantic Web Conference* (pp. 617–633), Monterey, US.: Springer.

- *Chapter 6*

    - Khare, P., Burel, G., & Alani, H. (2019). Relevancy identification across languages and crisis types. *IEEE Intelligent Systems Journal*.

2

Background and Literature Review

In the Chapter 1, we discussed the motivation and the opportunities that exist in analysing social media content, particularly Twitter during crisis situations. The work done in this thesis focuses on the textual data generated on such social media data. In this chapter we will cover the background of the techniques which are often used to handle and process text data, and cover an extensive literature on various research works that focus on processing crisis oriented data. Many related approaches exploit natural language processing tools and machine learning methods to generate insights from unstruc-

tured text data. Natural language processing tools are primarily used to break down a natural language text into several tokens which can individually be treated as features. Machine learning methods are used for classification tasks in order to identify posts as belonging to a certain class based on how the training data is labelled (explained more in section 2.2).

In the subsequent sections, we will cover fundamental approaches on natural language processing and semantics of the text, machine learning methods for text classification, and then gradually focus on a detailed literature review.

2.1 Natural Language Processing and Text Semantics

Twitter is one of the most prominent micro-blogging online service, with almost 326 million monthly active users, as of the year 2018[*]. The platform enables people to post short text posts called *Tweets*, of maximum length 280 characters (until the year 2017 it used to be 140 characters[†]) and also share photos and/or videos along with the text post. Other popular micro-blogging platforms are Facebook, YouTube, WhatsApp, LinkedIn, etc[‡]. Twitter makes public posts accessible via application programming interface (APIs)[§], while Facebook limits the access of user's posts based on privacy settings. Platforms such as LinkedIn are more focused towards professional discussions.

Let us look at a tweet posted during floods in Alberta in 2013:

RT @KaleighRogers: A 15-year-old High River (Calgary) boy is missing since floods . Call police if you see Eric #abflood

---

This tweet is a *Re-tweet* (re-shared by a user) of the original tweet posted by a user with the user

[*] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[†] https://www.washingtonpost.com/news/the-switch/wp/2017/11/07/twitter-is-officially-doubling-the-character-limit-to-280/?noredirect=on

[‡] Global social networks ranked by number of users 2019 https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[§] https://developer.twitter.com/en/docs/api-reference-index.html

handle @KaleighRogers. This post says that a *boy* named Eric aged 15 years, from the town *High River* in Alberta, is missing because of the floods and report to the police if seen by anyone. While it may not be a challenge for the humans, for the machines short text is not trivial to make sense of, due to often vague and open interpretation, particularly when building computational models. The contextual information is ambiguous and not substantive. Natural Language Processing (NLP) is the field of study which focuses on programming computers to process and analyze natural language data*. In this section, we will cover the Natural Language Processing (NLP) techniques, that will provide an overview of the basic computational methods to process the natural language (in text form).

To begin with, we will look at some of the key text processing operations, as also highlighted by Castillo (Castillo, 2016). These operations aim at converting an input text to a structured text segments.

– **Character encoding/decoding** - converts an input text string into an array of characters to an array of bytes and vice versa. A character encoding converts each character to its corresponding byte code and decoding will look up into the character table to convert it back to the character. Character encoding is done to ensure that the machine understands which particular character exists in the text. For instance, UTF-8 is a popular character encoding.

– **Tokenisation** is a process of sequencing a set of strings (as in a sentence) to an array/list of individual tokens.

> *'A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.'* (Manning et al., 2008)

For instance, if we look at the above tweet, it should result in a list of 23 individual tokens- which is each word in the sentence. An illustration below, in Table 2.1, shows another example.

One can also take into consideration the minimum length of the tokens. For instance, in the above

*Natural Language Processing, https://en.wikipedia.org/wiki/Natural_language_processing

Table 2.1: Illustrative example for tokenisation

Text	Breaking News: MET issues a storm warning
Tokenised	Breaking | News | MET | issues | a | storm | warning

case if we consider minimum length as 2, then the tokenised form will filter out the token *'a'* from the set of tokens. Trim (Trim, 2013) provides a more detailed approach towards tokenisation.

– **Normalisation** is the process of transforming the text to a single standard or canonical form, thereby ensuring that the data is consistent in its format before performing operations. This phase is application dependent, and can be designed as per the requirements. For instance, *'BREAKING NEWS'*, *'Breaking News'*, and *'breaking news'* can be normalised by converting all the characters from upper case to lower case. Another standard method is to handle abbreviations, so 'U.S.A' and 'USA' could be normalised by removing the punctuation marks and briging all such variations to a common representation. In some cases, the acronyms (for instance terms used in slang or urban vocabulary) can be resolved to their more standard forms, such as 'idk' can be resolved to 'i don't know'. But this is a very specific approach to focus on a certain type of data and it requires large scale lexicons and/or dictionaries dedicated for such problems.

– **Stop-word Filtering** is aimed at removing the tokens (words) from the data that are not treated as useful in a given scenario. In general, stop-words in any language are very commonly occurring words (quite often are function words* such as *the, is, for, of etc.*). Such function words are assumed to have no distinct context in themselves, and are usually used to connect different segments of a sentence. The scope of stop-words is languages as well as application dependent. A simple representation of stop-words filtering is shown below.

A large collection of stop-words across several languages can be collected †.

*https://en.wikipedia.org/wiki/Function_word

†https://code.google.com/archive/p/stop-words/

With Stop-Words	I started a T-Shirt campaign to benefit the #coloradoflood victims
Without Stop-Words	started T-Shirt campaign benefit #coloradoflood victims

– **Stemming and lemmatisation** – Often for grammatical requirements varying forms of a word are chosen for sentence construction such as *conducting*, *conducts conducted*, *conduct*. In this case, stemming would trim the word if it ends with *'s'*, *'ed'*, *'ing'* and replace it with the character *'i'*, thus bringing it to a form which can be consistent across a dataset. There are various rules to stemming across languages. In English, most popularly used stemmer is Porter's stemmer (Porter, 1980). Stemming may not necessarily result in an actual word (root word), but it helps in normalising the data. In certain cases, the words can be a derivative of a root word. For instance, *run*, *running*, *runs*, *ran* are all varying form of the lemma *run*. Lemmatisation will not trim the word, as in Stemming, and rather look for corresponding lemma of a given word depending on the part of speech a word represents. The aim of, both, stemming and lemmatisation is to bring the words in a standard form across a data set, and can also be treated as another form of *normalisation*.

– **Part of Speech (POS) tagging** – is the process of annotating words in a sentence with a part of speech tag the words belongs to. Part of speech is a language dependent classification of the class of words in any given sentence. Some well known and understood part of speech are *noun*, *verb*, *adjective*, *pronoun*, *adverb*, *conjunction* etc. There are computational models available for automatically annotating the part of speech in text. A piece of text is provided as an input and the automatic taggers provide sequence list of annotations as an output. Stanford NLP POS tagger (Toutanova & Manning, 2000; Toutanova et al., 2003) are widely used Part-of-Speech taggers for computational automatic tagging, and has tagger models for multiple languages. A representation of POS tagging is shown below.

Text	Thoughts with boulder flood affected				
Tokens	Thoughts	with	boulder	flood	affected
POS	Noun	Preposition	Noun	Noun	Verb

2.1.1 Named Entities and Knowledge base

Named entities are the references of *persons*, *locations*, and/or *organisations* that can be denoted with a proper name, in a text.

Text	*Obama* has declared emergency in *Colorado* after flooding

For instance, in the text above, *Obama* is a *Person* and *Colorado* is a *location*. Named Entity Recognition (NER) is about finding such named entities. Many NER approaches have been derived from statistical modelling methods such as Conditional Random Fields (CRF) sequence models. Several CRF approaches have been studied (Lafferty et al., 2001; Sutton & McCallum, 2006; Sutton et al., 2012). Stanford NER by Finkel et al. (Finkel et al., 2005) is a leading work for the contemporary methods on Named Entity Recognition.

Once the named entities are recognised, it would help to generate extra meaning or *semantics* of the entities, that will help in addressing the ambiguity regarding the exact concept being referred to in the text. For instance, let us look at the two texts below.

Table 2.2: Example- Entity Disambiguation

Text1	We saw a bright *Jaguar* speeding on the motorway
Text2	We saw a *Jaguar* chasing after a prey in the jungle

The *Jaguar* in Text1, in Table 2.2, is more likely to be referring to a Jaguar Car*, whereas the second

*https://en.wikipedia.org/wiki/Jaguar_Cars

reference in Text2 is supposedly about a Jaguar, the wild animal*. This phenomenon of connecting the recognised entities with a specific machine readable identifier is called *Named Entity Linking* (or also *Named Entity Disambiguation/Resolution*). The context of a given entity is determined by the type of concepts that co-occur with it in the text. In Text1, *Jaguar* was seen along with the concepts such as *speeding* and *highway*. Intuitively, if we are talking about a *Jaguar* on a highway, which is *speeding*, it is more likely to think of a *car/vehicle* instead of an animal. But this also requires a contextual background where we are already aware of a car named *Jaguar* and also a knowledge where usually the cars are seen on highways/motorways.

In order to link the entities to a machine readable identifier, it is required to have such a large contextual database. In the literature (Auer et al., 2007; Lehmann et al., 2015; Suchanek et al., 2007; Rebele et al., 2016), such contextual databases are termed as *Knowledge Base*, and are often developed on top of the knowledge extracted from encyclopedic form of information sources, such as crowd-sourced free online encyclopedia Wikipedia†. Wikipedia has often been used for studying Named Entity Disambiguation techniques because of its huge size of documentation on a wide range of topics (Bunescu & Paşca, 2006; Cucerzan, 2007; Ratinov et al., 2011; Zhou et al., 2010)

There are many popular knowledge bases such as DBpedia (Auer et al., 2007; Lehmann et al., 2015), YAGO (Suchanek et al., 2007; Rebele et al., 2016), and BabelNet (Navigli & Ponzetto, 2010; Navigli & Ponzetto, 2012) to name a few. Many of such knowledge bases are a multilingual semantic network of concept and entities, and extend beyond just Wikipedia in terms of knowledge source aggregation. For instance, YAGO and BabelNet are integration of lexicographic and encyclopedic knowledge from WordNet (a lexical database of English where words are grouped as per their meaning and in a hierarchy of hyponymy and synonymy), Wikipedia, and Wikidata. These knowledge bases enable us to iterate through a variety of semantic relationships for any given concept/entity.

*https://en.wikipedia.org/wiki/Jaguar

†https://en.wikipedia.org/wiki/Main_Page

There are several Named Entity Disambiguation services that are built on top of the aforementioned knowledge bases, such as AIDA (Hoffart et al., 2011) (built on YAGO), DBpedia-Spotlight (Daiber et al., 2013; Mendes et al., 2011) (built on DBpedia), and Babelfy (Moro et al., 2014) (built on BabelNet). These services not only perform Named Entity Disambiguation but also perform Word Sense Disambiguation. Word Sense Disambiguation is an approach to disambiguate the meaning of any word (instead of just Named Entities) in a given text, when the possibility is more than one. As an example, consider the following text:

Text	A 15 year old High River boy is missing due to flood. Call police if you see Eric St. Denis

Through the entity linking methods and further extension of information from knowledge bases, we can determine that *High River* is a locality/town in Calgary, Canada. Also, it can be determined that the word *missing* here implies being *lost*.

There are many such knowledge bases, also domain specific knowledge bases such as Geonames*, which is a geographical database. One of the ways to find out diverse knowledge bases is via Linked Open Data Cloud†, which links several open access knowledge bases.

We will subsequently use such natural language processing and semantic enhancing methods in formulating our methodology to classify crisis related data from social media.

2.2 Text Classification

Social media data is not *homogeneous* in nature. The context, that determines the relevancy of a post with a topic, might be very scattered on social media platforms. In addition, there is a substantial amount of noise on social media‡, which impacts the quality of data collection and analysis if

*https://www.geonames.org/
†https://lod-cloud.net/
‡https://blog.insightsatlas.com/noise-on-social-media-explained

not filtered out. Thus, the need to explore automatic text categorisation/classification methods is paramount. The classification methods are aimed at categorising the heterogeneous data.

There are two broad classes of classification methods studied: *supervised classification* and *unsupervised classification* (Castillo, 2016). For the *supervised classification*, it is first required to curate a set of manually classified data, using human annotators. This data is often termed as *labelled training data*. This labelled training data is used to train a model, to classify new unseen data into the trained categories. In the *unsupervised classification*, there is no prior labelled training data, instead the related items are determined based on a similarity score/metric. There is also a *semi-supervised* classification approach, where some of the available data is labeled but majority of it is unlabeled, and a combination of *supervised* and *unsupervised* techniques are used.

2.2.1 Supervised Classification

Most prominent use cases for supervised learning are: *binary classification*, *multiclass classification*, and *multilabel classification* (Castillo, 2016). *Binary Classification* is classification of data into either of two disjoint classes. *Multiclass Classification* is a scenario when there are three or more disjoint classes the data can be classified into. *Multilabel Classification* is a scenario when there are multiple classes/labels but not necessarily disjoint, therefore the data can simultaneously belong to multiple classes/labels. *Supervised Classification* are based on *supervised learning*. Supervised learning is dependent on some input variables (X) and a mapping function $f(x)$ that returns an output variable (Y).

$$Y = f(X)$$

Supervised learning approximates the mapping function and learns to predict the output variable for a given data. The term *supervised* indicates that the mapping learns to predict by being trained via a

training dataset. The function makes predictions and improves as it iterates through the training data. When the function achieves an acceptable level of performance the learning stops. The outcome of this learning, where the function is finally approximated, is also termed as a *model*. And this process of creating such models is called *machine learning*.

Broadly, the main elements of a supervised classification can be identified as follows:

- Labelled Training Data – Supervised classification methods require data/messages that are already labelled. These labels are the classes/categories the messages belong to. These labels are usually tagged by human annotators (volunteers or experts) on the subject. The size of labelled training data depends on the application or the system to be designed and also on the number of classes the data is classified into. Scenarios where the nature of the data is likely to be of a very diverse nature across the categories, as is often the case in social media data, the large size of labelled training data (from hundreds to thousands) is required. The impact of larger size yielding better results has been highlighted (Matykiewicz & Pestian, 2012). In the literature, training size used in supervised classification for social media data or text has ranged from hundreds (Yin et al., 2015) to thousands (Imran et al., 2014b) or even tens of thousands in some cases (Melville et al., 2013). Another important aspect of creating labelled training data is *sampling*. Ideally, for training a model it is better to have substantial representation from all classes/categories.

- Feature Selection – The input data is converted to a format, termed as *features*, that is suited for the chosen algorithm (function). In a huge data size, the feature space can become very high dimensional, which might impact the run-time of the system. Feature selection is the process of selecting the appropriate or most relevant features for the particular problem. These features are expected to represent the data as a whole, and are the subset of the input features. Some of the examples of feature selections methods are : *Chi squared test*, *information gain*,

pointwise mutual information, and *correlation coefficient scores*. A reference to feature selection methods can be found here (Guyon & Elisseeff, 2003). In our work we make use of some of these metrics for different analysis such as Information Gain and Spearman's rank correlation coefficient .

- Machine Learning Algorithms – There are several supervised classification algorithms for text classification. Among the well practiced algorithms are:
  - Linear Classifiers: Logistic Regression, Naive Bayes Classifier
  - Support Vector Machines
  - Decision Trees
  - Boosted Trees
  - Random Forest
  - Neural Networks

A detailed survey of text classification machine learning algorithms can be referred here (Sebastiani, 2002). For instance, Naive Bayes classifier (Dai et al., 2007) is based on Bayes theorem for conditional probabilities, that quantifies the conditional probability of a *class* variable, given the knowledge about the other set of variables (feature variables). To gather a fundamental understanding about the Bayes theorem, it is used to determine the probability P(A|B), when the probability P(A|B) cannot be determined directly from the data. But, if other prior probabilities such as P(B|A), P(A), and P(B) are evaluated from the data, then P(A|B) can be stated using the Bayes theorem as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Logistic regression is a statistical model, often used in text classification (Genkin et al., 2007), that uses a logistic function to model a binary class variable. However, it can also be used to model non-binary class variable as well. Consider there is a vector $\overline{\Theta} = (\Theta_0, \Theta_1 ... \Theta_d)$ of d+1 parameters. The i^{th} parameter Θ_i is the coefficient to the i^{th} dimension in the feature set. Then for a set of features $\overline{X} = (x_1, x_2 ... x_d)$, the probability of a class variable being, say, +1 or -1 is given as follows:

$$P(C = +1|\overline{X}) = \frac{1}{1 + e^{-(\Theta_0 + \sum_{i=1}^{d} \Theta_i x_i)}}$$

$$P(C = -1|\overline{X}) = \frac{1}{1 + e^{(\Theta_0 + \sum_{i=1}^{d} \Theta_i x_i)}}$$

Support Vector Machines (more widely known as SVM), use separating hyperplanes as the decision boundary between data from different classes. These are naturally tuned for *binary classification*. However there are different methods to tune it into *multiclass classification* (Aggarwal, 2015). In SVMs, the algorithm is optimised by determining the margin that separates the classes. The *maximum margin hyperplane* is assumed to be the optimal solution. Assuming that in the training data there are n data points each mapped with a class $(\overline{X_1}, Y_1) ... (\overline{X_n}, Y_n)$, where $\overline{X_i}$ is a d-dimensional vector representing the number of features and Y_i is the class label, say $Y_i \, \varepsilon \, \{-1, +1\}$. The separating hyperplane can be defined as:

$$\overline{W} \cdot \overline{X} + b = 0$$

$\overline{W}$ is a d-dimensional row vector representing a normal to the separating hyperplane. An optimal solution is such that,

$$\overline{W} \cdot \overline{X_i} + b \geq 0 : \forall_i Y_i = +1$$

$$\overline{W} \cdot \overline{X_i} + b \leq 0 : \forall_i Y_i = -1$$

An elaboration on various machine learning classification algorithms is provided by Aggarwal(2015).

- Evaluation Metrics – Evaluation metrics help in determining how well a classification system has performed in comparison to other approaches. They help in weighing the importance of different aspects in the results and thus influence the choices in the approach. Through metrics we can gauge how efficient a given approach is. Therefore, if we have labelled evaluation data, then a given classification approach can be evaluated simply by having a look at the number of correct and incorrect classified data points. If this is represented in a tabular form, then such a representation is called a *confusion matrix*. As an example look below, where there are two classes *positive* and *negative*. *True positives* are correctly labelled *positive* class data points, *False positives* are those data points which are originally *negative* but labelled as *positive* by the classifier.

  Accuracy is the simplest measurement of effectiveness. It calculates the proportion of correctly classified data points. In the above case, the accurancy can be evaluated as:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

  However, *accuracy* is not a sufficient metric to determine the performance of a classifier across the classes*. In that case, we look for more measures such as *Precision* and *Recall*. These metrics

*Accuracy Paradox, https://en.wikipedia.org/wiki/Accuracy_paradox

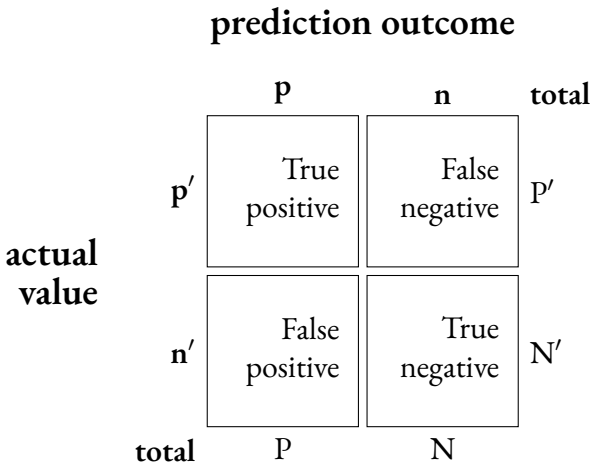

Figure 2.1: Confusion Matrix

can be measured for individual classes, for instance *Precision* for the positive class will be:

$$P_p = \frac{TruePositive}{TruePositive + FalsePositive}$$

This implies that it is a measure of number of correctly classified instances as *positive*, against total number of instances classified as *positive*. Thus, *Precision* evaluates specificity. While, *Recall* evaluates sensitivity, the responsiveness of a classification system. The *Recall* amounts for correctly identified instances belonging to a class, against all the instances that actually belong to that class. *Recall* for the positive class will be:

$$R_p = \frac{TruePositive}{TruePositive + FalseNegative}$$

Another parallel metric that combines both of these measures is F_1 measure or F measure. It is generally represented as a harmonic mean of *Precision* and *Recall*:

$$F_1 = 2\frac{PR}{P+R}$$

In practice, there are variations of this harmonic mean.

$$F_\beta = (1+\beta^2)\frac{PR}{\beta^2 P+R}$$

Here, β is a trade-off variable that can weigh precision over recall and vice-versa as required. There are other techniques to assess the performance of a classifier, such as the *Receiver Operating Characteristic* (ROC). The ROC curve is a plot of *true positive rate* (also a Recall) against *false positive rate*. It can provide a visualisation based analysis and a way to select optimal models. A detail insight into such evaluation metrics and a comparative study can be made here (Powers, 2011).

2.2.2 Unsupervised Classification

Unsupervised learning is an approach where the data is not labelled or categorised. The algorithm attempts to discover structures or patterns in data and categorises the data based on theme. In text classification, *clustering* is a well accepted method based on *unsupervised learning*. Clustering algorithms determine the similar data points (texts in text classification problems), based on certain similarity functions. *Similarity functions* quantify the similarity between two data points vectors. A prominently used similarity function is the *Cosine Similarity* (Baeza-Yates et al., 2011). The outcome of a clustering algorithm is the number of clusters in which the existing data points, within each cluster, are supposed to be similar or belonging to the same class. There can be two scenarios, one in which clusters are supposed to be disjoint, called *hard clustering*, and another in which the classes can be overlapping, called *soft clustering*. Some of the widely used clustering algorithms are described below.

- The *k*-means clustering – This method partitions the observations (data points) into *k* clusters, where each observation belongs to a cluster with a *mean* that is nearest to it. The approach operates in two steps: (i) assign an observation to a cluster which has the least Euclidean distance of its mean, also the centroid of the cluster, with the observation. (ii) re-calculate the new mean of each cluster with the new observations in the cluster. This process can be repeated until a fixed number of iterations or till the reassignment stops changing the centroids any further. The algorithm can be intialised by a random allocation of the initial centroids. This sort of approach where each data point belongs to any one of the clusters, is an example of hard clustering. A comparative study on intitialisation methods for *k*-means has been done (Celebi et al., 2013). There are alternatives to *k*-mean, and can be referred to (Hamerly & Elkan, 2002; Zaki et al., 2014).

- Topic Modeling Methods– These are statistical models for discovering abstract topics in a collection of data (documents). It assumes that each topic is represented by a probabilistic distribution of multiple words, and each document is represented by a probabilistic distribution of topics. This is a type of soft clustering. Some of the well known topic modelling approaches are Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Direchlet Allocation (LDA) (Blei et al., 2003). LDA is a generative probabilistic model, of a collection of documents and words. These approaches do not guarantee a semantically cohesive topics or topics built out of knowledge in a certain domain, but the topics are built on the word-document distribution which is unknown in a large scale data. Hence, the use of the term *latent* in the name of the methods, since these distributions are calculated using probabilistic methods such as Expectation propagation (Minka, 2001; Minka & Lafferty, 2002) or Gibbs Sampling (Griffiths & Steyvers, 2004).

2.2.3 Semi-supervised Classification

Getting labeled data is not trivial, as it is expensive and time consuming, while unlabeled data is often available in large quantities. Semi-supervised learning is an approach to make use of unlabeled data to improve the accuracy of the learning algorithms. The unlabeled data can be used to determine the low-dimensional structure of the data and also can be used to estimate the joint probability distribution of features (Aggarwal, 2015). Primarily, there are two types of techniques for semi-supervised learning: a) *meta-algorithms*- use the existing classification algorithm for enhancing the classification accuracy with unlabeled data. In this approach, we use the limited labeled data to classify the unlabeled data and use the most confidently labeled instances and add to the labeled data for re-training the classifier, b) make use of modified variations of some specific classifiers such as Bayes classifier using EM (expectation–maximization). In such approaches, the classifier is trained on the available labeled data, and then make predictions on the unlabelled data. Next, retrain the classifier and then determine the total likelihood of the model. This is repeated until the total likelihood of the model stops decreasing. A detailed study of the approach can be referred to (Aggarwal, 2015).

So far we have provided an overview of natural language processing methods, semantics via knowledge bases, and text classification approaches. Many of the related works discussed in the subsequent section have used such techniques. We will be using many of these approaches to propose our methodology and perform experiments. Next, we will cover the related work in the literature that covers different works in the areas of crisis data processing.

2.3 Dimensions of Social Media Data During Crises

During disasters, affected people turn to platforms such as Twitter. But given the enormous data people often do not know what kind of information they can expect to find on such platforms. Different works have studied and analysed the nature of the data that gets generated over Twitter during

crises situations. Vieweg (Vieweg, 2012) demonstrated that social media posts can provide situational awareness, that enhance the knowledge about a growing situation, which was also backed up by other studies (Imran et al., 2014a; Olteanu et al., 2014; Olteanu et al., 2015; Starbird et al., 2010).

Olteanu and colleagues (Olteanu et al., 2014) generated a lexicon of terms that usually appear in tweets during disasters. They gathered the data for six different crisis events based on keywords-based samples and location-based samples. Further, the tweets were crowd-sourced for their relevancy with the disaster scenario. The lexicon was created by annotating the unigrams and bigrams from the *positive* (crisis related) class, based on their likelihood to occur in disaster situations, and which tokens are specific to a given situation. The terms were further weighted based on their frequency across different crises. This work was done in view of locating useful information on Twitter during disaster events.

In another work, Olteanu and colleagues (Olteanu et al., 2015) analysed 26 different crisis events between 2012 and 2013. It was a very comprehensive analysis of nearly 25,000 tweets across these crisis events. They determined the type of *information types* and *informative* content occurring in the social media posts. The study revealed that there are various types of sources that post the content, many of them are *eye-witnesses*. Study also revealed that the *crisis related* content did contain *informative* content and was spread across categories such as *affected individuals, infrastructure, donations, caution & advice, sympathy*. Such categories were found to be prevalent across all the crisis events.

There are several studies that have shown the presence of valuable information in social media data in the course of crisis situations. Some of the examples quoted from the literature, that give an impression of such critical information, are hereby shown:

- "OMG! The fire seems out of control: It's running down the hills!" (Bush fire, France, 2009) (De Longueville et al., 2009).

- "Red River at East Grand Forks is 48.70 feet, +20.7 feet of flood stage, -5.65 feet of 1997 crest.

#flood09" (Red River Flood, 2010) (Starbird et al., 2010)

- "Anyone know of volunteer opportunities for hurricane Sandy? Would like to try and help in any way possible." (Hurricane Sandy, 2013) (Purohit et al., 2014)
- "My moms backyard in Hatteras. That dock is usually about 3 feet above water." (Hurricane Sandy 2013) (Leavitt & Clark, 2014)
- "Sirens going off now!! Take cover...be safe." (Moore Tornado, 2013) (Blanford et al., 2014)

These are only some of the cases that represent how social media posts can imbibe valuable pieces of information. Vieweg (Vieweg, 2012) analysed tweets from four different crisis events viz. 2009 Oklahoma Fires, 2009 Red River Floods, 2010 Red River Floods, and 2010 Haiti Earthquake. The study analysed what proportion of the sample data was *'off-topic'*, *'on topic and relevant to situational awareness'*, and *'on topic but not relevant to situational awareness'*. The data was further annotated for information types, and each tweet was categorised into one or more of following labels: *social environment*, *built environment*, and *physical environment*. These categories are very broad, and could cover for a large range of crisis impact on humans and their response. Thus yielding a wide spectrum of data. These environments have been defined as follows:

- *Social environment* is defined as any sort of human action/reaction during emergency event.
- *Built environment* is defined as information corresponding to civic and infrastructure in relation to the emergency situation.
- *Physical environment* relates to information about hazard, weather, and environment etc.

The information type was further subcategorised into 24 categories, which were at a more granular level for example *rescue*, *response*, *offer of help*, *missing*, etc. Across the four events, among the *related and contributing to situational awareness* labelled tweets, the top subcategories were *status-hazard*,

advice-info, *preparation*, *damage*, and *response*. These subcategories amounted for 37%-53% of *related and contributing to situational awareness* labelled tweets.

Two of the works (Olteanu et al., 2015; Imran et al., 2015) generated the subcategories of information types based on various categories of information identified across different crisis events analysed in several related research works. The resulting categories of information types and corresponding works are shown in table 2.3 below.

While in our approaches, proposed in this thesis, we have only focused on the content of the social media posts, some of the works also analysed the author's profile of the posts, who are the actual source of the broadcasted information. These works have categorised these sources based on their profiles, analysed by human annotators. The sources sharing information on social media during such events have been identified as *eye-witnesses* (Bruns et al., 2011; Olteanu et al., 2015), *government agencies* (Olteanu et al., 2014; Metaxas & Mustafaraj, 2013), and *NGO's* (De Choudhury et al., 2012; Thomson et al., 2012) among other categories. In two separate works (Olteanu et al., 2015; Imran et al.), a compilation of categories of *sources* explored across various works, as also shown in table 2.4 below, has been provided.

These information and the source categories shown in tables 2.3 and 2.4 highlight the fact that in the course of crisis events, there is an appropriate participation of public/sources that channel the relevant information on social media platforms. There are different information needs for different types of stakeholders (e.g. relief seeker, relief provider, fund raisers, civic authorities, etc.). Thus, justifying the need to explore the methods to identify crisis related information on social media.

Table 2.3: Categories and dimensions of crisis data in related works

Categories	Related Categories from other works
Affected Individuals	medical emergency, trapped people (Caragea et al., 2011); casualties, people missing & found (Imran et al., 2013a); self reporting (Acar & Muraki, 2011); fatality, injuries, missing people (Vieweg, 2012); missing people (Qu et al., 2011);
Infrastructure & Utilities	damage (Imran et al., 2013a); environment reports (Acar & Muraki, 2011); built environment (Vieweg, 2012); damage, fire & police services (Hughes et al., 2014); hospital services, sanitation, collapse structure (Caragea et al., 2011); road & traffic conditions (Truelove et al., 2015);
Donation & Volunteering	funds, goods, services (Imran et al., 2013a); donations and volunteering (Olteanu et al., 2014); help request, relief offer, relief coordination (Qu et al., 2011); relief and resource donations (Hughes et al., 2014) ; fundraising (Bruns et al., 2011; Shaw et al., 2013); shelter and food (Caragea et al., 2011); volunteer information (Vieweg et al., 2010);
Caution & Advice	caution and advice (Imran et al., 2013a); warnings (Acar & Muraki, 2011); advice, preparations (Olteanu et al., 2014; Olteanu et al., 2015); advice, warning, caution (Vieweg et al., 2010); tips (Leavitt & Clark, 2014); preparedness (Wukich & Mergel, 2015); advice and instructions (Shaw et al., 2013);
Sympathy & Support	condolences (Acar & Muraki, 2011); prayers (Olteanu et al., 2014; Olteanu et al., 2015); emotional support (Qu et al., 2011); gratitude and thanks (Shaw et al., 2013);
Other Useful Information	smoke and ash (Truelove et al., 2015); emergency location/fireline, visibility (Vieweg et al., 2010); other informative posts (Olteanu et al., 2014; Olteanu et al., 2015);

Table 2.4: Source Categorisation

Categories	Related Categories from other works
Eyewitness	citizen reporters & community (Metaxas & Mustafaraj, 2013; Olteanu et al., 2015); eyewitness (Bruns et al., 2011; Diakopoulos et al., 2012; Kumar et al., 2013); locals (Starbird et al., 2012; Vieweg et al., 2010); direct reporting (Shaw et al., 2013; Truelove et al., 2015);
Government	news org. & authority (Metaxas & Mustafaraj, 2013); govt/administration (Olteanu et al., 2014; Olteanu et al., 2015); police & services (Hughes et al., 2014; Denef et al., 2013; Bruns et al., 2011); public inst. & agencies (Starbird et al., 2010; Thomson et al., 2012);
NGO's	non-profit org (De Choudhury et al., 2012; Thomson et al., 2012); non-govt org (Olteanu et al., 2014; Olteanu et al., 2015);
Business	commercial (De Choudhury et al., 2012); enterprise (Thomson et al., 2012); for-profit org (Olteanu et al., 2014);
Media	news org (Metaxas & Mustafaraj, 2013; Olteanu et al., 2014); journalist, media (De Choudhury et al., 2012; Diakopoulos et al., 2012); professional news (Leavitt & Clark, 2014; Olteanu et al., 2015); alternate media, freelancers (Thomson et al., 2012); blogs, news-crawller bots (Starbird et al., 2010);
Others	sympathizers (Kumar et al., 2013); distant witness (Carvin, 2012); remote crowd (Starbird et al., 2012);

2.4 Crises Related Data Identification

Crisis data identification has been approached from more than one perspective. Often this perspective is dependant on the definition of the problem scope. Some approaches focus on identifying individual posts from a stream of data through supervised classification, while some others focus on clustering crises related posts based on certain criteria. A few other approaches begin by looking at when a crisis *event* evolves. These approaches are often termed as *Event Detection*. And the process of tracking the *events* and how they unfold over time is known as *Topic Detection and Tracking*. *Topic detection* includes aspects such as new topic detection, new event detection (first story detection), and topic tracking. Before looking into various works on crises data identification/classification, we will briefly cover literature on event detection from social media data.

2.4.1 Event Detection and Tracking

An event is said to be an occurrence of anything significant associated with specific time and location (Brants et al., 2003). On social media platforms, due to online presence of the masses, the occurrence of an event has also been defined by an increase in the volume of messages around a particular topic (Dou et al., 2012). Events have been categorised as *new event*, *specified event*, *unspecified event*, and *small scale event* (Atefeh & Khreich, 2015; Castillo, 2016). A *new event* is not similar to any of the earlier noted events. *Specific events* are predetermined type which can be monitored. *Unspecific events* are any events that are detected in the incoming data streams. *Small scale events* are generally those that do not generate too much traction for a particular situation, such as crisis events that last for a long time may include sub-events of smaller scale or similar independent events.

New Event Detection, also termed as First Story Detection (FSD), is a sub-task within Topic Detection and Tracking (TDT) (Allan, 2002). Event detection in TDT was traditionally meant for the newswire data, where each new topic was matched with the previous entries. The voluminous and

streaming nature of social media platforms such as Twitter warrant the usage of streaming algorithms. The streaming algorithm is a data processing model where the incoming data is chronologically arranged and is processed in bounded space and time as each new entry arrives (Muthukrishnan et al., 2005). Petrovic and colleagues (Petrović et al., 2010) applied the first story detection methodology, along with a clustering approach, on twitter data to identify new events. Becker and colleagues (Becker et al., 2011) also exploited clustering methods for identifying real-world events. They created clusters of related tweets and further classified a cluster as event or non-event. They extracted different types of features such as temporal (messages posted in an hour are used to create clusters), social (clusters refined using user interactions- retweets and replies), and topical features. For any *new event* they measure the cosine similarity between the new message and each cluster. They hypothesised that a high percentage of retweets and replies do not indicate an event, and also that events are built around a central topic, while the non-events clusters are formed around terms which do not form or reflect a central theme (e.g. *work, sleep, monday* etc). Phuvipadawat (Phuvipadawat & Murata, 2010) proposed grouping and ranking the messages collected via search queries (e.g. #breakingnews and/or #breaking news). The messages similar to each other are grouped together forming a cluster of news articles for a particular story. Message similarity was measured using TF-IDF, weight of nouns, and hashtags.

Another basic approach is the word frequency based method to detect events, when there is a rapid increase in the frequency of a single-word or multi-word tokens. The periodic counters of the number of messages are maintained, and as soon as the count of messages in a particular periodic window increases above a threshold, an *event* is said to be observed. The frequency based analysis can be extended to other activities which can reflect a sudden change in the masses' behaviour, for example web traffic. Osborne and colleagues (Osborne et al., 2012) took the previous approach (Petrović et al., 2010) as a baseline and enhanced it with considering the traffic on the relevant Wikipedia* pages in the same time intervals. They termed their approach as *multi-stream FSD*. However, one potential limitation

*https://en.wikipedia.org/wiki/Main_Page

of this approach is the dependency on web page traffic on third party platforms such as Wikipedia. As the authors themselves point out that Wikipedia lags behind Twitter, in terms of activity, by a few hours and hence it might not be suited for a real time event detection. Also, such approaches are aimed at identifying broad events, rather than identifying/classifying individual text documents into some classes/labels.

Another related work is a system *TwitInfo*, by Marcus and colleagues (Marcus et al., 2011), that collects posts based on an input keyword such as 'earthquake'. The system kept track of frequency of tweets per minute, and reported a potential event when the frequency in a particular time window exceeded the average frequency by two standard deviations. In a multi-word frequency based approach, a system *TwitterMonitor* proposed by Mathioudakis and Koudas (Mathioudakis & Koudas, 2010), identifies events by first exploring the rise in frequency of individual words, and then further grouping them together based on co-occurrence (in same tweets). Some of the variations of such an approach exploit multiple hashtags from the tweets (Corley et al., 2013). Another system, *Twevent* (Li et al., 2012a) relies on determining frequency of tweets which contain data segments, which are generated from segmenting the text into unigrams or bi-grams and extending them using Microsoft Web N-Gram service*. An expected frequency of segments is evaluated using a Gaussian distribution model†. The segments for which the actual frequency exceeds the expected frequency, they are termed as *bursty segments*. An obvious limitation of these approaches is that they are bounded by frequency threshold, which curbs the applicability of such systems in scenarios where the crisis related information are below the threshold and/or not carrying relevant vocabulary. Also, these approaches do not take into consideration different types of events (crisis) and the content language, which we focus on in this thesis.

The multi-word frequency can further be extended by generating graphs where nodes are words

*https://www.microsoft.com/en-us/research/project/web-n-gram-services/

†https://en.wikipedia.org/wiki/Normal_distribution

or phrases and edges indicate weights cross-correlation between different nodes. Further, these graphs can be segmented and clusters of nodes can be created (Sayyadi et al., 2009). Weng and Lee (Weng & Lee, 2011) proposed a system *EDCOW*, which computes the subgraphs from the cross-correlation graph, and label a subgraph as an event when there is a high cross-correlation between the nodes (which are the words). Interestingly, the cross-correlation graph is built on the criteria of words exhibiting a similar *burst pattern*, i.e., similar frequency pattern. This system focused on events from sports, music, politics etc. A similar burst detection approach was used to detect earthquakes (Robinson et al., 2013), where the frequency of posts was monitored for search queries such as '#earthquake' and '#eqnz'. We have already highlighted the difference between frequency based methods and the approaches we have adopted in the previous paragraph while comparing with the other work (Li et al., 2012a).

From the event detection perspective, Twitter has also been considered as a source of sensors, where the users are social sensors. Sakaki and colleagues (Sakaki et al., 2010), used Twitter social sensors (users) to detect earthquake events. They collected the tweets and performed semantic analysis for phrases such as *earthquake, shaking, now it is shaking*. They also used classification approaches to classify them as *positive* or *negative* class, i.e., they were either related to *earthquake* event or not. A potential limitation of this work lies in the assumption that people may share relevant information in only a certain variations of the text, and does not consider semantics at a more conceptual level. However, this is an example of *specific event detection*. Another domain specific event detection system, Twitter-based Event Detection and Analysis System-*TEDAS* was proposed by Li and colleagues (Li et al., 2012b). The system specifically detected crime and disaster events. *TEDAS* was partially a rule based system which crawled over tweets based on certain rules, such as specific keywords and hashtags. Next, the tweets are classified using a supervised learning. Within the event detection approaches, these works focus on crisis specific data, which either focused on specific crisis events (earthquake) or vocabulary (keywords and hashtags), thereby not scaling the applicability of the system to multiple

crisis type and multilingual crisis data. Table 2.5 shows a comparison of various works with regards to the research scope of this thesis.

Some of the works focused on extracting events in the form of entities, dates, etc. Ritter (Ritter et al., 2012) developed a system *TwiCal* to extract multi-type events relating to sports, politics, music release, from Twitter and generate open-domain calendar. They used an in-domain trained entity tagger (Ritter et al., 2011), instead of using Stanford Tagger. The system extracted entities, dates, event phrases from the Twitter data. The use of Natural Language Processing techniques has been exploited in more works to perform event detection. Elloumi and colleagues (Elloumi et al., 2013) designed a two-step model for performing event detection. The first step performs relation extraction and creates binary relations between entities in the text. The second step arranges these relations in a template, which can define an event. Popescu and colleagues (Popescu & Pennacchiotti, 2010) applied supervised machine, using Gradient Boosted Decision Trees (Friedman, 2001), learning to detect controversial events. For this they used a *controvery lexicon* from Wikipedia, *bad words lexicon*, and an *English dictionary*. The *English dictionary* comprised of 100k part-of-speech tagged English words, which was trained over Wall Street Journal and Brown Corpora*. The work reflected optimistic results, however it was not catered for controversial events from diverse domains. Alsaedi and colleagues (Alsaedi et al., 2016a) proposed a two stage classification system for identifying real-world events from Twitter in Arabic language. First stage was a classification task where the data is categorised into events or non-events. The second stage was a clustering stage to cluster the data into multiple potential events. For supervised classification task, a sample of 5000 Arabic tweets was manually annotated into categories *event* and *non-event*. While the work focused on Arabic data, it only demonstrated the event detection problem from a single language perspective. They used Arabic language specific stemmer for pre-processing the data.

So far, we have covered the segment of the literature where the focused domain of information on

*https://www.sketchengine.eu/brown-corpus/

social media is treated as an event, and the various approaches to detect the events. Next, we survey the works which have focused on identifying crises oriented information from social media.

2.4.2 Social Media Data Processing in Crisis Situations

When it comes to crises events (natural or man made disasters), as we have seen that social media emits information which can be valuable imminently to various stakeholders such as decision making bodies, first responders, impacted people, and even to the general public. But as we are aware of the potential challenges that we face in extracting, filtering (Gao et al., 2011), classifying, and/or ranking crisis related content from social media, we review various methods researched specifically to process such information. As a reminder, in Table 2.5 we compare various works with respect to the research questions being explored in this thesis.

Several systems have been proposed to extract crisis relevant information from social media. ESA- (*Emergency Situation Awareness*) system (Yin et al., 2012; Power et al., 2014) was aimed to enhance situational awareness with respect to natural disasters. The system performs this in a series of steps: (i) beginning with a burst detection of tweets; (ii) performing clustering leading to clusters that reflect events; (iii) filtering out tweets that are not high-value, via statistical classifiers using SVM; (iv) geo-tagging each tweet based on the location mentioned in the user profile. The events were geographically bounded to Australia and New Zealand. Lie and colleagues (Li et al., 2012b), as earlier mentioned, developed *TEDAS* which focused on crime and disaster related events on Twitter. The tweets were collected based on predefined keywords. Further, using statistical features such as mentions, hashtags, URL's, the tweets were classified using a supervised learning. The system ranked the tweets by training a function based on content features (if tweet contains certain words), usage features (number of re-tweets and likes), and user features (if it is a verified account, number of followers). To extract the locations they relied on the geographical references in the text, as we briefed the concepts of entities in the section 2.1. Rogstadius and colleagues (Rogstadius et al., 2013) proposed a disaster awareness

system *CrisisTracker*. Using the predefined filters the system collects the tweets and through the locality sensitive hashing techniques clusters them as stories. Jadhav and colleagues (Jadhav et al., 2010) developed *Twitris* by mining semantics of the tweets and considered spatio-temporal parameters of a tweet. In relation to the research questions posed in this thesis, these works do not aim to explore the applicability of proposed approaches in a new crisis type situation or crisis data in an unseen language.

Many other supervised learning approaches have been studied. Karimi and colleagues (Karimi et al., 2013) developed a statistical classifier for classifying crisis related data. They took the data from multiple crisis events and used human annotators to classify them as crisis related or not. Using statistical features such as n-grams, presence of hashtags, number of hashtags, and user mentions, they trained a SVM classifier. However, they analysed and validated their model only using k-fold cross vaidation (Geisser, 1974), which does not evaluate the scenario when the classification model is applied and tested on an entirely unseen data. They reported accuracy of roughly around 60%. Similarly, Stowe and colleagues (Stowe et al., 2016) annotated nearly 8000 tweets for: *sentiment, action, movement, preparation, reporting, information*. They opted for a supervised learning and used SVM for developing the classification model. Features such as whether a tweet is a re-tweet or not, base domain of URL, unigrams from the previous two tweets, and n-grams were used. They also augmented each n-gram with its corresponding part of speech and the named entity. Furthermore, a word embedding of all words was also augmented as a feature set. The embedding was generated by training a Word2Vec model (Mikolov et al., 2013) on nearly 22 million tweets from Hurricane Sandy. The model was validated on a 5-fold cross validation technique (with a reported F1 score of 0.72 for classifying relevance and a low F1 score ranging between 0.36-0.52 for individual categories), and thus did not demonstrate the validity of their model on a new type of crisis event.

Zhang and Vucetic (Zhang & Vucetic, 2016) proposed a semi-supervised approach for classifying crisis related data. A labelled corpus was used to train a logistic regression classifier, and an unlabelled corpus was used to create clusters as features (most related words for each word). The use of proba-

Table 2.5: A comparison across various works with respect to the problem scope in this thesis

Related Works	Focus	Cross-Linguality	Cross Crisis Type	Method	Semantic Features	Statistical Features	Translation
TwitInfo (Marcus et al., 2011)	Event Detection	No	No	Frequency Burst	No	No	No
Twevent (Li et al., 2012a)	Event Detection	No	No	Frequency Burst	No	Yes	No
EDCOW (Weng & Lee, 2011)	Event Detection	No	No	Clustering, Frequency	No	No	No
(Robinson et al., 2013)	E'quake Events	No	No	Frequency	No	No	No
(Sakaki et al., 2010)	E'quake Events	No	No	Classification	Neighbour words	Yes	No
TEDAS (Li et al., 2012b)	Crime & Disaster	No	No	Classification	No	Yes (content, user profile)	No
ESA (Yin et al., 2012; Power et al., 2014)	Natural Disasters	No	No	Burst, Clustering, Classification	No	Yes	No
CrisisTracker (Rogstadius et al.,2013)	Disaster	No	No	Clustering	No	Yes	No
(Karimi et al., 2013)	Disaster	No	No	Classification	No	Yes	No
(Stowe et al., 2016)	Natural Disaster	No	No	Classification	Yes (Word Embeddings)	Yes	No
(Zhang & Vucetic, 2016)	Natural Disaster	No	No (Only cross crisis)	Clustering, Classification	No	Yes	No
(Imran et al., 2013a; Imran et al., 2013b)	Natural Disaster	No	No	Classification, Extraction	No	Yes	No
(Agarwal et al., 2012)	Factory Fire	No	No	Classification, Extraction	No	Yes	No
(Schulz et al., 2013; Schulz et al., 2015)	Car Crash	No	No	Classification	Yes	Yes	No
STED (Hua et al., 2013)	Crisis, civic	No	No	Classification-semi-supervise	No	Yes	No
Twitcident (Abel et al., 2012)	Fire Events	No	No	Classification	Yes	Yes	No
Tweedr (Ashktorab et al., 2014)	Disaster	No	No	Classification, Clustering, Extraction	Yes (hypernyms)	Yes	No
(Li et al., 2015)	Natural Disaster	No	No (Only cross crisis)	Classification	No	Yes	No
(Imran et al., 2016b)	Natural Disaster	Yes (Only 2 language, lack rigorous evaluation)	Yes (Only 2 types)	Classification	No	Yes	No
(Pedrood & Purohit, 2018)	Natural Disaster	No	Yes (Only 2 events)	Classification	No	Yes	No
(Burel et al.,2017b Burel et al.,2017a)	Natural Disaster	No	No	Classification	Yes (Word Embeddings)	Yes	No
(Alam et al., 2018)	E'quake, Floods	No	Yes	Classification	Yes (Word Embeddings)	No	No
(Lorini et al., 2019)	Floods	Yes	No	Classification	Yes (Word Embeddings)	No	No
(ALRashdi & O'Keefe, 2019)	Crisis	No	No	Classification	Yes (Word Embeddings)	No	No

bilistic sequential models such as Hidden Markov Models, Conditional Markov Models, Conditional Random Fields (CRF), etc, has also been observed in this problem space. The evaluation considered the scenario where the training data was formed of tweets from different events than the test data. Although, the authors did not take into consideration the *type* of a crisis event when defining the training and the test data. The authors claimed that when the number of labeled tweets are less than 100, then their approach is superior to standard supervised classification approach based on bag of words representation. Imran and colleagues (Imran et al., 2013a; Imran et al., 2013b) had applied CRF to extract information from the tweets. They apply a two step process, where in step one they classify tweets using a Naive Bayesian classifier into categories: *infrastructure damage*, *donations*, or *caution and advice*. In the step two, relevant information for infrastructure, damages, or donations is extracted. They used several textual and statistical featuers, such as (i) presence of user mention, URL, hashtag, emoticon, any numeric character; (ii) length of the text; (iii) uni-grams, bi-grams, and part of speech.

Agarwal and colleagues (Agarwal et al., 2012) deployed a four step process to detect *factory fire* and *labour union strikes*: *detection*, *message correlation*, *extraction*, and *event correlation*. They used locality sensitive hashing, supervised classification, and post information extraction. The detection phase reports the messages that indicate occurrence of an event. It is a two step process: rejecting the tweets that follow a certain regular expression and then using a supervised classification and boosting as a next step. For supervised classification both Naive Bayes and SVM's are used. Following features were extracted and used by the classification model: (i) number of occurrences of location, people, organisation, and URL's in the text. (ii) occurrence of digital text (i.e. numbers), and further parsing the data within a range, and using it as a feature. (iii) after stemming and stop word removal, used remaining text as feature. *Locality Sensitive Hashing* is used to determine tweets similarity and tweets are treated as new events in case of similarity being less than 75%. They also customised standard NER tagger to extract out location entities from the text. Also, during the parsed tree traversal, if a

subtree had an article "a" or "the" before the term *factory* or *mill* or *plant*, then all the words between the article and the term would be extracted. As per the authors claim, this resulted in around 76% accuracy. To differentiate between event's locations, context of event was framed as a measure of time and distance. If two events were recorded within 24 hours and within 100 km radius, then they were treated as the same event.

Much like the above (Agarwal et al., 2012), there are more works which have focused on detecting events that might relatively be assumed as small scale events. Schulz and colleagues (Schulz et al., 2013) worked on proposing a real-time architecture for detecting car crashes from microblogs. The approach is a supervised learning which incorporates text classification and semantic web. For text processing, they relied on resolving the abbreviations via a dictionary compiled from an online slang resource*. As a pre-processing step they also focused on fixing spelling errors via Google Spellchecking API. The classifier was trained using several statistical features such as number of special characters (e.g. "!", "?"), capitalised characters, mention of spatial and temporal terms, and also used Linked Open Data features from *FeGeLOD* (Paulheim & Fürnkranz, 2012) to extract *types* and *categories* for instance of dbpedia:ENTITY. For identifying the temporal references in the text, the authors applied Heidel Framework (Strötgen & Gertz, 2013), which resolved the text into date and time. This work is close to the way we have approached the crisis classification problem. However, the approach did not target the associated problems of the crisis data with regards to multiple languages or a diverse range of crises, as we define in our research scope. Yet, this certainly makes use of Linked Open Data and some of the semantic properties. The authors report an accuracy of 89%. In another following up work, Schulz and colleagues (Schulz et al., 2015) proposed an approach to extract properties that define an event viz. *location*, *time*, and *type* while detecting small scale crises events such as *fires* and *car crashes*. Following a supervised learning, they label the tweets for their crisis relevance. Along with many of the statistical features (as highlighted in their previous work), they also use abstract features using Se-

*www.noslang.com

mantic Abstraction (an updated version of the work is - Schulz et al., 2017). Semantic Abstraction resulted in identifying all location entities and replacing them with a token "LOC". Further, the coordinates for locations were determined and polygons were drawn to narrow down the location area. Rule based clusters were formed, where the rules were described as a triple - <incident type, radius, time>. Those incidents falling under a particular radius and time interval were formed as one cluster of an event. The authors used datasets representing four classes - fire, car crash, shooting, and NOT Incident Related. It is important to note that most of social media posts are not geo-tagged, hence in order to determine locations for most of the posts, we need to curate methods to determine them from the text, as was demonstrated in this approach.

Hua and colleagues (Hua et al., 2013) proposed a system *STED*, to automatically detect *crises*, *civil unrest*, or *disease outbreak* events from Twitter. This was a semi-supervised approach. First, the labels are generated from public media sources by extracting *named entities* and *action words (verbs)* from the news description. Next, the labels are propagated to the tweets, by determining if a given tweet contains at least one of *named entities* or *action words*. Graph partitioning methods are used to create an event-related group of words and generating clusters of tweets. Auto-correlation between the words was used to filter out non-important words in the clusters. Support Vector Machine (SVM) is used for supervised classification. TF-IDF (term frequency-inverse document frequency) was calculated for each word and a threshold was used to only keep top words as the features. The location was extracted from the geo-tagged tweets, and the tweets which contained similar terms and hashtags were assumed to be from a similar location. The authors claim to have achieved 72% in precision and 74% in recall. However, the work only focused on events from Latin America and did not elaborate in detail about the labelled data (number of tweets) used from different events for classification.

Use of semantic web techniques has also been observed. Abel and colleagues (Abel et al., 2012) developed a system *Twitcident* which focused specifically on detecting fire related incidents from tweets. The system serves analysing, filtering, and searching of small scale incidents (which do not attract sig-

nificant web traction). The tweets were annotated using DBpedia Spotlight* (Mendes et al., 2011). The extracted concepts are represented as attribute-value pair such as *location, dbpedia:Austin Texas*. Similarly, various concepts for different categories are extracted. The tweets are classified as reporting about casualties, damages or risks. They are classified via hand-crafted rules which operate both on words in the text and attribute-value pairs. A recall of 0.61 is reported. While the usage of semantic features is demonstrated, the performance of the semantic features is not evaluated over the non-semantic approaches.

Ashktorab and colleagues (Ashktorab et al., 2014) proposed a system *TWEEDR*, which extracts disaster relevant information for relief workers. The system worked in three phases: classification, clustering, and extraction. The classification phase classified a tweet as disaster damage or casualty information. Clustering phase merged the similar tweets, and in the extraction phase the system extracted the phrases or tokens which contained particular information about different aspects of infrastructure. The authors collected data for twelve different crises events, by querying in two parts: (i) keyword queries based on terms and hashtags; (ii) geographical queries by bounding box coordinates around the event location. The authors experimented with and compared a number of classification algorithms such as k-nearest neighbours, decision trees, Naive Bayes, and Logistic Regression. The data was converted to standard unigram feature vector. In order to extract the nuggets of information, the authors employed Conditional Random Fields (CRF), and inspected for capitalisation, pluralisation of the word, whether the term is numeric in nature, and if the term belongs to a transportation lexicon. They also checked for the term's hypernyms from WordNet, and the part of speech. The authors reported a low average *precision* of 0.49. Evaluation was performed using a 10-fold cross validation approach on the entire data, and did not explore the cases when the system is applied on a new *type* of crisis.

Stowe and colleagues (Stowe et al., 2018) focused on classifying user-evacuation behaviour dur-

*DBpedia Spotlight, https://www.dbpedia-spotlight.org/demo/

ing hurricane events. The authors employed both SVM based and Convolutional Neural Network (CNN) based approaches to predict relevancy of tweets users produce during a certain event. They used temporal information, spatial information, and combined with word embeddings for generating vector representation of user behaviour. The evaluation was performed using a 10-fold cross validation approach and the authors observed that the deep learning approach show lower results than other classification approaches such as SVM and Naive Bayes. The authors explain that this was due to small size of the dataset. This work only focused on one type of event, i.e., hurricane, and did not explore the aspect of models being applicable to other forms of crises.

ALRashdi and O'Keefe (ALRashdi & O'Keefe, 2019) studied the application of different deep learning architectures with word embeddings to classify different types of crisis related content. The authors used CrisisNLP dataset (Imran et al., 2016a) for the study. The authors used two types of word embeddings- GloVe (Pennington et al., 2014) and crisis embeddings which are generated from almost 50k disaster related tweets. The dataset had labelled tweets from different crisis events and in English. The tweets were labelled across different category: *missing*, *infrastructure*, *sympathy*, *donation*, and *other information*. The training set across different classes ranged between 700-1500. The authors reported F1-score on the test data, which ranged around 59%-61%. This work did not explore the applicability of classification models based on variations in the crises types and/or languages. The evaluations did not consider whether or not the test data originated from the same crisis event or a new one.

Cross Crisis Adaptation

While most of the above works have explored methods to efficiently classify crises data from social media, many of them have not projected the problem of the applicability of the classification model on new types of crisis events, i.e., how effective a model is when it is tried on a new crisis event. The problem of developing crisis data classification models, and applying them to data from a new event

has been observed as a domain adaptation problem in some literature (Li et al., 2015; Imran et al., 2013b; Imran et al., 2016b; Li et al., 2017; Pedrood & Purohit, 2018; Li et al., 2018a). While domain adaptation has widely been seen in the field of *sentiment analysis* (Peddinti & Chintalapoodi, 2011; Tan et al., 2009; Blitzer et al., 2007), in crisis data classification it has been viewed from the perspective of applying the models to events from new languages and unseen events. Li and colleagues (Li et al., 2018a; Li et al., 2017) used a supervised learning model (Naive Bayes classifier). They used a popular crowd sourced labelled crisis dataset *CrisisLexT6* (Olteanu et al., 2014) and train-tested the events in pairs (based on timelines of the two events), i.e., the test event was not seen in the training data. A bag of word representation was used to represent the tweets as vectors. They adopted an iterative expectation-maximisation approach (Li et al., 2015), where the classifier iteratively learns from the target data by classifying a part of it and re-learning from it. However, the scope of the study did not take into consideration the similarity of different types of events (e.g., hurricanes and floods can have similarities in the social data given the nature of impact of events on people) and the languages. In another similar work by Li and colleagues (Li et al., 2015), they adopted a nearly similar approach on a much smaller data of two events: Hurricane Sandy and Boston Marathon. A Naive Bayes classifier was used to build the classification model.

Imran and colleagues (Imran et al., 2016b), analysed the classification performance when it was trained and tested on events from two types of crises events. While the authors collected the data from AIDR platform (Imran et al., 2014b) and CrisisLexT26 (Olteanu et al., 2015) dataset, the study was narrowed down to *earthquakes* and *floods*. Standard statistical features such as uni-grams and bi-grams were used. They demonstrated that a classifier, built on Random Forest algorithm, trained on Italian is more likely to perform well on test events from Spanish, instead of English. However, the scope was limited to two types of events which majorly originated in Italian and Spanish languages, and thus lacked a rigorous cross-crisis and cross-lingual evaluation. Pedrood and Purohit (Pedrood & Purohit, 2018) attempted a transfer learning approach which learns from one type of crisis event and classi-

fies a new type. They curated datasets from two events: Hurricane Sandy 2012 and Supertyphoon Yolanda 2013. They used a sparse coding model and compared it with bag-of-word representation as the feature representation. The scope was however limited to hurricanes and typhoons which might often result in similar impacts of flooding. In recent times, more popular deep learning methods have also been applied to such problems (Burel et al., 2017a; Burel et al., 2017b). Burel and colleagues (Burel et al., 2017b) adopted Dual-CNN (convolutional neural network) to develop a crisis relatedness classification model. This model was unique as it included two layers of word embedding, one via the Google's Word2Vec training model (trained on the data itself) (Mikolov et al., 2013) and the other was the semantic concepts layer. The semantic concept layer composed of entities extracted via Alchemy API * and their corresponding semantic sub-types, such as *location, politician, non-profit organisation*, extracted via knowledge bases (DBpedia, Freebase). They also used CrisisLexT26 dataset (Olteanu et al., 2015) to train and evaluate the model. In a following up work, Burel and colleagues (Burel et al., 2017a) adopted a nearly similar approach to identify different categories of information in the crisis data. However, both works (Burel et al., 2017b; Burel et al., 2017a) do not consider determining the adaptability of the model to unseen types of crisis or if the new crisis data was in an entirely new language. The applicability of neural networks on text classification and their domain adaptability was earlier demonstrated by Nguyen and colleagues (Nguyen & Grishman, 2015) on data from newswire, usenet, telephone conversations, and weblogs. Similarly, Alam and colleagues (Alam et al., 2018) demonstrated domain adaptation of crisis data classification models, by training and testing them on *earthquake* and *flood* events in iteration. They proposed a CNN architecture with word embedding to train a domain adaptive classifier. The word embeddings were generated from crisis data. They reported F1-score in the range of 59%-65%. Since there were only two events in the study, i.e., one *earthquake* and one *flood*, the study lacked a comprehensive analysis of the approach.

*Alchemy API, http://www.ibm.com/watson/alchemy-api.html

Multilingual Adaptation

Classifier adaptation (or domain adaptation) is a problem not only from the domain perspective, but also from the language or NLP perspective, where the applicability of a model trained on a certain language is determined in another language too. Acquiring training data in a new language each time is not a trivial task. This problem has widely been realised in research fields such as *sentiment analysis* (Ahmad et al., 2007; Araujo et al., 2016; Balahur & Turchi, 2014; Can et al., 2018; Dashtipour et al., 2016; Denecke, 2008; Mihalcea et al., 2007). The problem of multi-linguality in sentiment analysis has been addressed in various ways: translating the languages to one language (Araujo et al., 2016; Balahur & Turchi, 2014; Kanayama et al., 2004), weakly supervised models (Deriu et al., 2017), and using the lexical resources (such as SentiWordNet) (Denecke, 2008). In the crisis data scenario, variations in the language form an equally crucial aspect, as the variations in the *type* of crisis events. Crisis situations can occur around the world thereby resulting in data originating in different languages. In order to develop computational models that can identify crisis related content, we would also need to consider their dependency or lack of dependency on a diversity of languages. Imran and colleagues (Imran et al., 2016b) used crisis events from primarily in two languages and created classifiers using statistical features to test the language adaptation of the classifiers. Li and colleagues (Li et al., 2018b) used word embeddings for generalising the crisis data across several crisis events, however they used crisis events only in English, thereby excluding multilingual analysis from their study. Zielinski and colleagues (Zielinski et al., 2012) developed Naïve Bayes classifiers by mixing tweets from multiple languages and used a simple bag of words approach for training the classifier. The accuracy observed in their approach was fairly low across different language datasets. Alsaedi and colleagues (Alsaedi et al., 2016b) proposed a two stage approach to classify events from tweets originating from Arabic geographic locations, but data had a mix of tweets in English and Arabic. They create a bag of word model based on dictionary/lexicons representing words from different topics such as weather, energy,

health, politics etc. The evaluations were performed based on 10-fold cross validation approach, and did not specifically demonstrate a language adaptation aspect of the problem. Lorini and colleagues (Lorini et al., 2019) explored the impact of language agnostic and language aligned word embeddings to create classifiers that identify *floods* related posts from social media data. The experimental setting was a supervised binary classification problem. The authors experimented with SVM, Random Forest, and Convolutional Neural Network (CNN), and observed that performance of CNN was similar to that of SVM and Random Forest. They used GloVe embeddings derived from a tweet corpora as the language agnostic embeddings. The authors also used MUSE embeddings derived from Wikipedia (Conneau et al., 2017). The reported a F-1 score in cross-lingual classification (where the training language and target language were different), which ranged between 0.48-0.70, with an average F-1 score of 0.59. While the data was multilingual in nature, the entire data originated from common type of event (flood), which could have meant a significant overlap in the vocabulary (entities). The work also did not report the performance of models without the embeddings, which limits the judgement while determining the impact of embeddings.

Lo and colleagues (Lo et al., 2016) used multilingual lexicon (in English, Malay) to build a polarity detection approach in Singapore English (*Singlish*) language. The machine translation or any parallel corpus could not be used because it did not exist for Singapore English for detecting polarity in the content. From the text classification point of view, for the cross-lingual sentiment analysis task, Xiao and Guo (Xiao & Guo, 2014) used learning methods, by creating bi-lingual feature matrix between source language and target language. A similar approach of representation learning was observed by Zhou and colleagues (Zhou et al., 2016), where they map the semantic and sentiment correlations between the bilingual text in the same embedding space. However, the semantic correlations are established only by translated counterpart of the text. Duek and Markovitch (Duek & Markovitch, 2018) proposed generating language-independent features from knowledge sources such as Wikipedia to facilitate cross-lingual text classification. This work is in some ways similar to our ap-

proach of extracting broader semantics from the knowledge graph, however they train the classifiers on only language-independent features by relying on an ontology based on knowledge source. Although, this work does present a hierarchy of concept based approach, training the classifier only on those abstract concepts in a problem scenario where the data from Twitter is short in length (thus have limited context) might result in sparse information. In contrast to some of the works mentioned above, the approaches adopted by us consider not only the impact translated version of the data can have, but also explore the role contextual semantics (expanded via knowledge graphs along with retained original information (in text) for maximised context) can have on cross-lingual crisis data classification.

2.4.3 Semantics in text classification

From sections 2.1 and 2.1.1, we understand that *semantics* imply the added knowledge corresponding to the entities in the data. These semantics add context to the information, thereby enhancing the knowledge regarding the data from any given domain. Figure 2.2 shows a conceptual representation of adding the semantics to a text. From the text classification point of view, the supervised or unsupervised machine learning methods rely on the information that is existing in the data. This would imply that by enriching the contextual information in the data, the ability of the machine learning classifiers to classify data into classes will get better. Semantics enhance the chances of implicit or explicit relationships between different words in the data. Finding such relationships can encourage finding words from such classes that can give a more coherent representation of the vocabulary in the data. A coherent vocabulary is more likely to impact the accuracy of the classification algorithms on a data which is, otherwise, very diverse and scattered in its representation. These aspects of semantic knowledge can help us overcome the limitations of a prominently used bag-of-word approach on the actual data itself (Hu et al., 2008), in the machine learning methods. We have seen that knowledge bases are a valuable source to extract semantics for the words (Hu et al., 2008; Wang et al., 2016b).

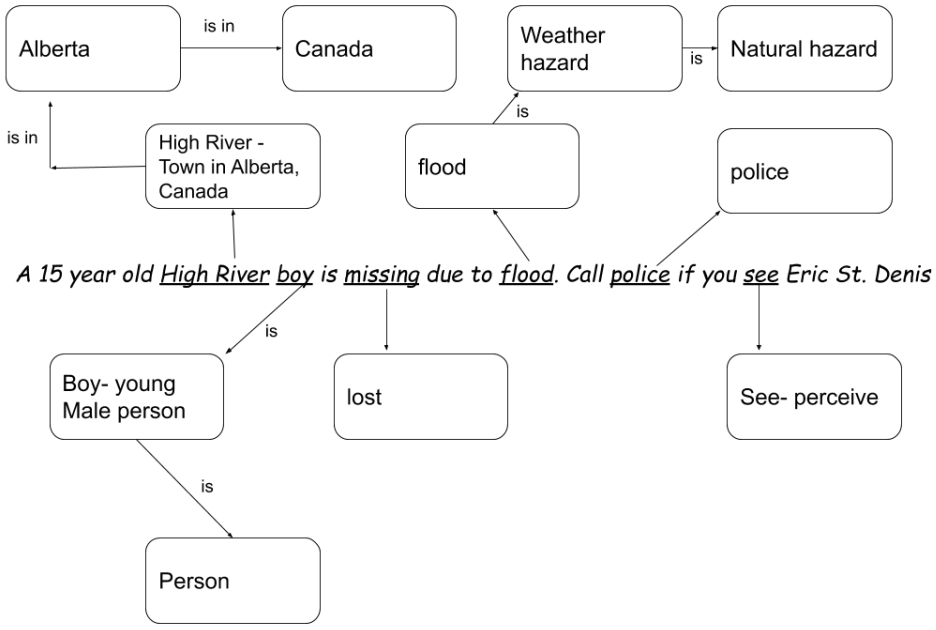

Figure 2.2: Conceptual representation - Semantic expansion of a tweet

Several works have established the phenomenon of semantic similarities between different words by exploiting knowledge bases such as WordNet or Wikipedia (Agirre et al., 2009; Zhang et al., 2011). In one of the earlier works, Siolas and colleagues (Siolas & d'Alché Buc, 2000) proposed a semantic kernel for text classification, for newsgroup database, using SVM classifier. They determined semantically closer concepts within the data based on their *semantic proximity* in WordNet (if present) where the *proximity* is defined as the inverse of the distance between the two words. Thus, it weighted the concepts/terms in the inverted index of the words based on semantic similarity apart from TF-IDF, and was reflected in the feature vector passed to the SVM kernel.

Hu and colleagues (Hu et al., 2008), proposed an enhanced clustering approach for text data by leveraging the semantic knowledge from Wikipedia. They highlighted the limitation of WordNet in terms of its limited coverage and overly simple relationships. They first extracted several semantic re-

lationships from Wikipedia such as synonym, hypernym, and associative relations and developed a framework to enhance the similarity measure for text clustering by exploiting the extracted semantic relationships. In order to develop the semantic thesaurus from Wikipedia, the authors exploited the *redirect* property and extracted anchor texts for synonymy, disambiguation pages for polysemy, extracted "is-a" relationship, based on another method (Ponzetto & Strube, 2007), to determine hypernymy, and derived associative relationships based on content similarity measure between pages and out-link category measure.

Abel and colleagues (Abel et al., 2011) experimented with enriching the tweets semantically and then augmenting them with the news articles. They extracted named entities from the tweets using Open Calais*. The annotations were in the form of DBpedia or Wikipedia URIs. The tweets are linked with news articles after determining the similarity between the tweets and articles based on TF-IDF similarity and URLs in the tweets. They provide a faceted search on such semantically enriched tweets.

Hu and colleagues (Hu et al., 2009) extracted external concepts from a knowledge base and internal concepts from the actual text, to improve the clustering of the shot text. They generate three levels of features: word level, phrase level, and external semantic features. A Solr† index of Wikipedia is created, and for each seed phrase the authors retrieve titles and bold terms (links) from each page returned by querying the index for any phrase query. The external features are filtered by applying heuristics, and also regulated in the total number of extracted features. By comparing their approach against the bag-of-words baseline, they were able to show that extracted semantic features improve the clustering accuracy in the range of 3-10%. They used two datasets, one from Reuters and the other from Google trends. In order to specifically work on short texts, they discarded the texts which contained more than 50 words. Two clustering methods were used: *k*-means, and Expectation Maximization Clustering.

*Open Calais, http://www.opencalais.com/

†Apache Solr, http://lucene.apache.org/solr/

Wang and collagues (Wang et al., 2016a) proposed text as a network classification framework. The text is represented as a heterogeneous information network. The structured and typed information network is generated via semantic expansion using the knowledge bases. In the approach, the authors not only take the extracted entities from the knowledge base (Freebase), but also the path (meta-path) between any two entities of the text in the knowledge base into the consideration. They use Naive Bayes algorithm for classification, and project the probability of the classifier as the product of two separate classification probabilities: one based on bag-of-words of entities, and other on the links generated by the path. In another work, Genc and colleagues (Genc et al., 2011) first map the tweets to corresponding Wikipedia pages, and then compute the distance between the pages to determine the semantic similarity between the tweets. The pages are determined by checking if there is a dedicated page for a word in a given tweet. This results in multiple candidate pages for each tweet. Then a score is calculated for each page, by determining the number of occurrences of words (of the tweet) in the page. The page with highest score gets assigned to the tweet. The similarity between pages is determined by calculating the number of links between the categories associated with any two pages. They also used String Edit Distance and Latent Semantic Analysis as alternative methods to measure the semantically closer tweets. They compared the three approaches by mapping the tweets on a two-dimensional plane by using the multi-dimensional scaling of distance between the tweets. This helped in visualising the clusters based on the three approaches. Using a Discriminant function analysis they measured which technique predicted the category of the tweets better.

In another approach Song and colleagues (Song et al., 2011) undertook a probabilistic approach on top of the knowledge base information to conceptualise the short social media posts (tweets). They used Probase (Wu et al., 2011; Wu et al., 2012) to determine the conceptual attributes and further applied Naive Bayes inference method to find out a more broader concept. For instance, they assumed, if the attributes refer to *population, language,* and *currency*, then there is a high probability of them referring to a subject as *country*. Although with no mention to any specific country. With the ex-

tracted semantic features using the approach, they perform clustering over the tweets using *k*-means clustering. They were able to show that their approach (probabilistic semantic conceptualisation) performed better than traditional bag-of-word statistical methods. A few other similar approaches indicate the usage of knowledge base oriented concepts for cluster labeling (Carmel et al., 2009) or enhancing classification and ranking of short social media posts (Wang et al., 2014).

Tang and colleagues (Tang et al., 2012) attempted to enhance the semantic information for the text by two methods: first augmented the knowledge with the translation of the original text to multiple languages, and then extracted *synonyms* from WordNet for each concept. They also extracted titles and keywords from Wikipedia as additional semantics for the words. They applied this on the data pulled from Facebook and Twitter corresponding to top 30 topics derived from Google Trends*. These topics were used as queries to Twitter and Facebook APIs to collect the data. Two clustering approaches *k*-means and LDA were applied on data sets to compare the baseline features and features generated from their proposed approach.

The TREC Microblog track (2011-15)† has boosted the research in the social media data classification by providing large size corpus (each track has millions of tweets) and hand-annotated subsets of it. Tao and colleagues (Tao et al., 2012), alongside the keyword and tweet syntax features such as hashtags, also exploited entity-based semantic features generated by DBpedia Spotlight to show better results in determining the relevance of a tweet with respect to a query. There are a few detailed surveys covering the usage of semantic techniques in mining the social media data (Bontcheva & Rout, 2014; Ristoski & Paulheim, 2016).

In recent times, word-embedding has become quite popular in the neural network based text classification approaches. Word-embeddings‡ are a distributed representation of words which are likely to have similar meanings or used in the same context. Individual words are represented as vectors formed

*Google Trends, https://trends.google.com/trends/
†TREC Microblog Track, https://trec.nist.gov/data/microblog.html
‡Word-Embedding, https://en.wikipedia.org/wiki/Word_embedding

of such terms (similar context), which are generated from a large corpus. These embeddings are then used as dense and high/low dimensional matrix in neural networks. These embeddings, conceptually, behave as external semantics generated from a large scale corpora. The obtained word embedding are meant to be combined with the original text data in a meaningful representation. They can either be converted to a vector using one-hot encoding or as in more popular approaches multi-layer perceptron or convolutional/recurrent neural networks (Hu et al., 2014). These approaches require a fixed length input, or use aggregation operations such as k-max pooling (Kalchbrenner et al., 2014; Xu et al., 2015) to bring it down the dimensionality for entire input.

Lai and colleagues (Lai et al., 2015) created a recurrent neural network for text classification. In their work, they created a recurrent structure for the text, which is a bi-directional recurrent neural network, to capture the context. Each word was structured with the word on the left of it and on the right of it, ensuring that each word is always structured with its neighbouring terms to establish the immediate context. Each neighbouring word is similarly defined by its neighbouring context. Additionally, they use a pre-trained word embedding which was trained on English and Chinese Wikipedia dumps. Similarly, in a work described earlier, Burel and colleagues (Burel et al., 2017b) used extracted semantic information from knowledge base along with word embedding layer by training on a Word2Vec model, to classify the social media data. Similarly, in some other related works, embeddings have been used to generate feature representations for the clustering of the short text (social media posts) (Xu et al., 2015).

2.5 Summary and Discussion

Social media platforms are now widely considered as a crucial source for mining critical information during crises situations. Distinct works have shown the validity of crucial information being present in social media data. But given the overwhelming amount of data getting generated in short time periods,

it is nearly impossible to manually monitor and record crisis related information. At the same time every crisis related information is informative with respect to enhancing awareness about an event. Such *related* information is pertinent to crisis situations in multiple dimensions such as relating to infrastructure damages, affected individuals/communities, medical support, donations, sympathy and support. Other than the images and videos, most of the data is available in text. There are several machine learning based approaches explored to classify such data into appropriate categories. Within the machine learning scope there are *supervised* and *unsupervised* learning methods. *Supervised* learning methods, as discussed in section 2.2.1, rely on a training data which is labelled in categories, and these labels are treated as the ground truth. The *unsupervised* learning methods, as discussed in section 2.2.2, look for similarity metrics between data points to create clusters which are similar in nature based on a certain aspect. The *supervised* and *unsupervised* methods are based on the features that are generated from the data. Features are the attributes in the data, which are passed to the algorithms in a certain format. In the scope of current research exploration we focus on the text data. In order to generate the features from the text, we require natural language processing techniques, as discussed in section 2.1. Some of the key text processing operations are *normalisation, tokenisation, character encoding, stop word filtering, stemming, lemmatisation, and part of speech tagging*. All these operations are critical in order to generate features from the text. Since text is basically the morphological representation of any language, the text processing techniques are strongly specific to different languages. For instance, there are POS tagger models available for different languages.

One of the key natural language processing techniques is *Named Entity Recognition* (NER). NER is the process of identifying the presence of named entities such as person, location, organisation, object in the text using NLP techniques. The extracted entities can further be enriched with more information about them, also referred to as *semantics*. Linking the entities to specific identifiers which can establish the exact reference to the entity is called *Named Entity Linking/Resolution*. This helps in determining the exact context of the concept and obtain contextual information. This is usually achieved

via *knowledge bases*. *Knowledge bases* are contextual databases that consolidate relational knowledge about various concepts, entities, etc., in a graph form. The nodes are entities and are connected via properties that establish their relationships. Named Entity Resolution addresses the ambiguity which might occur due to similar naming of different entities. To achieve this, the neighbouring concepts contribute to determine the exact context in which the entity is being referred to, thus disambiguate or resolve the conflict. Wikipedia, an online crowd-sourced encyclopedia, is a popular consolidated knowledge source often used for studying *Named Entity Disambiguation* methods. Wikipedia has also been used, along with other knowledge sources, to create large scale multilingual knowledge bases such as YAGO, BabelNet, DBpedia etc. There are also language based lexical resources such as WordNet. Several *Named Entity Disambiguation* services are known to function over these knowledge bases such as AIDA, DBpedia Spotlight, Babelfy, and IBM Alchemy, to name a few. These services also perform Word Sense Disambiguation along with Named Entity Disambiguation. But there are more than just the linked entities that contribute to extended knowledge when it comes to enriching the semantics. Knowledge bases can contribute to extracting several degrees of information related to an entity, such as what type of entity it is or what are its other similar connections in a certain context. For instance, if a person is born at some place we can determine other important people who were born at the same place or which country is that place in or what is the population of that place.

Further, to classify the text, as we mentioned earlier, there are *supervised* and *unsupervised* classification methods. Some of the widely used *supervised* classification methods are *Linear Regression, Logistic Regression, Naive Bayes, SVM, Decision Trees, and Neural Networks*. Among the *unsupervised* methods we have seen an extensive use of *k-means, LDA, PLSA* etc. It has been observed in the literature, when there is a labelled data available, that has been manually annotated and has inter-agreement by multiple annotators, supervised classification methodologies are generally followed since there are pre-defined classes available on the data that can be used to guide the algorithm. Supervised classification approaches suit the problems where the data has been well labelled into classes. In this thesis,

we have relied on humanly annotated datasets, and hence opted for supervised learning approaches instead of unsupervised. We rely on different evaluation metrics to determine the performance of any given method. Most widely used metrics are *precision, recall, F-1 score, Receiver Operating Characteristic*.

Classifying information from a stream of data is also perceived as an *Event Detection* problem (section 2.4.1). *New Event Detection* or *First Story Detection* is a subtask within *Topic Detection and Tracking* (TDT). TDT, initially meant for newswire data, was approached and applied on online data as well. Several approaches focused on identifying a new *event* based on clustering, comparing messages based on similarity functions, frequency of single word or multiple words, etc. We saw such methods being applied to identify crisis events such as *earthquakes*. These methods relied both on frequency burst and also clustering them together based on co-occurrence of words. Another interesting integration of frequency burst and co-occurring words is observed by analysing the burst frequency of multiple words which have a strong cross-correlation in a subgraph of words occurring in the data. Some of the approaches monitored the social media streams for keywords and phrases to gather the data and then use rule based methods to classify the data.

We found many machine learning based approaches to identify crisis related data from the social media data. Different systems have been proposed for filtering, classifying, searching the crisis data such as *TEDAS, ESA, Twitinfo, Twitcident, TWEEDR*, etc. These approaches show an extensive use of classical machine learning methods such as SVM and Naive Bayes algorithms. Some of the works that focus on extracting specific nuggets from the data also use the Conditional Random Fields. Majority of these works rely on statistical features and social media specific features (such as hashtags on Twitter). Other features such as presence of - *user mentions*, *URLs*, *emoticons*, *numeric characters* are also observed in some methods. While there are several machine learning based approaches to classify the crisis related data, not many of them attempt to contextualise the data by enriching the feature set through expanded semantic information. However, a couple of approaches do exhibit

extracting attributes such as *types* and *categories* from DBpedia but they did not target different types of crises or crises in different languages in their problem scope.

We have observed that there is plenty of literature which highlights the use of semantics in text classification problems (section 2.4.3). The semantics enhance the contextual information, which can result in a better definition of the boundaries between data from different classes. Semantics tend to improve the implicit or explicit relationship between different words in the data, thereby yielding a more coherent representation of the overall vocabulary in the data, which might otherwise be a diverse set. Thus, knowledge bases serve as a valuable source to extract the semantic information of the words. Many works established the increase in semantic similarities between different concepts by exploiting knowledge bases. Different types of knowledge bases serve different purposes. For instance, WordNet is an English lexical database. Nouns, verbs, adjective, and adverbs are grouped together into sets of cognitive synonyms, called *synsets*, where each group represents a certain concept. The *synsets* are interlinked in a network via conceptual relations and lexical relations. Such a knowledge base allow us to bridge and relate concepts at a macro or a micro level based on synonymy or hypernymy. Another type of knowledge base are built on large scale encyclopedia like Wikipedia. Some of the popular knowledge base are DBpedia, BabelNet, Freebase, Google Knowledge Graph, YAGO, etc. These knowledge bases compound large scale entities and relationship between them in a graph format. Extensive research has been done to classify text after incorporating semantics from such knowledge bases. Some of the approaches enrich the data and provide a faceted search on top of the enriched data. An effective approach is to first perform NER on the text though NER APIs such as Babelfy, DBpedia Spotlight, IBM Alchemy, etc. These API's can return a specific Wikipedia/DBpedia URI which in turn can be used to extract further relationships for any given entity. Most of the classification methods enhance the vector space by enriching the vocabulary with the semantics. Some works apply additional filtering or refining methods to incorporate the semantics, such as probabilistic approaches to refine the concepts. Recently, the TREC Microblog (2011-15) catalised the research in social media data

classification by providing a large scale corpus, and annotated subsets.

Word-embeddings are another popular approach. Word-embeddings are distributed representation of words which are likely to occur together or represent similar context. They are generated from large scale text corpus. These embedding can be used as dense and high dimensional matrix in neural network based classification approaches. These embedding behave as explicit semantics generated from large scale corpora based factors such as co-occurrence.

Social media crisis data represents widespread classes of data and all of which are crisis relevant, whether it is a post which mentions donation drives or posts that report any emergency situation or posts conveying sympathy with the affected ones. Different crises situations might yield a different nature of data and can often occur in multiple languages. It is neither practical nor always feasible to train a new classifier each time a new type of crisis occurs or a new language that is seen or for specific classes such as *donations* or *infrastructure*, etc. In the literature study we saw several classification methods but most of the approaches do not take the adaptability to a new type of crisis or to a new language into consideration. The applicability of semantics in the crises domain has not been much explored either, as it has been observed in other text classification problems. This motivates us to exploit the scope of semantics to address the diversity in data across different types of crises and across different languages. In the subsequent chapters we will answer the research questions which we have posed earlier.

3

Classifying Crisis Data - A Hybrid Statistical Semantic Approach

In Chapter 1 we highlighted the scope of the problem pertaining to the crisis relevant information on social media platforms. We defined our research scope and proposed the research questions and hypothesis in Section 1.2. In Chapter 2, we provided a study of relevant techniques and related literature, and compared different works covered in the literature with the research scope defined in this

thesis. As we move ahead, we perform experiments to address every individual research question one by one. In this chapter, we focus on addressing the first research question -

- RQ1 - *How could the addition of semantics improve the binary classification of Tweets with regards to their relevancy to crises?*

Different crisis events generate a varying vocabulary of data. Often there might be contextual similarities across crisis data, but this is not easy to capture. For machine learning based classification systems, this variation in the data across different crisis data poses a challenge when they are applied to unseen crisis data, particularly when they are trained on some crisis events and applied to unseen crisis events. Semantic features can enhance capturing the context of such data, as they can capture context of a piece of information contained in a text. To address research question *RQ1*, we explore the use of *semantic features*, extracted via named entity recognition techniques and knowledge bases, in enhancing a binary classification model's adaptability in classifying crisis *related* tweets. In the literature, a wide range of statistical features and machine learning methods have been researched in recent years to automatically classify such information. We compare the semantically enriched model with a baseline statistical features model, and demonstrate that *semantic features* enhance the models' ability to identify crisis *related* content from new crises events. The contributions of the work done in this chapter can be summarised as follows:

- Show the impact of adding different types of semantic features to the feature set for training a classification model which can identify crisis related information from Twitter.
- Exhibit that using a hybrid combination of semantic features and statistical features improves the classifier's performance when classifying the data from new crisis events which were not part of the training data.

3.1 Introduction

The 2016 World Humanitarian Data and Trends report by UNOCHA* reported around 102 million people, across 114 countries, being affected by natural disasters in the year 2015 alone, and causing an estimated damage of $90 billion. There is a massive surge of real time content on social media platforms during such scenarios, often containing information valuable to many stakeholders. There was a 500% increase in the frequency of tweets observed in Japan during 2011 earthquake†. As we have earlier seen in Chapter 2, many of such messages hold relevance to crises scenarios with respect to the information they convey and enhance the situational awareness. This information brings in value to various stakeholders such as impacted communities, relief agencies (for example American Red Cross‡, All Hands Volunteers§), civic authorities etc. But given the voluminous nature of data generated on social media platforms, particularly on Twitter¶, it is nearly impossible to manually filter or sieve relevant and actionable content (Gao et al., 2011). Hence, it is essential to develop automated tools that can robustly perform such filtering. In practice, such tools are largely unavailable and in addition the social media data characteristics (short length, colloquialism, lack of syntactic structure) make it even more challenging to automatically process and generate understanding.

In this thesis, the larger goal that we aim to achieve is to propose classification approaches that are able to identify crisis related information from voluminous social media steams, and filter out irrelevant content. While we have seen a number of approaches focusing on crisis data identification/classification in the literature, a key aspect of adaptability of such systems to *new crisis events* (or new types of crisis or new in new languages) has largely been missing. It is important for such classification systems to be valid and adaptive to new crisis events. When the classifier is applied on to a new

*UNOCHA https://data.humdata.org/dataset/world-humanitarian-data-and-trends
†https://blog.twitter.com/official/en_us/a/2011/global-pulse.html
‡American Red Cross, https://www.redcross.org
§All Hands and Hearts, https://www.allhandsandhearts.org
¶Twitter, https://twitter.com/

crisis event, the test data might differ from the training data (which the classifier has been trained on) in terms of the vocabulary in the data. Despite the variation the data vocabulary between the training and the test events, there might be a contextual similarity between the data sets. But capturing this context is not straightforward, when vocabulary of the data differs.

We hypothesise that *semantic features*, extracted through knowledge bases, can enhance capturing of the context, and alleviate the variations between the training and the test data. Semantic features can also align different concepts which are inter-related via different relationships in knowledge bases (e.g. hypernyms, synonyms, same as, type of, etc.). Such features can boost the keyword-query based search of information. As earlier cited in Chapter 1, section 1.2, in course of crisis situations we may look for relevant information using keywords such as "*building damage AND/OR airport, building, hotel etc.*", and eventually figure out that it is difficult to cover every possible infrastructure oriented aspect. The challenge of varying crisis data can become more evident when there is a new crisis event. In such scenarios, when a trained classification model is applied on unseen crisis data from new crisis events, the classifiers are likely to under perform due to inconsistency between the vocabulary of the training and the test data. *Semantic features* should be able to incorporate the contextual consistency across varying crisis data (events) in machine learning based classification models.

Following this hypothesis, in this chapter we aim to answer the first research question:

> RQ1 - *How could the addition of semantics improve the binary classification of Tweets with regards to their relevancy to crises?*

As earlier seen in Chapter 2, much of previously explored research on classification of crises data into *related* and *not related* has focused on supervised (Li et al., 2012b; Karimi et al., 2013; Stowe et al., 2016; Zhang & Vucetic, 2016) and unsupervised (Rogstadius et al., 2013) machine learning methods. Many of these methods use features such as n-grams and statistical features (text length, POS presence in the text, presence of URL's, number of hashtags). As mentioned earlier, in this chapter

we aim to explore the impact of extracted semantic information as features in identifying crisis related information, in a binary classification system. Most of these approaches use Support Vector Machines, Naive Bayes, and/or Conditional Random Fields (Power et al., 2014; Stowe et al., 2016; Imran et al., 2013b). Unsupervised methods usually rely on clustering and keyword processing approaches. In this chapter, we propose a hybrid approach where both *statistical* and *semantic* set of features play a role in building the binary classification model. The semantic features, explained in more detail in section 3.2.2, include extracted entities and hypernyms from knowledge base *BabelNet* (Navigli & Ponzetto, 2010; Navigli & Ponzetto, 2012).

In this chapter we will explore whether the semantic features are effective in improving the applicability of the classification models on previously unseen crisis events. We use a labelled dataset of multiple crisis events named *CrisisLexT26* (Olteanu et al., 2015). We elaborate more on the data used in the experiments in this chapter in section 3.2.1. The results show that adding semantic information to the model along with statistical features enhances the classifier's performance to identify crisis *related* tweets when applied to unseen crisis events as compared to the baseline of only statistical features.

The rest of the chapter is organised as follows: Section 3.2 elaborates on our classification approach. Section 3.2.1 describes the dataset used, and selection of the labelled data and events. Section 3.2.2 describes the feature engineering, and types of features: *statistical* and *semantic*. Section 3.3 details our experimental set up and results. We discuss the findings in section 3.4 and summarise the work in section 3.5.

3.2 Crisis Related Information Classification

To differentiate crisis *related* content from *not related* in social media data, we propose a binary classification approach. In our case, we perform experiments on tweets from Twitter. Tweets are publicly

shared posts by users on Twitter platform. In this section we will elaborate on dataset, feature engineering, and classifier selection.

3.2.1 Dataset

To this end, we have used *CrisisLexT26* (Olteanu et al., 2015) dataset. This data set has been used in some of the works (Imran et al., 2016b; Burel et al., 2017b; Burel et al., 2017a) covered in the literature study in Chapter 2. This is an annotated dataset of 26 crisis events that occurred between 2012 and 2013. Each event has 1000 labelled tweets. The tweets were originally collected using the standard techniques of using specific hashtags and/or impacted location name paired with canonical form of disaster such as *Queensland Flood* or a meteorological name. The labels are categorised into four categories: *Related and Informative*, *Related and Not Informative*, *Not Related*, and *Not Applicable*. The label *Related and Informative* meant that a given tweets conveyed some useful information which assists in understanding about the crisis event. *Related and Not Informative* meant that while the tweet was conveying an information which was referred to the crisis event but did not contain useful information. *Not Related* were the ones, as the name suggests, were not related to a crisis, and *Not Applicable* were the ones that were not readable or too short. Additional information about the CrisisLexT26 data set can be found on the CrisisLex website*.

In this particular work, we focus on tweets in English. Hence from 26 crisis events, we selected the events which were dominantly in English. We selected the following events: *Australia Bushfire* (ABF), *Boston Bombing* (BOB), *Colorado Flood* (CFL), *Colorado Wildfire* (CWF), *Los Angeles Shooting* (LAS), *Queensland Flood* (QFL), *Savar Building Collapse* (SBC), *Singapore Haze* (SGH), and *West Texas Explosion* (WTE). In order to facilitate a binary classification system, we need a two label dataset. For this to happen we merged the tweets those labelled as *Not Related* and *Not Applicable* as *Not Related* class, thus obtaining a total of 1539 tweets as *Not Related*. For the other class, we merged

*CrisisLex, http://www.crisislex.org/

Table 3.1: Event data distribution per class

Event	Class -1,0 Size		Total
	1 (Related)	0 (Not Related)	
West Texas Explosion (WTE)	111	89	200
Colorado Wildfire (CWF)	247	247	494
Colorado Flood (CFL)	89	75	164
Australia Bushfire (ABF)	250	250	500
Boston Bombing (BB)	79	71	150
Los Angles Shooting (LAS)	130	120	250
Queensland Flood (QFL)	320	281	601
Savar Building Collapse (SBC)	261	239	500
Singapore Haze (SGH)	80	67	147

Related and Informative with *Related and Not Informative*, thus creating the *Related* class and obtained 7461 *Related* class tweets. We can see there is a huge disparity between the size of *Related* and *Not Related* tweets. Thus, to reduce this disparity we further randomly selected 1667 crisis *Related* tweets. This disparity was addressed at each individual event level. This gave us a near balanced dataset of 3206 binary labelled tweets from two classes *Related* and *Not Related*. Table 3.1 shows final data distribution across classes for each selected event (class *Related* represented by '1' and class *Not Related* represented by '0').

3.2.2 Features

In our binary classification approach to classify social media posts as crisis *Related* and *Not Related*, we generate two type of features; *Statistical Features* and *Semantic Features*. As elaborated in Chapter 2, *Statistical Features* have widely been studied and used in several text classification methods (Li et al., 2012b; Karimi et al., 2013; Stowe et al., 2016; Zhang & Vucetic, 2016). We use the *Statistical Features* as a baseline approach for the binary classification task. These statistical features reflect the quantifiable properties of the text as well as linguistic properties. *Semantic Features* reflect the named

entities emerging from the tweets, as well as their hierarchical information (hypernymy) extracted using an external knowledge graph.

Statistical Features

Given a text post, we extract the following as statistical features:

- *Number of nouns*: Nouns generally can refer to entities such as location, person, organisations, etc, involved in the scope of crisis event. It forms the part of Part of Speech (POS) features, as explained in Section 2.1.(Imran et al., 2013a; Imran et al., 2013b; Stowe et al., 2016)

- *Number of verbs*: Verbs can indicate that any action is being undertaken or occurring in course of the crisis event. It forms the part of Part of Speech (POS) features.(Imran et al., 2013a; Imran et al., 2013b; Stowe et al., 2016)

- *Number of pronouns*: Much like nouns, pronouns may also refer to the actors, locations, or resources that are named in a given text posted during the crisis event. It forms the part of Part of Speech (POS) features.

- *Tweet Length*: Total number of characters in a given post. The length of a post may be related to the amount of information contained in it.(Imran et al., 2013a; Imran et al., 2013b; Sakaki et al., 2010)

- *Number of words*: Similar to the length to the post, number of words may also be an indicator of the amount of information present in the post.(Imran et al., 2013b; Karimi et al., 2013)

- *Number of Hashtags*: Hashtags are social media specific features, which often indicate the themes of the post and are manually generated by the posts' authors. These features are indicated by any alph-numeric phrase beginning with a '#' sign in the text. The presence or ab-

sence of the number of hashtags can be important discriminatory feature.(Imran et al., 2013a; Imran et al., 2013b; Karimi et al., 2013)

- *Readability*: We use Gunning fog index which uses average sentence length (ASL) and the percentage of complex words (PCW): 0.4 * (ASL + PWC). This feature aims to determine how complex a post is for humans to parse*.

- *Unigrams*: Unigrams provide a keyword-based representation of the content of the posts, thus enabling a vector based representation of the overall data.(Imran et al., 2013a; Imran et al., 2013b; Karimi et al., 2013; Li et al., 2012b; Stowe et al., 2016; Zhang & Vucetic, 2016)

We used Weka data mining software† to perform pre-processing of the data and transforming into unigrams, by using its *StringToWord* functionality. Further we converted all the tokens into lower case and performed *stemming* (using Lovins' algorithm)‡, stopword removal, and tf*idf transformation. We have explained these pre-processing techniques on text data in Chapter 2 in section 2.1. This resulted in a total unigram size of 10655. For extracting the Part of Speech (POS) tags and the statistical features listed above (top five), we availed a widely used tool, the Stanford Core NLP software§ in Java. We count the number of Hashtags by identifying the number of times the character '#' is used in the text, and readability is computed using the Gunning fog index in Java.

Semantic Features

We generate the semantic features in multiple steps, as shown in Fig. 3.1. This extraction of semantics is done in three steps: (i) semantic annotation, (ii) semantic expansion, and (iii) semantic filtering.

*https://en.wikipedia.org/wiki/Gunningfogindex
†Weka, https://www.cs.waikato.ac.nz/ml/weka/
‡http://www.mt-archive.info/MT-1968-Lovins.pdf
§Stanford Core NLP, https://stanfordnlp.github.io/CoreNLP/

Each individual step generates a different kind of semantic feature and we explore each of them by trying different combinations of features, for the binary classification task.

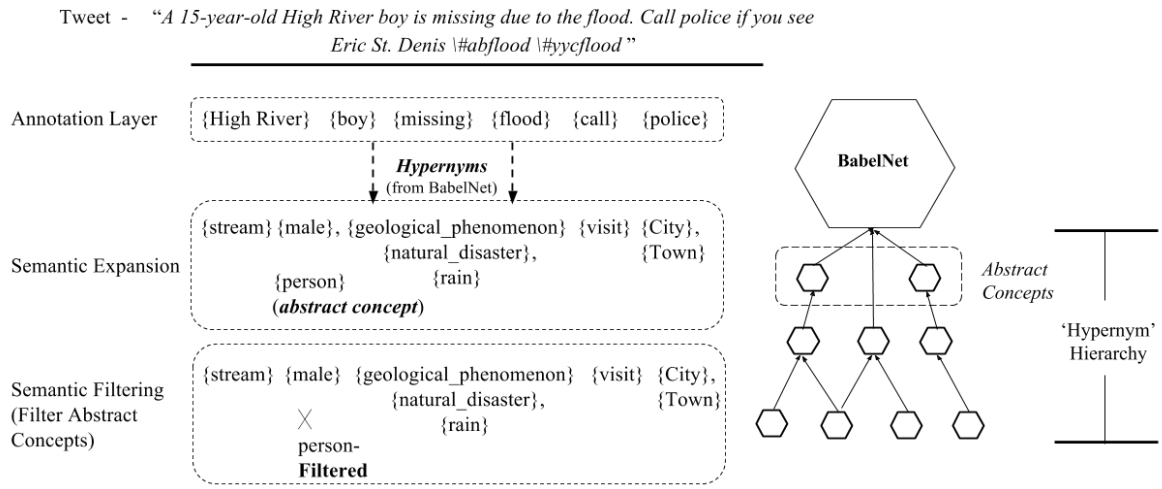

Figure 3.1: Semantic Features: Annotation, Expansion, & Filtering

- *Semantic Annotation Features* (SemAF): The first step is to extract the annotated entities in the tweets via Named Entity Recogniser (NER) services. We used Babelfy* (Moro et al., 2014) for this purpose. Babelfy performs multi lingual word sense disambiguation and entity linking, by linking the entities to BabelNet- a multi lingual knowledge base (Navigli & Ponzetto, 2010; Navigli & Ponzetto, 2012). For each entity that Babelfy annotates in a given text, it returns a unique identifier in the form of a *Synset ID* for each identified entity. This *Synset ID* is a unique identifier for any particular concept/entity in the knowledge base BabelNet. For each *Synset ID* BabelNet stores multi-dimensional semantic information such as multi-lingual senses, hypernyms, synonyms, similar-as, etc, relationships and these can be extracted from the knowledge base. In this semantic feature set, once we get a Synset ID from Babelfy, we extract main sense of each Synset ID in English from BabelNet. As an example if we look at Fig. 3.2 (a screenshot of Babelfy API's web interface), for a given post,

*Babelfy, http://babelfy.org/

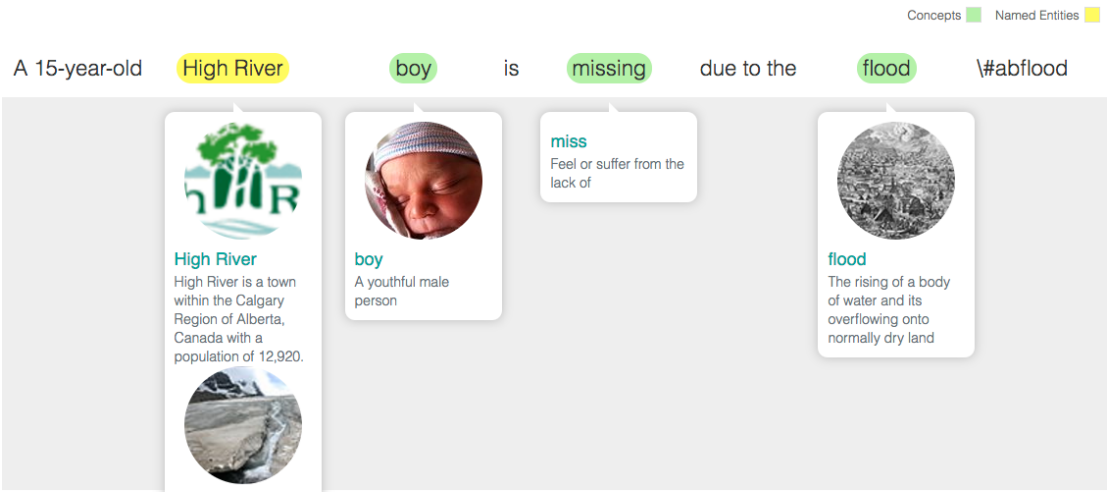

Figure 3.2: Semantic Annotation Example via Babelfy

> *"A 15-year-old High River boy is missing due to the flood. Call police if you see Eric St. Denis #abflood"*

Babelfly identifies and annotates entities such as *High River*, *Boy*, *Flood*, etc. Annotating the entire data resulted in 12,006 unique concepts.

- *Semantic Expansion Features* (SemEF): In this step, after extracting the annotations, we expand the semantics via BabelNet knowledge base. For each extracted entity/concept (Synset ID) we retrieve every *hypernym*, at a distance 1 (which implies direct hypernyms only, and not hypernyms of hypernyms), of these entities. *Hypernyms* are the words with a broader meaning of another word, thus constituting a category into which words with more specific meanings fall*. For example, *fruit* is a hypernym of *apple*. We hypothesise that *hypernyms* reflect the a broader/upper level concept to each entity, thus encapsulate the broader semantics of the crisis related information. As an example, let us consider the entities *fireman* and *policeman* often occur in the crisis related posts. If we expand the semantics to the hypernym level of both of

*https://www.lexico.com/en/definition/hypernym

these entities, we observe that both have at least one common hypernym - *defender*. Hence, when a new post arrives containing an entity *MP (Military Police)*, then it is more likely to be crisis related since it also has *defender* as a hypernym. We expanded semantics for each concept that got annotated and yielded in an additional 7032 unique concepts.

- *Semantic Filtering Features* (SemFF): The process of semantics expansion of extracted entities, sometimes can yield very generic or broad level of concepts which eventually hold a very low discrimination power between crisis *related* and *not related* content. For instance, the concept *Person* is hypernym to many entities and it appears in both crisis *related* and *not related* posts. Considering that every single concept which relates to a person/individual will have its hypernym as *person*, and instance of every such concept will have its hypernym as *person*. For instance, concepts such as *neighbour*, *sportsman*, *relative*, *collector*, *baby*, *socialiser* among many others, have their hypernym as *person*. This make a concept such as *person* a very broad concept in itself. For such issues we propose a filtering approach which aims to curb on semantics from expanding to a very broad range of concepts. Our filtering approach is based on computationally determining the depth of a concept in the hierarchy of BabelNet. To determine the depth of a concept, we iterate through the hierarchy of BabelNet through REST API, via 11653 unique BabelNet Synset IDs collected after annotation and hypernym extraction. To create this hierarchy, for each Synset ID, we iterate through BabelNet via 2 relationships- *hypernyms* (thus generating a network of concepts above it) and *hyponyms* (generating a network of concepts below it). This process resulted in a network of 3.9 million relations, for nearly 3.5 million concepts, which are put in a Directed Graph, where the node which has the highest *betweeness centrality* is determined as the most abstract concept of the network. To this end, we used NetworkX* graph library in Python. We identified the most abstract node as the fol-

*Networkx, https://networkx.github.io/

lowing Synset ID - 'bn:00031027n', which relates to the concept with its main sense as *'Entity'*. Next, the *shortest path* between any given concept, out of 3.5 million concepts, and the node 'Entity' is defined as the *depth* of that particular concept in the hierarchy. We defined level 0 (zero) as the depth of the node 'Entity', and found the maximum depth reaching 21. We plotted the discriminative features from the data, by calculating Information Gain score, and plotted them against the depth in the hierarchy. We did this across each event and observed the depth/levels at which the features tend to be most the informative (based on Information Gain) were between 3 and 7. In Fig. 3.3 and Fig. 3.4, we show these plots for the training data corresponding to Singapore Haze and Australia Bushfire events. The darker and bigger dots show features with higher Information Gain. We attach the plotted graphs for all the events in the Appendix A, from Figure A.1-A.7. In the filtering phase, we filter out concepts whose depth does not fall between the level 3 and 7. This resulted in 574 concepts getting filtered out across the balanced data from the selected 9 events.

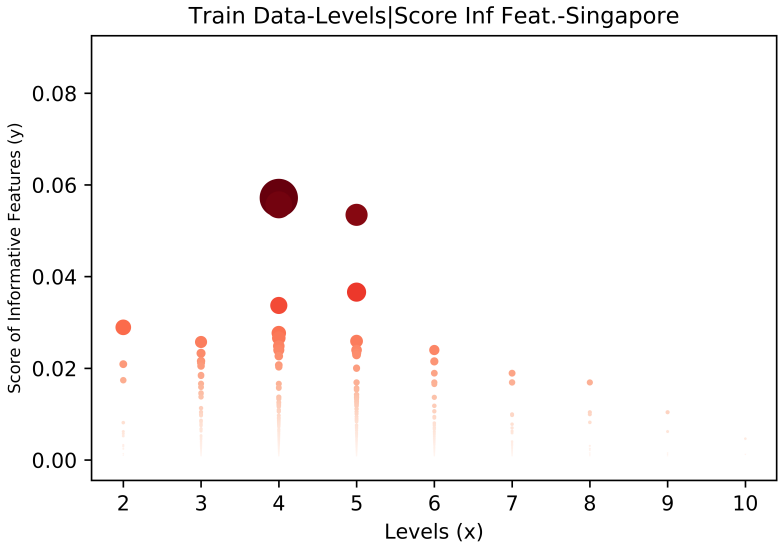

Figure 3.3: Information Gain/Level:Training Data-Singapore Haze

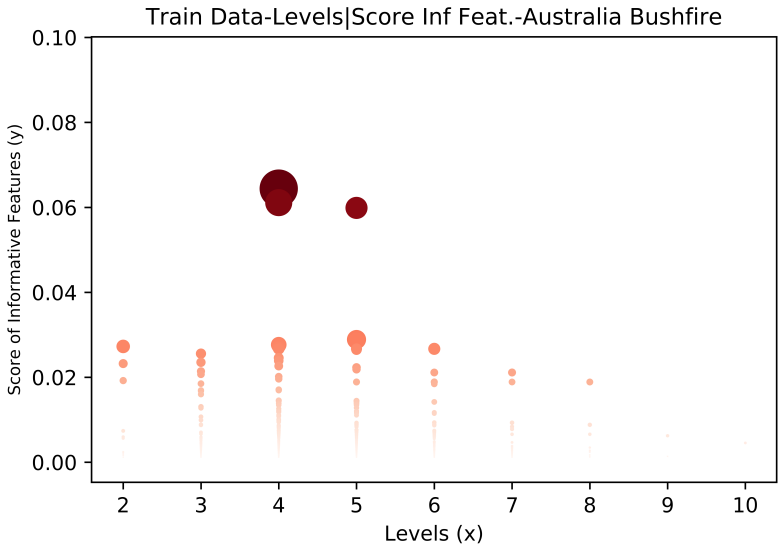

Figure 3.4: Information Gain/Level:Training Data-Australia Bushfire

3.2.3 Classifier Selection

While selecting the binary classification algorithm, we kept in mind the following:

A. For future unseen data, it is important to avoid over-fitting approaches.

B. Not aiming to perform a memory costly operation.

C. Limited training instances (nearly 3200) & high dimensionality (unigrams).

Keeping the above aspects in consideration, we opted for Support Vector Machine (SVM) (Cristianini et al., 2000) with Linear Kernel for classification. Also, the use of SVM in text classification problems is a widely followed and acceptable methodology (Lorini et al., 2019; Stowe et al., 2018; Stowe et al., 2016; Agarwal et al., 2012). In the recent times, some of the related works have also proposed crisis data classification methods based on deep learning approaches such as Convolutional Neural Networks (CNN) using *word embeddings* (ALRashdi & O'Keefe, 2019; Lorini et al., 2019;

Alam et al., 2018; Burel et al., 2017a; Nguyen et al., 2017). Some of these approaches (Lorini et al., 2019; Alam et al., 2018) used pre-trained large scale word embeddings such as GloVe embeddings (Pennington et al., 2014) or Google word embeddings (Mikolov et al., 2013). However, if the size of dataset is quite small the deep learning approaches do not exhibit better performance than other classification approaches such as SVM and/or Naive Bayes (Stowe et al., 2018).

For our experiments, the training data always varied between 2800-3200 depending on the combination of the events for training and excluded test dataset. The features (unigrams of tweets or unigrams of tweets + semantic expansion) varied in the range of 15,000-18,000. So it is evident that the number of features were exclusively high in comparison to the number of training samples. Explanations can be referred to understand the usage of linear kernel over other kernels*. Radial Basis Function (rbf) kernel or a Polynomial Kernel may cause an over-fitting problem, hence we opted for a linearly separable hyperplane. Also, we compared the SVM (with linear kernel) with a standard Convolutional Neural Network (CNN) architecture using word embeddings (Kim, 2014), and found out that in the given dataset the CNN based model does not perform as well as the baseline SVM based model built on statistical features. The details of the CNN with word embeddings architecture set up is provided in section 3.3.1 under *Crisis Classification Model* scenario.

We also validated this by comparing the statistical significance of SVM Linear Kernel over RBF and Polynomial (degree 3) kernel using the Paired-T Test. Over 10 iteration of 10-fold cross-validation over the entire dataset (1667 crisis related, 1539 not related) under all 5 feature sets (explained further), on an average - SVM Linear Kernel had an accuracy of 88%, Polynomial (3 degree) 68%, and RBF had an accuracy of 66%.

*http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

3.3 Experiments

In this section we describe the design of our experimental set up in which we create different models based on a combination of different features, and selection of training and test data. Further, we report our findings and discussions on how semantics impact the classifier performance in classifying crisis related data from new/unseen events. And thus answer the first research question *RQ1*.

3.3.1 Experimental Setup

In this work, we design two main experimental scenarios:

- *Crisis Classification Model*: In this set up we aim to compare the performance of the classifier based on *statistical features* with the classifier based on combination of *statistical* and *semantic features*, and analyse if adding semantics boosts the classifier's performance. For this experiment, we take the entire data as mentioned in section 3.2.1 and train the model and validate it following a 10 iteration of 10-fold cross validation approach. To this end, we used the WEKA software (v.3.8)* to generate the classifiers. We create the following models:
  - **SF** : A classifier generated with statistical features; our baseline.
  - **SF+SemAF**: A classifier generated with statistical features, and semantic annotation features.
  - **SF+SemAF+SemEF**: A classifier generated with statistical features, semantic annotations, and their hypernyms, i.e., the Semantic Expanded Features.
  - **SF+SemFF**: A classifier generated with statistical features, and filtered semantic annotations, along with their hypernyms, i.e., the Semantic Filtered Features

*http://www.cs.waikato.ac.nz/ml/weka/

As mentioned in section 3.2.3, we also compare the above mentioned models with a deep learning model based on CNN architecture using word embeddings. The architecture of the CNN model is similar to a widely adopted architecture proposed by Kim (Kim, 2014) using an embedding layer. We used GloVe embeddings that are a vector (of dimensionality 100) obtained from large scale corpora of tweets (2 billion tweets, 27 billion tokens, and 1.2 million vocabulary) (Pennington et al., 2014). The input sequence of data (tweets in the dataset) and embeddings are fed to a sequential network of convolutional layers with 128 convolutional filters of sizes 3, 4, and 5. The output layer performs a binary classification (of crisis relevancy) using a sigmoid activation and used RMSprop optimizer. We performed 10 iterations of 10-fold cross validation. We used Keras* (based on Tensorflow†) for building the model in Python 3. The results are shown in Table 3.2, with the model named as **CNN-embeddings**.

- *Unseen-Crisis Event Classification*: In the second scenario of the experimental set up, we retest the classifier models we built above, by applying them to a new crisis data that was not part of, or not seen, in the training set. In this experimental set up, we generate the four classifiers, as mentioned in the previous task. However, since the model is to be tested and applied on a new crisis event each time, we use 8 out of 9 crisis events to train the model, and apply the model on the single crisis event that was left out of the training data for validation (test data). Since there are four types of models and 9 overall events, we end up performing 36 different classification experiments.

3.3.2 Results: Crisis Classification Model

In this part we highlight the results from the first experimental set up, where we analyse the performance of various feature models based on 10 iterations of 10-fold cross validation. The results are

*Keras, https://keras.io/
†Tensorflow, https://www.tensorflow.org/

shown in the Table 3.2, where it presents Precision (P), Recall (R), F-measure(F1) value (from 10-fold cross validation), mean of F-measure ($F1_{mean}$) of 100 results from 10 iterations, standard deviation in F-measure (σ), $\Delta F/F_{mean}$ which shows the increment of F_{mean} over the baseline, and p-value which reflects the significance of the F_{mean} score of any given model in comparison to the baseline. p-value is calculated using 2 sample t-test*, where we take into consideration the mean F_{mean} score and standard deviation of two compared models, i.e. baseline and any other model.

From the Table 3.2, if we compare the F_{mean}, the two semantic models, SF+SemAF and SF+SemAF+SemEF, show an improvement over the baseline SF model. However SF+SemAF (annotations) shows a better performance than SF+SemAF+SemEF (annotations and hypernyms), with still a very marginal gain of 0.6% over the baseline. The CNN-embeddings model shows F_{mean} score of 0.845 which is significantly less than the baseline model SF. This is similar to the observations made by Stowe and colleagues (Stowe et al., 2018), as the overall size of data is merely around 3200 tweets, with almost 1500 tweets in each class. The deep learning models are likely to under-perform in such a small size of data. Based on the performance of CNN-embeddings, and the amount of time it requires to train a deep learning model, we chose to focus on rest of the SVM based classification models and treat SF model (based on statistical features) as the baseline in further experiments.

*2 sample t-test, https://select-statistics.co.uk/calculators/two-sample-t-test-calculator/

Table 3.2: 10 iterations of 10-fold Cross Validation, showing performance of statistical semantics classifiers vs statistical classifier.

Features	P	R	F1	$F1_{mean}$	Std. Dev. σ	$\Delta F/F1_{mean}$	p-value (for $F1_{mean}$)
SF (Baseline)	0.864	0.865	0.865	0.872	0.017	-	-
SF+SemAF	0.871	0.871	0.87	0.877	0.017	0.0057	0.039
SF+SemAF+SemEF	0.868	0.868	0.868	0.873	0.017	0.0011	0.67
SF+SemFF	0.863	0.864	0.864	0.873	0.018	0.0011	0.68
CNN-embeddings	0.851	0.846	0.843	0.845	0.049	-0.09	< 0.001

3.3.3 Results: Unseen-Crisis Event Classification

In Table 3.3 we report the results of experiments conducted for this unseen-crisis event classification scenario, where we leave one crisis event entirely out for validation, and train the classification model on rest of the eight events*. In the table each row depicts the particular crisis event that was left out of the training data and is used as a test data to validate the model created on the remaining nine crises events. The size of each test data and corresponding training data can be determined from the Table 3.1. We created four different classification model, as shown above, for each of these 9 events in the dataset- the one using only statistical features (SF), which we use as baseline, and the semantically enhanced ones: (i) SF+SemAF, (ii) SF+SemAF+SemEF and, (iii) SF+SemAF+SemFF.

The results, in the Table 3.3, report Precision (P), Recall (R), F1-measure (F) and the increment of F measure over the baseline, $\Delta F/F$ for each of the nine tested crisis event across all the four models.

*Each model was tested on the 8 event dataset it was trained on using 10 fold cross-validation to ensure its accuracy before applying it to the 9th event data. There accuracy drops around 17% on average when applied to new events.

Table 3.3: Unseen-Crisis Event Evaluation- SF, SemAF, SemEF, and SemFF feature sets (best set of features highlighted in bold)

	SF			SF+SemAF				SF+SemAF+SemEF				SF+SemFF			
Test	**P**	**R**	**F**	**P**	**R**	**F**	$\Delta F/F$	**P**	**R**	**F**	$\Delta F/F$	**P**	**R**	**F**	$\Delta F/F$
Event															
WTE	0.806	0.805	0.804	0.813	0.81	0.808	0.005	0.819	0.815	0.812	0.010	0.828	0.825	**0.823**	0.024
CWF	0.643	0.64	0.638	0.633	0.623	0.617	-0.033	0.716	0.715	**0.714**	0.119	0.712	0.711	0.71	0.113
CFL	0.784	0.774	0.774	0.796	0.793	0.793	0.025	0.79	0.787	0.787	0.017	0.797	0.793	**0.793**	0.025
ABF	0.776	0.774	0.774	0.782	0.778	0.777	0.004	0.811	0.8	**0.798**	0.031	0.803	0.79	0.788	0.018
BB	0.713	0.707	0.702	0.693	0.693	0.693	-0.013	0.734	0.733	0.732	0.043	0.761	0.76	**0.759**	0.081
LAS	0.811	0.808	**0.808**	0.777	0.776	0.776	-0.040	0.777	0.776	0.775	-0.041	0.789	0.788	0.787	-0.026
QFL	0.699	0.694	0.694	0.702	0.696	**0.695**	0.001	0.702	0.691	0.69	-0.006	0.704	0.692	0.691	-0.004
SBC	0.618	0.594	0.58	0.651	0.64	**0.636**	0.097	0.619	0.584	0.561	-0.033	0.617	0.586	0.565	-0.026
SGH	0.716	0.66	0.648	0.744	0.68	0.669	0.032	0.737	0.68	**0.67**	0.034	0.732	0.673	0.662	0.022
Avg.			0.714			0.718	0.009			0.727	0.0194			0.731	0.0251
%							0.9%				1.94%				**2.51%**

The three semantic feature models, i.e. SF+SemAF, SF+SemAF+SemEF, SF+SemFF, on average enhance classification results in all cases. As an observation, SF+SemAF improves the classification over the baseline SF, in 6 out of 9 case, with an average of 0.9% increase in F-1 measure (notice the number of positive $\Delta F/F$). As opposed to our observation in 10-fold cross-validation setup, SF+SemAF+SemEF performs much better in the unseen-crisis event scenario, where the model is applied on a new crisis event. SF+SemAF+SemEF shows improvement over the baseline in 6 out of 9 tested events, with an average gain of 1.94% over the baseline. Also, to be noted is that SF+SemAF+SemEF improves over SF+SemAF in 5 out of 9 cases. The semantic filtering of abstract concepts approach SF+SemFF (applied over SF+SemAF+SemEF) results in improvement of performance over SF+SemAF+SemEF in 7 out of 9 cases (average of 0.6% gain over SF+SemAF+SemEF). This observation validates the assumption that certain abstract concepts potentially induce noise and thus filtering them out could aid in enhancing the classifier's performance. The filtering model SF+SemFF, on an average, gains 2.51% in performance over the baseline.

Table 3.4: Examples of posts that were misclassified by the statistical classifier, but classified correctly by the semantic classifiers.

PostID	Text	Label
Post1	I GET 5078 REALL FOLLOWERS! http://t.co/qrF5dpD3 #BestRap,#boulderflood,#PutinsFlik,#Rem #in	Not Related
Post2	@Stana_Katic Can we get some loveballs in Colorado? We need it after all the flooding! Love you! Xo	Not Related
Post3	RT @LarimerCounty: #HighParkFire burn area map as of Monday night 10 p.m. http://t.co/1guBTcXX	Related
Post4	Colorado wildfires their worst in a decade http://t.co/RtfLmfds	Related
Post5	RT @RedCross: Thanks to generosity of volunteer blood donors there is currently enough blood on the shelves to meet demand. #BostonMarathon	Related

3.3.4 Feature Analysis

We manually analyse some of the tweets that were *misclassified* by the statistical baseline model SF, but classified correctly by the semantic models. This will help us in better understanding the impact of the semantics in such a problem scenario. In this context, we manually analyse some of the tweets that were *misclassified* by the statistical baseline model, but were correctly classified when using semantics (see Table 3.4). The analysis is complimented by taking into consideration the Information Gain (IG) score as well, to determine the discriminatory nature of any given feature.

When the Information Gain (IG) was calculated for the features corresponding to the baseline classifier SF, we found that *number of hashtags* was the most relevant feature. On performing a manual analysis of some of the tweets, we noticed that tweets belonging to the *Not Related* class tend to either have zero hashtags (e.g. Post 2 in Table 3.4) or might contain too many hashtags (e.g. Post 1 in

Table 3.4). Among the other discriminative statistical features we found *number of nouns* and *number of pronouns*. This explains our hypothesis behind using the statistical features such as *nouns* and *pronouns*, as crisis *related* posts tend to mention persons, resources, and/or locations in the course of crisis. Further, in the semantic models (*semantics feature* added with the *statistical features*) we observed that semantic annotations, and hypernyms were among the highly ranked features based on the Information Gain (IG) score. Hypernyms such as *Happening* and *Event*, which are hypernyms of concepts such as *incident*, *fire*, *crisis*, *disaster*, and *death*, were among the top 10 IG score features (among almost 800 features which had IG > 0). Annotations such as *Structural_and_Integrity_Failure* were also among such features.

If we look at Post 3, in Table 3.4, it was wrongly classified by the statistical feature model. The post contains the term *burn*, which does not (or barely) occur in the corresponding training data, instead *fire* is more prominent in the training data. In statistical feature model, there is no semantic information to relate the term *burn* with the crisis *related* class. However, this post is correclt classified by the SF+SemAF model, because the term *burn* is enriched by the semantic annotation process and now adds the concept *Fire* to the post. *Burn* and *Fire* belong to the same BabelNet synset and are therefore identified by the same ID within the semantic annotation features.

Post 4 was wrongly classified by the SF+SemAF model, but semantic expansion by adding hypernyms in the SF+SemAF+SemEF model resulted in correct classification of this instance. The semantic expansion of the annotated term *wildfire* returned the concept *Fire* which is a feature with a high IG score.

Post 5 shows the case where expanding the semantics via adding the hypernyms did not help in correct classification, instead it brought in concepts which had low discriminatory powers. Expanding to hypernyms of the annotations such as *Thanks* and *Meet* returned concepts such as *Virtue* and *Desire*, which were not only very abstract but very low in their IG score. Thus it contributed towards adding the noise. The semantic filtering model SF+SemFF aided in removing the abstract concepts

and thus resulting in increased informative nature of rest of the features such as *Volunteer*, *Benefactor*, and *Donor*.

3.4 Discussion

In this chapter, we demonstrated that when a classifier model trained on some crisis data is applied to identify crisis *related* information from a new crisis event, the model trained via mixing the *semantic features* and the *statistical features* performs better at the task in comparison to the model built only on *statistical features*. This addressed the research question *RQ1* we had posed. Through various iterations of experiments and feature analysis we were also able to prove our hypothesis that semantic features can improve the classification performance when applied to an unseen crisis event. We used knowledge graphs to enrich the semantic and thus enhance the vocabulary by incorporating entity sense and hypernyms. This captured the broader context of tweets which are otherwise very limited in their context due to shorter length. Although, the inclusion of semantics can also potentially induce noise by including very broad and abstract concepts, which We observed in the feature analysis section. One of the reasons for expansion to bring in very broad/abstract concepts could be the unsymmetrical mappings of *hypernym-hyponym* relationship in BabelNet, due to which the hierarchical expansion returned abstract concepts. Perhaps using a strictly symmetrical resource such as WordNet* might prevent such issues from happening, but in that case it will not cater to information about other entities which a large scale knowledge graph such as BabelNet or DBpedia† can offer. Expanding the semantics to beyond the hypernymy or synonymy and incorporating more information about entities and their types can also play a role in classification. Also, expanding this model to a bigger data size can boost the classifier's learning abilities.

While the experiments have enabled us to answer the research question *RQ1*, we need to further

*WordNet, https://wordnet.princeton.edu/

†Dbpedia, https://wiki.dbpedia.org/

examine how the classifier would behave if it was not only a new crisis event it was applied to, but also an entirely new *type* of crisis event. For example, a crisis classification model trained on the data from *earthquake* type crisis events, is applied on data from *floods* or *wildfires* type of crisis event. This forms part of our research question *RQ2*, and we will expand our approach from this chapter to address the research question *RQ2* in the next chapter.

3.5 Summary

Semantics help in broaden the contextual information, thus enhancing the classification algorithms' performance in identifying crisis *related* information from new crisis events. In this chapter we presented our work on creating a hybrid system which incorporates both *statistical* and *semantic* features to classify crises events when they have not been seen in the training data, thus aiming to answer the first research question *RQ1*- "*How could the addition of semantics improve the binary classification of Tweets with regards to their relevancy to crises?*".

We extracted the *semantic features* in two stages: (i) main sense of the entities annotated via Babelfy; (ii) extracted hypernyms and their senses for each annotated entities via BabelNet. The *statistical features*, which are usually the quantified linguistic and structural properties of the text, were computed. A binary classification approach, based on SVM classifier, was adopted to classify the tweets as crisis *related* and *not related*. We also observed that in some cases the semantic enrichment resulted in addition of very abstract concepts, which are not descriminative in their nature. To address the problem of abstract concepts, we proposed a filtering model which filters out abstract concepts based on their hierarchy. The hierarchy was created using the hypernym-hyponym relationship network in the BabelNet, and later identified the depth of the informative features (based on Information Gain score) in the hierarchy.

We were able to demonstrate, and thus answer the research question *RQ1*, that adding the seman-

tics to the statistical features in a binary classification model enhances the performance and helps the classifier in identifying crisis *related* information from new unseen crises events.

4

Classifying Crisis Information Relevancy Across Crisis Types

Semantic features enhance the accuracy of the classifier, as seen in Chapter 3, when classifying crises posts from new events according to their relevancy to crises. In this chapter, we take the problem scenario a step beyond the scope of Chapter 3, where the training and test data were independent of

the *type* of the event such as *floods*, *earthquakes*, and *train crashes*. This chapter focuses on addressing second research question -

- RQ2 - *To what extent could semantics improve Tweets classification for new types of crisis events?*

For instance, how will the classification model behave if it was applied to data from events such as *floods*, while it was trained on different *type* of events such as *earthquakes* and *train crashes*. In this chapter, we will analyse how do *semantic features* impact the adaptability of the classifiers when the *types* of crises events tested by the classifier are different than the *type* of crises events the classifier is trained on.

4.1 Introduction

In the previous chapter, we explored the role of semantics in classifying the crisis *related* information when the tested event was not seen in the training data. However, what we did not take into account was the *type* of crisis the event reflected such as *floods*, *earthquakes*, *train crashes* etc. If the model was strictly trained on certain *types* of events, they are not likely to perform well when tested on different *types* of crises events, given the model is built only on *statistical features*. Given the real time need of identifying the *relevant* information in the midst of crises, re-training the model on a new *type* of crises is not a viable solution due to lack of well labelled training data for a new type. Some previous works focused on domain adaptive classifiers which were trained on some crises type events and applied to different types (Li et al., 2018a; Pedrood & Purohit, 2018; Imran et al., 2013b). However, a limitation of these approaches is centered around not considering enough types of events (limited to 2). Imran and colleagues (Imran et al., 2013b), tried a domain adaptive approach by considering two disasters: Joplin 2011 tornado and Hurricane Sandy. A model trained on a subset of Joplin tornado data was applied to Hurricane Sandy and the remaining part of Joplin tornado data. As we can notice, this work was limited to only two crises, a hurricane and a tornado, which can often result in similar type of data

due to the similarity in their nature of impact. Also, they did not consider the aspects of semantics, that could have improved the model's adaptability in multiple *types* of crises. Some approaches such as adversarial training and graph embedding have also been seen in domain adaptive problems to different crisis events (Alam et al., 2018), although they were tested with just two crisis types with one event in each type.

In this chapter we aim to address the second research question:

> RQ2 - *To what extent could semantics improve Tweets classification for new types of crisis events?*

We aim to analyse the performance of a model trained on certain types of events (e.g., earthquakes and train crashes), when applied to types of events which were not seen in training data (e.g., floods, typhoons, etc.). We will further analyse whether adding the semantics can boost the performance of the classifier model in such a scenario. Our hypothesis for using the semantics in cross crises domain classification is that adding the concepts and properties of entities (e.g., type of an entity, category of an entity, hypernyms) improves the classifier's adaptability in identifying crisis information content across different crises domains, by creating a non-event specific contextual semantic abstraction of crisis-related content. The contributions of the work done in this chapter can be summarised as follows:

- Build a hybrid statistical-semantic classification model with semantics extracted from two knowledge bases: *BableNet* and *DBpedia*.
- Conduct the experiments for classifying relevancy of tweets from 26 crisis events of various types.
- Create classification models with multiple combination of features.
- Analyse the classifier models when crisis *types* are included/excluded from the training data.

- Demonstrate that adding the *semantic features* increases the classification accuracy on unseen crisis *types* by +7.2% in F1 in comparison to non-semantic models.

The rest of the chapter is organised as follows: Section 4.2 elaborates on our classification approach. Section 4.2.1 describes the dataset used, and selection of the labelled data and events. Section 4.2.2 describes the feature engineering, and types of features: *statistical* and *semantic*. Section 4.3 details our experimental set up and results. We discuss the findings in section 4.4 and summarise the work in section 4.5.

4.2 Semantic Classification of Crisis Relevancy Across Crises Types

To create an automated crisis relevancy binary classification model which is adaptive across crises types, we require a labelled dataset spanning across various crises *types*, various *statistical* and *semantic* features, and a machine learning classification algorithm. In the following sub-sections, we present (i) the dataset used for training and testing the classifiers in Section 4.2.1, (ii) the statistical and semantic set of features used for building the classifiers in Section 4.2.2, and (iii) the classifier selection process in Section 4.2.3.

4.2.1 Dataset

As in the previous chapter, we use the CrisisLexT26 dataset* (Olteanu et al., 2015) for this study as well. The data contains 1000 labelled tweets for each of the 26 different crises events in the following categories: *'Related and Informative', 'Related but not Informative', 'Not Related' and 'Not Applicable'*. For this study, we selected all 26 crises events. To create a binary classification system we merged the *Related and Informative* and *Related but not Informative* into the *Related* class, and merged the *Not Related* with *Not Applicable* to create the *Not Related* class.

*CrisisLexT26 http://crisislex.org/data-collections.html#CrisisLexT26

Table 4.1: Crisis events data, balanced between related and not-related classes

			Category						Category		
Nb.	Id	Event	Related	Not-Related	Total	Nb.	Id	Event	Related	Not-Related	Total
1	CWF	Colorado Wildfire	242	242	484	2	COS	Costa Rica Earthquake	470	470	940
3	GAU	Guatemala Earthquake	103	103	206	4	ITL	Italy Earthquake	56	56	112
5	PHF	Philippines Flood	70	70	140	6	TYP	Typhoon Pablo	88	88	176
7	VNZ	Venezuela Refinery	60	60	120	8	ALB	Alberta Flood	16	16	32
9	ABF	Australia Bushfire	183	183	366	10	BOL	Bohol Earthquake	31	31	62
11	BOB	Boston Bombing	69	69	138	12	BRZ	Brazil Nightclub Fire	44	44	88
13	CFL	Colorado Floods	61	61	122	14	GLW	Glasgow Helicopter Crash	110	110	220
15	LAX	LA Airport Shoot	112	112	224	16	LAM	Lac Megantic Train Crash	34	34	68
17	MNL	Manila Flood	74	74	148	18	NYT	NY Train Crash	2	1	3
19	QFL	Queensland Flood	278	278	556	20	RUS	Russia Meteor	241	241	482
21	SAR	Sardinia Flood	67	67	134	22	SVR	Savar Building	305	305	610
23	SGR	Singapore Haze	54	54	108	24	SPT	Spain Train Crash	8	8	16
25	TPY	Typhoon Yolanda	107	107	214	26	WTX	West Texas Explosion	81	81	162

Further, we reduced the data redundancy by removing the replicated instances of the tweets by comparing each tweet in pairs after removing the user-handles (i.e., '@' mentions), URL's, and special characters. After removing duplicates there were 21378 documents (tweets) annotated with the *Related* label and 2965 annotated with the *Not Related* label. This distribution between the two classes was highly skewed. Thus, to avoid classification bias caused by the imbalance in the size of the classes in the data, we balanced the data by matching the number of *Related* documents with the *Not Related* ones across each crisis event. Following this, the overall size of the data resulted in 5931 tweets (2966 *Related* and 2965 *Not Related* documents). Table 4.1 shows the distribution of selected tweets for each event.

4.2.2 Features

As in the previous chapter, we generate two types of features for the binary classification problem to classify the tweets as crisis *related* and *not related*: *statistical features* and *semantic features*. The importance of *statistical features* was highlighted in Chapter 2, and was also used in addressing research question RQ1 in Chapter 3. For research question RQ2 addressed in this chapter, we use the *Statis-*

tical Features as the baseline approach as well. The statistical features contain the linguistic properties and quantifiable properties of the text. In the *Semantic Features* we use two types of semantic features from two knowledge base, i.e., *BabelNet* and *DBpedia*. Further, we provide more details about both type of features.

Statistical Features

For every tweet in the dataset, we use the same statistical features as those in Chapter 3:

- *Number of nouns*: Nouns generally refer to entities such as location, person, and organisations involved in the scope of crisis event (Imran et al., 2013a; Imran et al., 2013b; Stowe et al., 2016).
- *Number of verbs*: Verbs can indicate that an action is being undertaken or occurring in the course of a crisis event (Imran et al., 2013a; Imran et al., 2013b; Stowe et al., 2016).
- *Number of pronouns*: Much like nouns, pronouns may also refer to the actors, locations, or resources that are named in a given text posted during the crisis event.
- *Tweet Length*: Total number of characters in a given post. The length of a post may be related to the amount of information contained in it (Imran et al., 2013a; Imran et al., 2013b; Sakaki et al., 2010).
- *Number of words*: Similar to the length of the post, number of words may also be an indicator of the amount of information present in the post (Imran et al., 2013a; Karimi et al., 2013).
- *Number of Hashtags*: Hashtags are social media specific features, which often indicate the themes of the post and are manually generated by the posts' authors. The presence or absence or the number of hashtags can be important discriminatory features (Imran et al., 2013a; Imran et al., 2013b; Karimi et al., 2013).

- *Unigrams*: Unigrams provide a keyword-based representation of the content of the posts, thus enabling a vector based representation of the overall data (Imran et al., 2013a; Imran et al., 2013b; Karimi et al., 2013; Li et al., 2012b; Zhang & Vucetic, 2016; Sakaki et al., 2010).

To extract the Part of Speech (POS) features we used the spaCy library*. The tokens were converted to lower case. Stop-words were removed using a stop-word list†. The tokens were stemmed using the Porter Stemmer. Converted the data to the unigrams using the regexp tokeniser provided in the NLTK library‡. In the end, we applied the TF-IDF normalisation on the tokens to weigh the important words (tokens) in the data as per the relative importance within the entire dataset. This process resulted in the generation of a total of 10757 unigrams (total vocabulary size) for the overall balanced data.

Semantic Features

As explained in earlier chapters, *semantic features* are aimed at generalising the crises information representation across the data. In this case, we hypothesise that the semantic features can generalise the information across various *types* of crises events. For this work, we extracted the named entities using Named Entity Recogniser (NER) service Babelfy,§ and used two knowledge bases for expanding the semantics: (1) BabelNet,¶ and; (2) DBpedia‖

- *Babelfy Entities and BabelNet Senses (English)*: the NER and word sense disambiguation service Babelfy built on top of BabelNet extracted the entities (e.g., *news*, *sadness*, *terremoto*). For each of these entities (returned by Babelfy in the form of Synset IDs), we extract the associated *English* sense/labels from BabelNet (e.g., *news*→*news*, *sadness*→*sadness*, *terremoto*→*earthquake*).

*SpaCy Library, https://spacy.io
†Stop Words List, https://raw.githubusercontent.com/6/stopwords-json/master/stopwords-all.json
‡Regexp Tokenizer (NLTK), http://www.nltk.org/_modules/nltk/tokenize/regexp.html
§Babelfy, http://babelfy.org
¶BabelNet, http://babelnet.org.
‖DBpedia, http://dbpedia.org.

- *BabelNet Hypernyms (English)*: we extract the English sense/label of all the direct hypernyms (at distance-1), of each annotated entity, from BableNet. Hypernyms, by their nature, can broaden the context of an entity, thereby enhancing the semantics of a document (e.g., *broadcasting*, *communiucation*, *emotion*).

- *DBpedia Properties*: Babely returns a DBpedia URI for each annotated entity (if available). We extract a list of properties associated with each DBpedia URI* by querying the SPARQL endpoint: `dct:subject`, `rdfs:label` (only in English), `rdf:type` (only of the type http://schema.org and http://dbpedia.org/ontology), `dbo:city`, `dbp:state`, `dbo:state`, `dbp:country` and `dbo:country` (the location properties fluctuate between `dbp` and `dbo`) (e.g., `dbc:Grief`, `dbc:Emotions`, `dbr:Sadness`).

In the previous chapter we saw that hypernyms enriched the context of the text by adding the semantics. The text documents with different entities but with similar hypernyms can be correlated. Consider the following entities *fireman*, *policeman*, *MP (Military Police)*, and *garda* (an Irish word for police). These four entities, while uniquely different in their morphological representation, share a common English hypernym: *defender*.

Also, there are multilingual tweets in the datasets, and formulating the semantics in *English* helps in preventing the data sparsity which might, otherwise, result from diverse morphological forms of entities and concepts across different languages (refer to Table 4.2 to see an example). The entity senses and hypernyms are extracted from BabelNet. The semantic expansion via *BabelNet semantics* resulted in vocabulary expansion by an additional 3057 unigrams (in comparison to statistical features).

DBpedia properties were extracted to obtain more information for each entity which were reflected by subject, label, and location specific properties. The semantic expansion via *DBpedia semantics* expanded the vocabulary size by 1733 unigrams (in comparison to statistical features).

*Ontology Namespaces: `dct`: http://purl.org/dc/terms/; `dbo`: http://dbpedia.org/ontology/; `dbp`: http://dbpedia.org/property/; `rdf`: http://www.w3.org/1999/02/22-rdf-syntax-ns#; `rdfs`: http://www.w3.org/2000/01/rdf-schema#

Table 4.2: Semantic expansion with BabelNet and DBpedia semantics.

	Post A	Post B
Feature	*'Sad news to report from #Guatemala -at least 8 confirmed dead, possibly more, by this morning's major earthquake.'*	*'Terremoto 7,4 Ricther Guatemala deja 15 fallecidos,casas en el suelo, 100 desaperecidos, 100MIL personas sin luz FO'*
Babelfy Entities	*news, sadness, dead, describe, earthquake*	*terremoto, casas, suelo, luz, fallecidos*
BabelNet Sense (English)	*news, sadness, dead, describe, earthquake*	*earthquake, house, soil, light, dead*
BabelNet Hypernyms (English)	*broadcasting, communication, emotion, feeling, people, deceased, inform, natural disaster, geologica phenomenon*	*natural disaster, geological phenomenon, building, Structure, residential_building granular material, people, deceased*
DBpedia Properties	`dbc:Grief`, `dbc:Emotions`, `dbr:Sadness`, `dbc:Demography`, `dbr:Death`, `dbc:Communication`, `dbr:News`, `dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`	`dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`, `dbc:Home`, `dbc:Structural_system`, `dbc:Light`, `dbr:Death`, `dbc:Demography`

To create a binary classification system for classifying crisis *related* information across crises types, we use both types of semantics features *BabelNet semantics* (SemBN) and *DBpedia semantics* (SemDB) as individual semantic features, and in combination as well (SemBNDB). Combination of both types of semantic features resulted in vocabulary expansion of 3824 unigrams.

4.2.3 Classifier Selection

For a binary classification problem, the rationale behind the choice of classification algorithm was explained in Chapter 3. Combination of various features resulted in a high dimensionality, in the range of 10-15k, in comparison to the relative size of the training data (around 6000). Considering this high dimensionality and the need to avoid over fitting, we opted for Support Vector Machine (SVM) with a Linear Kernel as the classification algorithm. SVM has been found effective for such classification tasks*.

In addition, we re-validated the suitability of SVM Linear Kernel to our task in comparison to RBF kernel, Polynomial kernel, and Logistic Regression. By performing 20 runs of 5-fold cross-validation of different feature combination, we found that SVM Linear Kernel was statistically significant with a higher mean F_1 value of 0.8118 and a p-value of < 0.00001 (via 2 tailed t-test). The codebase and data generated in this chapter is accessible from the shared Github repository†.

4.3 Experiments

In this section, we elaborate on the experimental set up, combination of features for creating the classification models, and selection of different *types* of events for training and test data for cross-crisis classification scenario.

*A Practical Guide to Support Vector Classification, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

†https://github.com/pkhare/crisc_codebase

4.3.1 Experimental Setup

We design the experiments in two scenarios:

- *Crisis Classification Models*: For the first experiment, we train and evaluate the classification models on the entire dataset comprising of all the 26 crisis events (Table 4.1). We create various models, by combining different features, and aim to analyse whether the impact of semantics boosts the binary classification. We perform this on the entire data and validate the model through numerous iterations of 5-fold cross validation. To this end, we used scikit-learn library* for the task. The various classification models based on different feature combinations are as follows:
  - **SF**: This classification model is built on the *statistical features* only. This also happens to be the baseline model for these experiments.
  - **SF+SemBN**: This classification model is built on the combination of *statistical features* and the semantic features from *BabelNet Semantics* (entity sense, and their hypernyms-in English).
  - **SF+SemDB**: This classification model is built on a combination of the *statistical features*, and the semantic features from *DBpedia Semantics* (label, type, and other DBpedia properties).
  - **SF+SemBNDB**: This classification model is built on a combination of the *statistical features*, and the semantic features from both *BabelNet and DBpedia Semantics*.
- *Cross-Crisis Classification*: In the second scenario of the experiment design, we aim to evaluate the models on *types* of events which were not observed in the training data. For instance,

*Scikit-learn, http://scikit-learn.org

training the model on data from *flood* type events, and testing it on data from *earthquake* type events. The models are again created based on a combination of different features as shown in the above experiment design. However, in this case we define the training and test data based on two different criteria:

A. Identify crisis *related* posts from a crisis event, when the *type* of event is already included in the training data (e.g., apply the model on tweets from a new *flood* type incident when tweets from other *flood* type crises are in the training data).

B. Identify crisis *related* posts from a crisis event, when the *type* of the event is not included in the training data (e.g., apply the model on tweets from a new *flood* type incident when training data do not contain tweets from any other *flood* type crises).

The criteria in the *Cross-Crisis Classification* are based on the *type* of events. To enable such an analysis, we distributed the 26 crisis events broadly in 11 *types*, as shown in Table 4.3. The categorisation of the events into types is based on personal understanding of the nature of any given crisis event, and how related the events might be based on their effects. For example, *floods* and *typhoons* are quite similar considering that typhoons often result in floods.

4.3.2 Results: Crisis Classification

We report the results from the first experiment where we perform 20 iterations of 5-fold cross validation on each feature model on the entire dataset (26 crises events across all 11 event types). Table 4.4 presents the results and reports the mean of *Precision* (P_{mean}), *Recall* (R_{mean}), and F_1 score (F_{mean}) from 20 iterations of 5-fold cross validation, the standard deviation in F_1 score distribution (σ), and percentage change of F_1 score compared to the baseline ($\Delta F/F$).

*NYT has only 3 tweets in total.

Table 4.3: Types of events in the dataset

Event Type (Nb.)	Event Instances	Event Type (Nb.)	Event Instances
Wildfire/Bushfire (2)	CWF, ABF	Haze (1)	SGR
Earthquake (4)	COS, ITL, BOL, GAU	Helicopter crash (1)	GLW
Flood/Typhoon (8)	TPY, TYP, CFL, QFL, ALB, PHF, SAR, MNL	Building collapse (1)	SVR
Terror Shooting/Bombing (2)	LAX, BOB	Location Fire (2)	BRZ, VNZ
Train crash (2)	SPT, LAM*	Explosion (1)	WTX
Meteor (1)	RUS		

Table 4.4: Crisis-related content classification results using 20 iterations of 5-fold cross validation, $\Delta F/F$ (%) shows percentage gain/loss of the statistical semantics classifiers against the statistical baseline classifier.

Model	P_{mean}	R_{mean}	F_{mean}	Std. Dev ()	$\Delta F/F$ (%)	Sig. (p-value)
SF (Baseline)	0.8145	0.8093	0.8118	0.0101	-	-
SF+SemBN	0.8233	0.8231	0.8231	0.0111	1.3919	$<$ 0.00001
SF+SemDB	0.8148	0.8146	0.8145	0.0113	0.3326	0.01878
SF+SemBNDB	0.8169	0.8167	0.8167	0.0106	0.6036	0.000011

In Table 4.4 we can see that the semantic feature classifiers show a gain in F_{mean} in comparison to the baseline classifier, although very small. A noticeable gain (improvement) against the baseline classifier is observed in SF+SemBN (1.39%) and SF+SemBNDB (0.6%). Both the improvements from SF+SemBN (1.39%) and SF+SemBNDB (0.6%) are found to be statistically significant ($p < 0.05$) based on a 2-tailed one-sample t-test, where the F_{mean} of SF is treated as the null-hypothesis. A t-test can be evaluated as:

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{4.1}$$

$\bar{x}$ is the mean (F_{mean}) of the 100 results from each classifier, μ is the mean of the null hypothesis (which is the F_{mean} of the baseline classifier), s is the standard deviation of the sample (which is the new classifier, other than the baseline), and n is the size of the sample (100 results).

4.3.3 Results: Cross-Crisis Classification

In this section we look at the results when different classification models had to deal with the *type* of crisis events. As described in Section 4.3.1, we have set up 2 criteria for evaluating the classifiers: (i) when the model has seen the *type* of the tested event in the training data (Criteria 1), (ii) when the model has not seen the *type* of the tested event in the training data (Criteria 2).

Criteria 1 - Already seen event types

In this sub-task, the models were evaluated on a new crisis event instance of an event *type* which already existed in the training data of the model. For example, we evaluated a new *earthquake* type event on a model which was trained on a data that contained other *earthquake* type events. In this task, we train each classifier on 25 crisis events out of 26 events, and use the 26^{th} event as a test event data. To carry out the evaluation, we select the following event types and events as test data events:

- *Flood/Typhoons* - Typhoon Yolanda (TPY), Typhoon Pablo (TYP), Alberta Flood (ALB), Queensland Flood (QFL), Colorado Flood (CFL), Philippines Flood (PHF) and Sardinia Flood (SAR).
- *Earthquake* - Guatemala Earthquake (GAU), Italy Earthquake (ITL), Bohol Earthquake (BOL) and Costa Rica Earthquake (COS).

Table 4.5: Cross-crisis relatedness classification: criteria 1 (best F_1 score is highlighted for each event).

Test event	Instances		SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
	Train	Test	P	R	F_1	P	R	F_1	$\Delta F/F$ (in %)	P	R	F_1	$\Delta F/F$ (in %)	P	R	F	$\Delta F/F$ (in %)
TPY	5717	214	0.808	0.804	0.803	0.777	0.776	0.776	-3.44	0.772	0.771	0.771	-4.01	0.780	0.780	0.780	-2.83
TYP	5755	176	0.876	0.864	0.863	0.853	0.841	0.840	-2.66	0.831	0.83	0.829	-3.84	0.861	0.852	0.851	-1.29
ALB	5899	32	0.72	0.719	0.718	0.754	0.75	0.749	4.25	0.845	0.844	0.844	17.41	0.845	0.844	0.844	17.41
QFL	5375	556	0.791	0.784	0.783	0.80	0.793	0.792	1.18	0.780	0.772	0.77	-1.66	0.789	0.782	0.781	-0.22
CFL	5809	122	0.82	0.803	0.801	0.835	0.828	0.827	3.28	0.806	0.762	0.754	-5.88	0.796	0.77	0.765	-4.41
PHF	5791	140	0.764	0.764	0.764	0.769	0.764	0.763	-0.13	0.772	0.771	0.771	0.93	0.744	0.743	0.743	-2.83
SAR	5797	134	0.684	0.612	0.570	0.747	0.694	0.677	18.79	0.702	0.664	0.648	13.70	0.696	0.664	0.650	14.10
GAU	5725	206	0.788	0.782	0.780	0.739	0.728	0.725	-7.1	0.798	0.786	0.784	0.51	0.779	0.772	0.770	-1.30
ITL	5819	112	0.595	0.589	0.583	0.619	0.589	0.562	-3.58	0.667	0.634	0.615	5.49	0.659	0.616	0.588	0.98
BOL	5869	62	0.743	0.742	0.742	0.732	0.726	0.724	-2.38	0.758	0.758	0.758	2.20	0.684	0.677	0.674	-9.07
COS	4991	940	0.794	0.790	0.790	0.773	0.770	0.770	-2.56	0.740	0.739	0.739	-6.42	0.751	0.750	0.750	-5.08
Avg.					0.745			0.746	0.52			0.753	1.67			0.745	0.50
SDV					0.091			0.077	6.89			0.068	7.78			0.079	8.05

For example, to understand this list of event types and events, consider the case when test event is *Typhoon Yolanda (TPY)*, the classification model is trained on the rest of 25 crisis events. It is understood from the events list (see Table 4.1) that there are multiple *typhoons* and *floods* events in the dataset.

The results for Criteria 1 are shown in Table 4.5. We observe that the improvement shown by the semantic feature classifiers is small and inconsistent across the test cases. SF+SemBN improves over the baseline in 4 out of 11 test cases. SF+SemDB improves over the baseline in 6 out of 11 test cases. On an average the percentage gain ($\Delta F/F$) ranges between +0.52% (SF+SemBN) and +1.67% (SF+SemDB) with a standard deviation varying between 6.89% to 7.78%. This shows that when the model has already seen a *type* of crisis in the training data, the semantic features are not too responsive in improving classifier's performance significantly, against the baseline statistical features model's F_1 score.

Criteria 2 - Unseen event types

Unlike Criteria 1, in this task we ensure that the classification model is validated and applied to the *type* of crisis events that it is not trained on, i.e., not seen in the training data. To ensure that the training dataset never observes any data from a crisis *type* on which the classification model is applied, the training and the test data sets are always selected from different *types* of crises. For carrying out the evaluation, we select the following event types and events as training and test data events:

- Training data: All the event *types* excluding *Terror Shooting/Bombing and Train Crash*. Test Data: *Los Angeles Airport Shooting (LAX), Lac Megantic Train Crash (LAM), Boston Bombing (BOB)*, and *Spain Train Crash (SPT)*. All the test events are *shooting/bombing/train crash* type of crises.

- Training data: All the event *types* excluding *Flood/Typhoon*. Test Data: *TPY, TYP, ALB, QFL, CFL, PHF,* and *SAR*. All the test events are *flood/typhoon* type of crises.

- Training data: All the event *types* excluding *Earthquake*. Test Data: *GAU, ITL, BOL,* and *COS*. All the test events are *earthquake* type of crises.

The results for Criteria 2 are shown in Table 4.6. From the table we can see that the average best performance is exhibited by the *DBpedia semantics* model SF+SemDB, with an average gain of +7.2% (with a Std. Dev. of 12.83%) in F_1 score ($\Delta F/F$) over the baseline SF classifier. The SF+SemDB shows improvement in F_1 score, against the baseline, in 10 out of 15 test cases. Out of the remaining 5 test cases where the improvement is not seen, in 2 test cases the percentage loss in F_1 score ($\Delta F/F$) is -0.034% and -0.56%.

The combination of both the semantic features and the statistical features SF+SemBNDB model produced an improvement, over the baseline, in 9 out of 15 test cases with an average percentage gain of +2.64% in F_1 score. Comparing this with Criteria 1, semantic features (particularly *DBpedia*

semantics) improve the classifier's performance significantly and consistently, over the statistical features, when the model has not seen the *type* of test event in the training data. This shows that while semantics may not improve the classifier's performance much when the type of events in the training and testing data are the same, however, semantic feature appear to be making a significant impact when the model is applied to a totally new *type* of crisis, that the model was not trained on. This makes semantic feature-based models more appropriate for such situations, where the model needs to be applied to a new type of crisis.

Table 4.6: Cross-crisis relatedness classification: criteria 2 (best F_1 score is highlighted for each event).

	Instances		SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Test event	Train	Test	P	R	F_1	P	R	F_1	$\Delta F/F$ (in %)	P	R	F_1	$\Delta F/F$ (in %)	P	R	F	$\Delta F/F$ (in %)
LAX	5407	224	0.664	0.656	0.652	0.681	0.679	0.677	3.90	0.666	0.665	0.665	1.95	0.657	0.656	0.656	0.58
LAM	5844	68	0.655	0.632	0.618	0.642	0.632	0.626	1.2	0.619	0.618	0.616	-0.34	0.638	0.632	0.628	1.62
BOB	5407	138	0.669	0.630	0.608	0.663	0.645	0.635	4.40	0.613	0.609	0.605	-0.56	0.628	0.616	0.607	-0.19
SPT	5844	16	0.573	0.563	0.547	0.690	0.688	0.686	25.56	0.767	0.750	0.746	36.5	0.69	0.688	0.686	25.56
TPY	4409	214	0.714	0.664	0.642	0.715	0.640	0.606	-5.67	0.69	0.664	0.651	1.39	0.676	0.617	0.582	-9.45
TYP	4409	176	0.769	0.699	0.678	0.802	0.705	0.679	0.12	0.742	0.682	0.661	-2.54	0.733	0.642	0.603	-10.99
ALB	4409	32	0.727	0.719	0.716	0.771	0.719	0.705	-1.63	0.833	0.813	0.81	13.02	0.742	0.719	0.712	-0.63
QFL	4409	556	0.734	0.694	0.681	0.728	0.676	0.657	-3.51	0.733	0.707	0.698	2.58	0.741	0.707	0.696	2.23
CFL	4409	122	0.792	0.779	0.776	0.736	0.713	0.7060	-9.04	0.707	0.705	0.704	-9.27	0.755	0.754	0.754	-2.87
PHF	4409	140	0.589	0.564	0.532	0.672	0.607	0.566	6.52	0.662	0.643	0.632	18.9	0.617	0.586	0.556	4.67
SAR	4409	134	0.663	0.590	0.537	0.660	0.597	0.553	2.93	0.658	0.619	0.595	10.69	0.691	0.642	0.617	14.84
GAU	4611	206	0.610	0.553	0.487	0.584	0.549	0.495	1.62	0.692	0.650	0.630	29.39	0.667	0.621	0.593	21.79
ITL	4611	112	0.546	0.536	0.509	0.632	0.571	0.516	1.26	0.633	0.589	0.553	8.54	0.661	0.598	0.555	8.93
BOL	4611	62	0.732	0.726	0.724	0.656	0.645	0.639	-11.73	0.684	0.677	0.674	-6.86	0.606	0.597	0.588	-18.77
COS	4611	940	0.595	0.560	0.515	0.626	0.554	0.480	-6.71	0.618	0.578	0.538	4.56	0.645	0.580	0.527	2.33
Avg.					0.615			0.615	0.61			0.652	7.2			0.624	2.64
SDV					0.090			0.076	8.66			0.071	12.83			0.065	11.74

4.3.4 Feature Analysis

In order to gain a better understanding of the impact of the semantic features on the classification models, we analysed the most informative features across the designed statistical feature model and two semantic feature models. The informative features were derived by calculating the Information Gain (IG) score for all the features in each of the three model scenario (over the entire data across 26 crisis events, see Table 4.1). A list of top informative features (IG score) across different models is shown in Table 4.7. We notice very event-specific features in the statistical feature SF model, such as *collapse, terremoto, fire, earthquake, #earthquake, flood, typhoon, injured,* and *quake*. In the top 50 features we observe at least 7 hashtags, which indicates that event specific vocabulary is more crucial for the classifier in determining the crisis relevancy of the tweet. This may impact the performance of classifiers when the data is from new types of crises and contains a different type of vocabulary.

We also observed that *No.ofHashtag* turned out to be a key statistical feature, across all the models. Exploring further, we found that out of 2966 crisis *related* tweets 1334 tweets contained zero hashtags (45% of the crisis *related* tweets), while only 15% of *not related* tweets had zero hashtags (471 out of 2965 tweets). In the two semantic models, i.e., SF+SemBN and SF+SemDB, concepts such as *natural_hazard, structural_integrity_and_failure, conflagration, geological phenomenon, perception, dbo:location, dbo:place, dbc:building_defect,* and *dbc:solid _mechanics* were seen to be amongst the IG score features (Table 4.7). When we looked deeper, we found that *Structural_integrity_and_failure* represented the annotated entity form for terms such as *building collapse* and *collapse*. These terms occurred often in several crisis events such as earthquake, floods, and building collapse. Since there are many floods and earthquake events, such semantics are expected to be informative. *Natural_disaster* is a hypernym and a *type* (DBpedia property) to several concepts occurring in the data such as *flood, landslide,* and *earthquake*. This shows that adding semantics not only homogenises the vocabulary, but also enhances the ability of the classification model to correctly identify crisis related content from

Table 4.7: IG-Score ranks of features for: SF, SF+SemBN and SF+SemDB.

	SF		SF+SemBN		SF+SemDB	
R.	IG	Feature	IG	Feature	IG	Feature
1	0.106	No.OfHashTag	0.106	No.OfHashTag	0.106	No.OfHashTag
2	0.046	costa	0.056	costa	0.044	No.OfNouns
3	0.044	No.ofNoun	0.044	No.OfNouns	0.036	costa_rica
4	0.044	rica	0.044	rica	0.035	dbc:countries_in_central_americ
5	0.035	collapse	0.036	costa_rica	0.035	collapse
6	0.033	terremoto	0.035	central_american_country	0.031	terremoto
7	0.026	TweetLength	0.032	collapse	0.027	dbo:place
8	0.025	7	0.031	terremoto	0.026	TweetLength
9	0.024	#earthquake	0.026	TweetLength	0.024	#earthquake
10	0.023	bangladesh	0.026	fire	0.024	dbo:location
11	0.022	No.OfVerb	0.024	#earthquake	0.023	dbo:populatedplace
12	0.022	#redoctober	0.023	structural_integrity_and_failur	0.023	dbc:safes
13	0.021	No.OfWords	0.023	coastal	0.022	structural_integrity_and_failure
14	0.018	tsunami	0.022	information	0.022	dbc:building_defect
15	0.017	fire	0.022	financial_condition	0.022	dbc:solid_mechanics
16	0.016	building	0.022	No.OfVerbs	0.022	dbc:engineering_failure
17	0.016	rt	0.022	#redoctober	0.022	bangladesh
18	0.015	factory	0.021	No.OfWords	0.022	dbc:flood
19	0.014	toll	0.020	shore	0.022	dbr:wealth
20	0.014	flood	0.020	building	0.022	No.OfVerbs
21	0.013	#bangladesh	0.019	anatomical_structure	0.021	No.OfWords
22	0.013	#colorad	0.019	phenomenon	0.02	dbc:coastal_geography
23	0.012	alert	0.018	natural_disaster	0.019	dbc:article_containing_video_clip
24	0.012	hit	0.018	failure	0.018	dbc:natural_hazard
25	0.012	typhoon	0.017	conflagration	0.017	fire

unseen events, by considering highly informative semantic features.

While the semantic models have not shown to significantly outperform the statistical feature models when the model is applied to already seen *types* of events, we have been able to demonstrate potential limitations of statistical feature models when they are applied to new unseen event types. It appears that the features in statistical feature models are quite tied to event specific features whereas semantic features overcome that limitation.

4.4 Discussion

In this chapter we demonstrate the impact of mixing the statistical features and the semantic features to address the problem of classifying crisis *related* content from new and unseen *types* of crisis events. Two sources of semantic features, *DBpedia Semantics* and *BabelNet Semantics*, were found to enhance the classifier's accuracy for most of the test case events. However, *DBpedia Semantics* were seen to be more consistent and significant in their impact, more likely due to a wider coverage of extracted semantics provided by DBpedia.

We analysed some of the tweets which were wrongly classified by either the statistical classifier (SF) or the semantic classifiers (SF+SemBN and SF+SemDB) in Criteria 1 and 2. We made the following observations: (i) semantic features tend to generalise the context in comparison to the event specific vocabulary as seen in the statistical feature models and thus are more adaptable to new *types* of events. For instance, the following tweet- *"EU, Canada release aid money for PHL flood victims: European Union and Canada are supporting Philippine efforts."* is a crisis-related tweet from Philippines Flood (PHF) which was wrongly classified by the statistical model. In the statistical SF model, none of the terms occurring in the text were observed to be informative features (based on IG-score). However, when *DBpedia semantics* were added to the training and test sets, in SF+SemDB feature classifier, properties such as `dbc:flood` and `dbc:weather_hazard` were found amongst the informative features.

These properties are related to *flood* and *money aid* in the original tweet; (ii) semantic features can also bring in too generic and broad entities/concepts and such features may not contribute towards improving a classifier's accuracy. Such features can be weakly discriminative features in the training data and can be found in tweets from both the classes; (iii) often the semantic extraction tools can yield rather non-relevant entities and thus expand the semantics towards irrelevant aspects. For instance, the following tweet- *"Scary. RT @AmyFreeze7: Super Typhoon in Philippines is 236 mph It's roughly the top speed of Formula 1 cars"* is from event *Typhoon Yolanda (TPY)*. The semantic feature based classifiers misclassified this tweet, while it was correctly classified by the SF classifier. A look into the features and information gain shows that the terms *typhoon* and *scary* were highly ranked feature in SF features' training data. On adding the semantics, the tokens in the post expanded to multiple related entities about *Formula 1*, which were not relevant to the crisis related features in the training data. This indicates the type of challenges which semantic expansion pose.

The experiments were performed across different crisis event *types*. One of the limitations of the event type distribution, as seen in Table 4.3, is the imbalanced number of events across each type. Some crisis *types* have more events than the others. The imbalanced distribution in the number of events (leading to number of tweets) across crises *types* could lead to classification bias. Having a wider range of crisis types with a higher number of events across each type, should help in making the classifiers more adaptive to various domains.

In this work, we considered that different crisis types are distinct from each other. The *type* of each crisis is basically the officially identified nature of any given event (e.g., flood, typhoon, earthquake). However, it is not a strict condition that different crises *types* will always yield uniquely distinct content, as there are chances of an overlap in the nature of the content. In our experiments, we have not taken into account the actual difference in the content that different crisis events generate, and rather only segregated the training and test data based on crisis types. Therefore, we cannot rule out the possibility that while the training and test data are distinct in their crisis *types*, there might be a certain

overlap or similarity in the content. We discuss the potential way to address this as a future work in Chapter 7 (*Discussion and Future Work*) in Section 7.2.

The work done in this chapter has contributed towards answering research question *RQ2*. The data used in this work originated in multiple languages. As a next step, we aim to analyse how the semantics or translation techniques can assist the classifiers to become adaptive to multilingual crisis data to identify crisis related information. This forms part of our research question *RQ3*, and we will expand our study to answer this research question in the next chapter.

4.5 Summary

The work done in this chapter is aimed towards answering research question *RQ2 - "To what extent could semantics improve Tweets classification for new types of crisis events?"*. We hypothesised that adding semantics in the form of entities, properties of entities (e.g., type of an entity, category of an entity, hypernyms) will enhance the ability of classification models to identify crisis *related* information in new *types* of crises events. We demonstrated this by creating mutiple classification models by merging *statistical features* with the *semantic features*. We created semantic features using two different external knowledge bases: *DBpedia* and *BabelNet*. To conduct the experiments we created two criteria: (i) apply the classifier to a new crisis *type* event, when the classifier has already seen another event of a similar *type* of crisis in the training data, (ii) apply the classifier to a new crisis *type* event, when the classifier has not seen a similar *type* of crisis in the training data. We observed that semantics, particularly *DBpedia semantics*, enhance the classifier's accuracy when applied to a new *type* of crises which was not seen in the training data. On average the *DBpedia* features, when combined with statistical features, show a performance of F1 score of around 0.652 (Table 4.6), which when compared to other parallel works on cross-domain crisis data classification (Pedrood & Purohit, 2018; Alam et al., 2018; Imran et al., 2016b) is noteworthy. For instance, the F-measure in the cross-domain model adapta-

tion (cross-crisis classification), in the work by Imran and colleagues (Imran et al., 2016b), which used textual features such as uni-grams, bi-grams, and part of speeches, varied between 0.22-0.58.

We also performed feature analysis and an initial error analysis to understand how the semantics played a role. The semantics make the vocabulary representation of the events more broader and less event specific. This results in an increment of the discriminative/informative property of such features (entities) which are more likely to be existing in multiple types of crises events, instead of event specific vocabulary (as observed in the statistical feature models).

We were able to demonstrate, and thus answer research question *RQ2*, that adding the semantics (particularly *DBpedia semantics*) to the statistical features in a binary classification model enhances the performance and helps the classifier in identifying crisis *related* information from new *types* of crises events.

5

Classifying Crisis Information Relevancy Across Multiple Languages

Crisis Data is multilingual in nature. Not only crisis events occur globally, resulting in online data sources from various languages, we also observe multilingual data getting generated within a single crisis event as well. Thus, language forms a very important aspect of creating automated classification

tools to identify crisis *related* information. In the previous chapters we explored the type of classifiers that can enhance the classification accuracy while identifying crisis *related* information from new/unseen crises events and crises *types*. In this chapter, we take the problem of crises data classification towards the *language* aspect of the data, where the classification models are analysed for their performance when applied to the data from a new unseen language. This chapter focuses on addressing the third research question -

- RQ3 - *To what extent could semantics improve crisis-relevancy classification of Tweets written in a new language?*

For instance, if a model is trained on crises data in *Italian*, how will the classifier perform when the test data, on which it is applied to, is in *English* or *Spanish*. This is the problem of cross-lingual classification of the crises data. We will determine whether or not adding the semantics or translating the test data to the language of the training data, or a combination of the translation and addition of semantics, would help the classification model tackle the problem of data in a new language.

5.1 Introduction

In Chapter 4 we demonstrated that adding semantics such as *DBpedia* properties enhanced the adaptability of the classification models to new *types* of crises, in order to identify crisis *related* information. The data used for creating and evaluating different classification models comprised of 26 crises events, which were spread across several crises *types* (we categorsied them in 11 different *types*). These events occurred at diverse geographical locations. It is evident that the overall dataset is multilingual in nature, which we did not fully investigate in the previous chapters while addressing research questions *RQ1* and *RQ2*. It is crucial to ensure the applicability of crises classification models to new languages for multiple reasons. Firstly, the data can always come in a new language, not only in the course of a new crisis in a different geographical location, but also within the same crisis events, data can occur in

multiple languages. Secondly, it is not feasible to train a new model for a new language every time due to lack of time and labelled data in real time. Also, it is infeasible to produce a model that is trained on all languages. Language adaptive classification tasks are, in general, NLP problems given the lack of sufficient data across languages to train the classification models on. In this chapter we aim to address the third research question:

> RQ3 - *To what extent could semantics improve crisis-relevancy classification of Tweets written in a new language?*

Imran and colleagues (Imran et al., 2016b) had shown that a classifier trained on data in *Italian* language is more likely to perform better when applied to posts in the *Spanish* language than in the *English* language. Although, their approach lacks in a rigorous cross-language analysis, as it focused on two *types* of events occurring in only two languages. Broadly following types of solutions to the problem of cross-linguality have been seen: (a) translation of the data/resource from one language to the target language, and then train the models (Araujo et al., 2016;Mihalcea et al., 2007); (b) using weakly-labelled data (without supervision) to build the models (Deriu et al., 2017); (c) using multilingual word representation using knowledge resources such as Wikipedia (Wick et al., 2016). To answer research question *RQ3* in this chapter, we test two similar approaches aimed to classify the cross-lingual crisis data for their *relevancy* in crisis situations: (a) translate the data to a single language; (b) use semantic features in English to supplement the training data of target language(s). We analyse how the addition of semantics and translation of the data to a common language amplifies the performance of classifiers, while dealing with data from a new language, for identifying crisis *related* information.

The contributions of the work done in this chapter can be summarised as follows:

- We generate hybrid statistical-semantic classification model by extracting semantics from two different knowledge bases: *DBpedia* and *BabelNet*.

- We conduct experiments to classify relevancy of tweets spanning across 30 crises events in 3 languages (English, Spanish, and Italian).
- We perform relevancy classification of tweets by translating them into a single language, as well as with performing the classification on the cross-lingual datasets.
- We are able to demonstrate that adding semantics enhanced the accuracy of cross-lingual classification by 8.26%-9.07% in average F_1 score when compared to the traditional statistical models.

The rest of the chapter is organised as follows: Section 5.2 elaborates on our classification approach. Section 5.2.1 describes the dataset used, and selection of the labelled data and events. Section 5.2.2 describes the feature engineering, and types of features: *statistical* and *semantic*. Section 5.3 details our experimental set up and results. We discuss the findings in section 5.4 and summarise the work in section 5.5.

5.2 Cross-Lingual Classification of Crisis Data

To build a language adaptive crisis relevancy binary classification model, we require a labelled dataset spanning across multiple languages to train the model on. As shown in the problems addressed in the previous chapters, we also require different *statistical* and *semantic* features, and a machine learning classification algorithm. In the following sub-sections, we present (i) the dataset used for training and testing the classifiers in Section 5.2.1, (ii) the statistical and semantic set of features used for building the classifiers in Section 5.2.2, and (iii) the classifier selection process in Section 5.2.3.

5.2.1 Dataset

To conduct this study, we chose multiple datasets from the CrisisLex library*. We shortlisted 3 data collections: CrisisLexT26 (Olteanu et al., 2015), ChileEarthquakeT1 (Cobo et al., 2015), and SOSItalyT4 (Cresci et al., 2015). We have used CrisisLexT26 in the previous chapters as well, it is a labelled dataset of tweets spanning across 26 different crisis events which occurred between 2012 and 2013. The dataset has 1000 labeled tweets for each events categorised into the following labels: *'Related and Informative', 'Related but not Informative', 'Not Related', and 'Not Applicable'*. Since these events occurred at diverse geographical locations around the world, they covered a range of languages. ChileEarthquakeT1 is a dataset of 2000 labelled tweets in Spanish collected during the 2010 Chilean earthquake). In ChileEarthquakeT1 all the tweets were labeled for their relatedness (*relevant* or *not relevant*). The SOSItalyT4 contains the set of labelled tweets for 4 different natural disasters (2 earthquakes and 2 floods) which occurred in Italy between 2009 and 2014. It contains almost 5.6k tweets labeled based on the type of information they convey ("damage", "no damage", or "not relevant"). As per the guidelines of the labeling (as provided by the authors), both *"damage"* and *"no damage"* indicated relevance of the tweet to the crisis.

Labelled tweets from all the 3 collections were considered. As in the previous chapters, we converged some of the labels into binary class labels, in order to create a system for binary classification as we are interested in classifying the tweets as crisis *related* or *not related*. The labels in CrisisLexT26 were merged as follows: *'Related and Informative'* and *'Related but not Informative'* were merged into the *Related* category, and *Not Related* and *Not Applicable* were merged into the *Not Related* category. The ChileEarthquakeT1 labeled dataset is already binary labeled for crisis relatedness. From the SOSItalyT4 data, we treated the tweets labelled as *damage* and *no damage* to the *Related* category (the original guidelines considered the label *no damage* as relevant to crisis but not indicating tweets

*CrisisLex, crisislex.org/

pertaining to damage), and tweets labelled as *not relevant* to the *Not Related* category.

Next, we filtered out duplicate instances from each dataset to prevent redundancy in the data and bias in the model. The duplicates were identified by matching the tweets, one by one, after removal of special characters, URLs, and user-handles (i.e., '@' mentions). As and when the two strings matched, we discarded the new one. Following this there were 21,378 *Related* and 2965 *Not Related* tweets in the CrisisLexT26 data set, 924 *Related* and 1238 *Not Related* in the Chile Earthquake data set, and 4372 *Related* and 878 *Not Related* in the SOSItalyT4 data set.

As a next step, we detected the language each tweet was written in. To this end, we used 3 different language detection APIs: detectlanguage[*], langdetect[†], and TextBlob[‡]. For each tweet, the language label was the one which was agreed by at least 2 of the language detection APIs. Following this, the entire data showed to have been constituted of more than 30 languages, where the major proportion of almost 92% of the tweets (29,141 out of 31755) were composed of English (en), Spanish (es), and Italian (it). Considering this aspect of the language distribution, we chose to focus our study on the tweets from these 3 languages. First, we created an unbalanced data set (unbalanced in terms of mutual distribution between the languages) by randomly selecting tweets across the 3 languages (see Table 5.1-*unbalanced*). We tested the unbalanced set, as in the real world scenario an imbalance in the data between different languages might occur. Further, we remove the imbalance across the languages and also the tweets for *Related* and *Not Related* classes, and create a balanced data with an equal distribution throughout, to avoid any kind of bias (Table 5.1- *balanced*).

Additionally, we also provide an overview of the language distribution across each crisis event in the original data sets, in the Table 5.2.

[*] https://detectlanguage.com
[†] https://pypi.python.org/pypi/langdetect
[‡] http://textblob.readthedocs.io/en/dev/

Table 5.1: Data size for English (*en*), Spanish (*es*), and Italian (*it*)

	Unbalanced		Balanced			
			Train		Test	
Language	Not Related	Related	Not Related	Related	Not Related	Related
English (*en*)	2060	2298	612	612	201	200
Italian (*it*)	813	812	612	612	201	200
Spanish (*es*)	1039	1124	612	612	201	200
Total	3912	4234	1836	1836	603	600

Table 5.2: Language Distribution (in %) in Crises Events Data

	Language (%)					Language (%)			
Event	*en*	*it*	*es*	Other	Event	*en*	*it*	*es*	Other
Colorado Wildfire	99.30	0	0.09	0.61	CostaRica Quake	45.67	1.96	44.03	8.33
Guatemala Quake	23.84	1.20	69.56	5.40	Italy Quake	18.53	71.10	9.70	0.77
Philippines Flood	91.31	0	0.98	7.71	Typhoon Pablo	81.22	0.22	4.40	14.17
Venezuela Refinery	8.93	0.22	89.8	1.06	Alberta Flood	99.48	0	0	0.52
Australia Bushfire	98.94	0.0	0.10	0.97	Bohol E'quake	86.5	0.12	0.12	13.25
Boston Bombing	93.22	0.21	2.12	4.34	Brazil Club Fire	31.6	0	1.79	66.61
Colorado Floods	99.67	0	0.11	0.22	Glasgow Helicopter	99.86	0	0.11	0.03
LA Airport Shoot	97.07	0.11	1.30	1.52	LacMegantic Train	52.57	0.21	1.16	46.06
Manila Flood	72.40	0.22	0.22	27.16	NY Train Crash	99.86	0.14	0	0
Queensland Flood	99.56	0.09	0	0.35	Russia Meteor	87.56	0.64	2.56	9.24
Sardinia Flood	10.93	88.49	0.12	0.46	Savar Building	86.90	0.82	5.19	7.09
Singapore Haze	97.47	0.0	0	2.53	Spain Train Crash	43.13	0	54.67	2.20
Typhoon Yolanda	91.59	0.11	1.83	6.47	Texas Explosion	94.99	0	3.00	2.01
L'Aquila Quake	4.89	88.58	1.43	5.10	Emilia Quake	1.02	87.99	0.34	10.65
Genova Flood	2.09	95.12	0	2.79	Chile Quake	10.82	0.19	82.00	6.99

5.2.2 Features

Similar to the work in the previous chapter, in this chapter as well we generate two types of features for the binary classification problem to classify the tweets as crisis *related* and *not related*: *statistical features* and *semantic features*. For the research question RQ3 addressed in this chapter, we treat the *Statistical Features* as a baseline approach. The statistical features represent the linguistic properties and quantifiable properties of the text. The *Semantic Features* are associated with the named entities and the hierarchical contextual information. Further, we elaborate on both types of features.

Statistical Features

For every tweet we extract the same statistical features seen earlier in Chapter 4, which can be referred to in the Section 4.2.2, under subsection Statistical Features.

As we are calculating the number of nouns, verbs, and pronouns, which represent the Part of Speeches (POS), we rely on the spaCy library* to extract the POS. To tokenise the data into unigrams, we use the regexp tokenizer provided in NLTK†. In order to remove the stop words a dedicated list‡ is used. We also convert the tokens to lower case. In the end, we perform the TF-IDF normalisation on the unigrams to weigh the relevance of tokens in the documents (tweets) as per their relative importance within the overall data. The data is represented as vectors. Following this, the vocabulary size of unigrams (for overall individual language data set in the balanced data) is as follows: *English (en)*-7495, *Spanish (es)*-7121, and *Italian (it)*-4882.

*SpaCy Library, https://spacy.io
†http://www.nltk.org/_modules/nltk/tokenize/regexp.html
‡https://raw.githubusercontent.com/6/stopwords-json/master/stopwords-all.json

Semantic Features

In this research problem, the *semantic features* are aimed towards generalising the data representation of the crises situations across the languages. The semantic features are extracted in a way in order to be more generic in context and less crisis-specific, thus addressing the problem of data sparsity. To this end, we extract the same semantic features, from BabelNet and DBpedia, as seen earlier in Chapter 4, and can be referred to in the Section 4.2.2, under subsection Semantic Features.

In the previous chapters we showed how semantics can bridge the contextual gaps between diverse concepts. By generalising the semantics to one language (in this case English), we overcome the challenge of data sparsity which might arise from different morphological forms of entities across different languages. As an example we can look at Table 5.3, where we compare the two tweets, originally in different languages, exhibiting similarities in context after inclusion of semantics. We also show how the Google translation service can impact the representation. Inclusion of semantic information, through *BabelNet Semantics* (entity sense and hypernyms), resulted in a vocabulary size of unigrams across each language data set as follows: *English (en)*-12604, *Spanish (es)*-11791, and *Italian (it)*-8544.

Similarly, extraction of DBpedia properties of the entities, as mentioned above, resulted in a vocabulary size of unigrams across each language data set as follows: *English (en)*-21905, *Spanish (es)*-15388, *Italian (it)*-10674. We analyse both sets of semantics features, *BabelNet* and *DBpedia*, individually with the statistical features as well as in combination to build a binary classification model.

5.2.3 Classifier Selection

In the problems explored in previous chapters, we highlighted the reasons of opting for the Support Vector Machine (SVM) with a Linear Kernel as the classification model such as high dimensionality of the data and avoiding the over-fitting problem. In the earlier chapters we also demonstrated that SVM Linear Kernel was statistically significant in performing better than other kernels such as RBF

Table 5.3: Semantic expansion with BabelNet and DBpedia semantics

	Post A	Post B
Feature	*'#WorldNews! 15 feared dead and 100 people could be missing in #Guatemala after quake http://t.co/uHNST8Dz'*	*'Van 48 muertos por terremoto en Guatemala http://t.co/nAGG3SUi vía @ejeCentral'*
Babelfy Entities	*feared, dead, people, missing, quake*	*muertos, terremoto*
BabelNet Sense (English)	*fear, dead, citizenry, earthquake*	*slain, earthquake*
BabelNet Hypernyms (English)	*geological_phenomenon, natural disaster, group*	*geological_phenomenon, natural disaster, dead*
DBpedia Properties	`dbr:Death`, `dbc:Communication`, `dbr:News`, `dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`	`dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`, `dbr:Death`
Google Translation	To *es-' #¡Noticias del mundo! 15 muertos temidos y 100 personas podrían estar desaparecidas en #Guatemala después terremoto http://t.co/uHNST8Dz'*	To *en-'48 people killed by earthquake in Guatemala http://t.co/nAGG3SUi vía @ejeCentral'*

kernel, Polynomial kernel, and Logistic Regression. In this case as well the training data instances varied between 1200-4500 through different experiments, and the dimensionality of the features ranged between 9000-20000. Also, Burel and colleagues (Burel et al., 2017a) discuss a better performance shown by SVM in comparison to other common methods such as classification and regression trees (CART) and Naive Bayes in related classification problems. They also discuss the near similar performance of SVM and CNN models in the classification of tweets. Based on this, we opted for SVM Linear Kernel for this task as well.

5.3 Experiments

In the following subsections we will provide details of the experimental set up where we create and validate multiple classification models based on statistical features, semantic features, and translation of the data.

5.3.1 Experimental Setup

We design the following classification models:

- *SF*: This model is built on only the statistical features; this model is our baseline.
- *SF+SemBN*: This model is built on the combination of statistical features with semantic features from *BabelNet* (entity sense, and their hypernyms in English, as explained in Section 5.2.2).
- *SF+SemDB*: This model is built on the combination of statistical features with semantic features from *DBpedia* (label in English, type, and other DBpedia properties).
- *SF+SemBNDB*: This model is built on the combination of statistical features with semantic features from both *BabelNet* and *DBpedia*.

Next, we define following experimental scenarios for applying and validating the classification models for cross-lingual crises data:

A. **Monolingual Classification with Monolingual Models:** In this scenario, the model is trained on data from one particular language, and then applied and validated on a test data in the same language. Here, we analyse the value of semantic features over the baseline classifier, when the language of the training and test data is the same.

B. **Cross-lingual Classification with Monolingual Models:** In this scenario, the model is trained on data from a particular language, and then applied and validated on a test data in a different language. For instance, the classifier is evaluated on crises data in *Italian*, while it is trained on crises data in *English* or *Spanish*.

C. **Cross-lingual Classification with Machine Translation:** In the third scenario, a model is trained on data in a particular language (say *Spanish)*, and then applied to crisis data which has been translated using automated tools from other language(s) (say *English* or *Italian*) to the language of the training data. For automated translation we use the Google Translation API*. In order to perform this experimental scenario, we translate the crises data in each of the three languages to other two languages one by one.

We perform all the experiments on both the (i) *unbalanced* data set, to analyse the scenario where the languages might be imbalanced in their distribution and (ii) on the *balanced* dataset, to analyse the scenario where there is no bias in the training of the classifier with regards to any language. It is to be noted, that the default reference to the results in the chapter is of the *balanced* data set, unless specifically mentioned about the *unbalanced* data set results. We report the results by providing the metrics *Precision* (P), *Recall* (R), F_1 score (F_1), and ΔF_1. We calculated *macro* values for these metrics,

*https://cloud.google.com/translate/

Table 5.4: Monolingual Classification Models – 5-fold cross-validation (best F_1 score is highlighted for each model). *en*, *it*, and *es* refer to English, Italian, and Spanish respectively.

Unbalanced Data (from Table 5.1-*unbalanced*)																
		SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Test	Size	P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en	4358	0.833	0.856	0.844	0.84	0.858	0.849	0.59	0.826	0.844	0.835	-1.07	0.829	0.845	0.836	-0.95
it	1625	0.703	0.721	0.711	0.712	0.714	0.713	0.28	0.696	0.706	0.701	-1.4	0.702	0.715	0.708	-0.42
es	2163	0.801	0.808	0.804	0.812	0.809	0.810	0.75	0.799	0.795	0.797	-0.87	0.798	0.798	0.798	-0.75
Avg.				0.786			0.791	0.54			0.778	-1.1			0.781	-0.71
Balanced Data (from Table 5.1-*balanced*)																
		SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Test	Size	P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en	1224	0.832	0.830	0.831	0.835	0.805	0.820	-1.32	0.835	0.799	0.816	-1.80	0.829	0.808	0.818	-1.56
it	1224	0.690	0.729	0.709	0.703	0.722	0.712	0.42	0.689	0.716	0.702	-0.99	0.708	0.718	0.712	0.42
es	1224	0.798	0.765	0.781	0.794	0.783	0.789	1.02	0.779	0.754	0.766	-1.92	0.780	0.773	0.776	-0.64
Avg.				0.774			0.774	0.04			0.761	-1.57			0.769	-0.59

which is an unweighted mean of metric for each label (in our case a balanced representation of two labels). ΔF_1 is the % change (gain or loss) in comparison the baseline- $\frac{(\textit{semantic model } F_1 - \textit{SF } F_1) * 100}{\textit{SF } F_1}$, where $\textit{SF } F_1$ is the F_1 score in SF model.

5.3.2 Results: Monolingual Classification with Monolingual Models

In this scenario, we tested the model on data from the same language as the model was trained on. To this end, we adopted a 5-fold cross validation approach and conducted the experiments across the indiviaul datasets of each language, i.e., *English*, *Italian*, and *Spanish*. Results, in the Table 5.4, indicate that when the language of the training and the test data are similar, the addition of *semantic features* does not impact the classification accuracy over the baseline model (SF model).

5.3.3 Results: Cross-lingual Classification with Monolingual Models

In this scenario, we tested the model on data from another language, as the model was trained on. In this case, we trained the model on one of the three languages, one by one, and evaluated them on the crisis data in the other two languages. From the results, as shown in the Table 5.5, it can be seen that the baseline model SF has an average F_1 score of 0.557. The addition of *semantic features* enhances the classification accuracy, in terms of ΔF_1, on an average by 8.26%-9.07%, with a standard deviation (SDV) between 10.9%-13.86% across all three semantic models and all test cases. On the *unbalanced* dataset the semantic models increase the classification accuracy, on an average, by 7.44%-9.78%. SF+SemDB shows an average gain in F_1 score (ΔF_1) of 8.71% with a standard deviation of 13.86% over the baseline (for the *balanced* data). While, the SF+SemBN shows an average gain in F_1 score (ΔF_1) of 8.26% with a standard deviation of 10.94% over the baseline.

5.3.4 Results: Cross-Lingual Crisis Classification with Machine Translation

In this scenario, we train the models on crises data in one of the three languages, and apply it to crisis data in the other two languages, but only after translating them into the language which the model has been trained on. We evaluate these models one by one across all three languages. For instance, when the training data is in Engslish (*en*), the Italian (*it*) posts are first translated to English (*it2en*) and then used as a test data. There are two aspects that we aim to analyse: (i) the impact of the *semantic features* on the classification of the translated data; and (ii) the performance of the classifiers on the translated data as compared to cross-lingual classification as seen in the previous section 5.3.3.

The results are presented in the Table 5.6, and average F_1 and % change ΔF_1 for the translated cases (*it2es*, *en2es*, etc.) are provided. SF+SemBN performs better over the baseline in 4 out of the 6 cases (when both the test and training data are both in the same language after translation). SF+SemDB

Table 5.5: Cross-Lingual Classification Models (best F_1 score is highlighted for each model).

Unbalanced Data (from Table 5.1- *unbalanced*)																	
		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en		4358															
	it	1625	0.576	0.522	0.417	0.598	0.562	0.518	24.2	0.595	0.576	0.553	32.6	0.609	0.588	0.568	36.2
	es	2163	0.674	0.633	0.604	0.663	0.654	0.645	6.79	0.653	0.649	0.643	6.46	0.649	0.641	0.633	4.8
it		1625															
	en	4358	0.469	0.474	0.449	0.547	0.545	0.538	19.82	0.508	0.508	0.504	12.25	0.516	0.516	0.516	14.9
	es	2163	0.635	0.610	0.586	0.643	0.627	0.612	4.43	0.601	0.60	0.596	1.70	0.625	0.620	0.614	4.78
es		2163															
	en	4358	0.633	0.62	0.604	0.60	0.572	0.532	-11.9	0.623	0.618	0.610	0.99	0.606	0.592	0.571	-5.46
	it	1625	0.536	0.533	0.521	0.529	0.529	0.528	1.34	0.526	0.526	0.526	0.96	0.539	0.539	0.539	9.78
Avg.					0.530			0.562	7.44			0.572	9.16			0.573	9.78
SDV					0.082			0.053	13.08			0.053	12.3			0.044	14.47
Balanced Data (from Table 5.1-*balanced*)																	
		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		P	R	F_1	P	R	F_1	ΔF_1	P	R	F_1	ΔF_1	P	R	F	ΔF_1
en		1224															
	it	401	0.539	0.515	0.429	0.588	0.571	0.549	28	0.569	0.568	0.568	32.4	0.578	0.576	0.572	33.3
	es	401	0.689	0.688	0.688	0.669	0.668	0.668	-2.9	0.647	0.644	0.641	-6.8	0.666	0.661	0.659	-4.2
it		1224															
	en	401	0.521	0.521	0.521	0.581	0.581	0.580	11.3	0.558	0.552	0.539	3.5	0.550	0.546	0.538	3.3
	es	401	0.655	0.646	0.640	0.672	0.655	0.647	1.1	0.638	0.636	0.635	-0.78	0.637	0.633	0.631	-1.4
es		1224															
	en	401	0.609	0.593	0.578	0.657	0.620	0.597	3.3	0.667	0.666	0.665	15	0.660	0.653	0.650	12.4
	it	401	0.529	0.522	0.489	0.534	0.534	0.532	8.8	0.551	0.546	0.533	9	0.555	0.551	0.543	11
Avg.					0.557			0.596	8.26			0.597	8.71			0.599	9.07
SDV					0.096			0.053	10.94			0.057	13.86			0.054	13.6

Table 5.6: Cross-Lingual Crisis Classification with Machine Translation (best F_1 score is highlighted for each event).

Unbalanced Data (from Table 5.1- *unbalanced*)																	
		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		*P*	*R*	F_1	*P*	*R*	F_1	ΔF_1	*P*	*R*	F_1	ΔF_1	*P*	*R*	*F*	ΔF_1
en		4358															
	it2en	1625	0.644	0.613	0.591	0.635	0.611	0.593	0.34	0.582	0.568	0.548	-7.27	0.597	0.580	0.561	-5.0
	es2en	2163	0.681	0.681	0.681	0.667	0.667	0.667	-2.0	0.669	0.661	0.659	-3.2	0.664	0.661	0.660	-3.1
it		1625															
	en2it	4358	0.609	0.601	0.588	0.636	0.618	0.597	1.53	0.570	0.570	0.569	-3.2	0.575	0.574	0.571	-2.9
	es2it	2163	0.647	0.629	0.612	0.675	0.636	0.607	-0.81	0.609	0.595	0.578	-5.5	0.620	0.603	0.583	-4.7
es		2163															
	en2es	4358	0.643	0.626	0.609	0.661	0.634	0.610	0.16	0.654	0.654	0.653	7.2	0.649	0.648	0.646	6.07
	it2es	1625	0.585	0.584	0.583	0.590	0.590	0.589	1.03	0.581	0.580	0.580	-0.51	0.586	0.585	0.584	0.17
Avg.					0.611			0.611	0.03			0.598	-2.1			0.60	-1.6
SDV					0.036			0.029	1.3			0.046	5.1			0.04	4.2
Balanced Data (from Table 5.1-*balanced*)																	
		Size	SF			SF+SemBN				SF+SemDB				SF+SemBNDB			
Train	Test		*P*	*R*	F_1	*P*	*R*	F_1	ΔF_1	*P*	*R*	F_1	ΔF_1	*P*	*R*	*F*	ΔF_1
en		1224															
	it2en	401	0.624	0.583	0.546	0.622	0.598	0.577	5.7	0.561	0.558	0.554	1.46	0.594	0.588	0.581	6.4
	es2en	401	0.675	0.671	0.669	0.704	0.696	0.693	3.6	0.701	0.671	0.658	-1.6	0.695	0.674	0.664	-0.74
it		1224															
	en2it	401	0.583	0.578	0.572	0.639	0.631	0.625	9.3	0.547	0.546	0.545	-4.7	0.551	0.551	0.551	-3.6
	es2it	401	0.638	0.621	0.609	0.703	0.668	0.653	7.2	0.619	0.603	0.590	-3.1	0.610	0.596	0.582	-4.4
es		1224															
	en2es	401	0.686	0.678	0.675	0.691	0.670	0.661	-2.0	0.691	0.691	0.691	2.3	0.683	0.683	0.683	1.2
	it2es	401	0.594	0.594	0.593	0.586	0.586	0.586	-1.2	0.580	0.576	0.570	-3.9	0.579	0.576	0.571	-3.7
Avg.					0.610			0.633	3.75			0.601	-1.59			0.605	-0.83
SDV					0.052			0.045	4.57			0.059	2.9			0.054	4.14

performs better over the baseline in 2 out of the 6 cases (when the test and training data are both in the same language after translation). On average, SF+SemBN shows an improvement over the baseline (SF) of 3.75% in F_1 score with a standard deviation (SDV) of 4.57%.

If we compare the translation models with the overall baseline for the cross-lingual classification, i.e., with the SF model (without translation) from Table 5.5, the SF+SemBN (with translation, Table 5.6) shows an average F_1 gain (ΔF) of 15.23% (with a standard deviation of 12.6%). Similarly, SF+SemDB (with translation, Table 5.6) shows an average F_1 gain (ΔF) of 9.82% (with a standard deviation of 14.6%) over the baseline SF (without translation, Table 5.5). Also, the SF (with translation)

shows an increment of 11.25% against the overall baseline SF (without translation) with a standard deviation of 13%.

5.3.5 Cross-Lingual Ranked Feature Correlation Analysis

In order to get a better understanding of how the addition of semantic features and translation of the data impacted the informative/discriminatory nature of the cross-lingual data, we performed a correlation analysis of the ranked features between the datasets of all three languages and across all the models. We took the entire balanced datasets of each language, used in each model (by merging the training and the test data into one). Following this, we calculated the Information Gain (IG) score over every discrete dataset, across all the 4 models (SF, SF+SemBN, SF+SemDB, SF+SemBNDB). To be reminded, that the datasets for the semantic models (SemBN, SemDB, SemBNDB) had semantics included in them. Also, the IG was calculated for the translated datasets as well (*en2es*, *it2es*, *en2it*, *es2en*, *es2it*, and *it2en*). This resulted in a ranked list of features, based on IG score, across each dataset.

Next, we take each pair of datasets, say English (*en*) and Spanish (*es*) (represented in the Table 5.7 as *en - es*), and determine the common ranked features in the ranked feature list based on $IGscore > 0$. This provides us two different ranked lists of common features with $IGscore > 0$. We calcualte the Spearman's Rank Order Correlation (ranges between $[-1, 1]$) between the two lists of ranked features. In the cases where the translation was applied, we considered the pairs where the second language was translated to the first language. For example, if English (*en*) and Italian (*it*) are to be correlated, we considered English (*en*) and Italian translated to English (*it2en*). Next, we repeat this process in the other order too, i.e., Italian (*it*) and English translated to Italian (*en2it*).

The analysis in Table 5.7, shows variations in the correlation across different datasets. Such variations can result from a number of factors. One of the key factors is the overlap of crises events in the data samples for each language. It is to be noted that while segregating the language data we only took the language into consideration, and did not take the discreteness of the crises events (Table 5.2)

into consideration. This could imply that under different datasets of the languages, there might be an overlap of some of the crises events. This could be observed in the correlation between *en-es*, which does not increase in the SF+SemBN and SF+SemDB, while just translating them mutually to each other's language does not impact the correlation much (slightly decreases). In fact, the highest correlation between *en-es* is observed in the SF model (without translation). This also justifies the better performance of the SF model (without translation), in comparison to the semantic models, in the cross-lingual classification (Table 5.5). In the SF model the correlation between *en-it* is ~-0.179, that reflects a near 'no correlation'. The addition of semantics enhances the correlation for *en-it* in both SF+SemDB and SF+SemBN, and also with the translation. Similarly for *es-it*, the correlation increases after semantic inclusion (which justifies the performance of semantic features in Table 5.5).

The greater correlation between discriminative features of data in different languages can be attributed to the addition of semantics in English (Section 5.2.2), which resulted in the cross-lingual vocabulary to match semantically as well as linguistically. It is important to be reminded again, that we considered features with $IGscore > 0$ to be more specific with the discriminatory features in cross-lingual datasets.

Translation also helped in an increased correlation between the, otherwise, cross-lingual data. This is an expected outcome for multiple reasons. Firstly, translation to the same language enables having similar features such as verbs, adjectives as well along with nouns across the datasets. Secondly, as we earlier mentioned that there is a potential overlap of different types of events covered in different languages such as *earthquakes* and *floods*, which can trigger a contextual overlap in the nature of the information.

Table 5.7: Spearman's Rank Order Correlation between ranked informative features (based on IG) across models and languages

Data/ Model	SF	SF+ SemBN	SF+ SemDB	SF+ SemBNDB	Translation	
en − es	0.573	0.385	0.349	0.373	0.515(en-es2en)	0.449(es-en2es)
en − it	-0.179	0.402	0.111	0.315	0.266(en-it2en)	0.594(it-en2it)
es − it	0.418	0.222	0.503	0.430	0.678(es-it2es)	0.612(it-es2it)

5.4 Discussion

In this chapter, to answer research question *RQ3*, we analysed the impact of adding the semantic features with the statistical features, and translating the data to a common language, in facilitating a cross-lingual crises data relevancy classification. The aim was to explore the methodologies which can assist in developing language-agnostic classification models. However, given the nature of the data, this analysis was limited to three languages: *English*, *Italian*, and *Spanish*. Gathering large scale annotated crises oriented data across several languages is challenging. One of the ways to address this could be translating the data to multiple languages using automated machine translation tools.

From the results, we see that when the data is cross-lingual (not translated) adding the semantics (both *DBpedia* and *BabelNet*) improve the cross-lingual classification accuracy in comparison to the baseline. Also, just translating the data also enhances the classification performance in comparison to the baseline SF (statistical features model). Adding the semantics after translation enhances the classification performance, but not by much. We can say that, if the data was to be translated then just the translation models might be sufficient. In the case where translation is not viable (if the translation accuracy is too inaccurate) then adding the semantics (without translation) can be recommended along with the statistical features for its higher accuracy. We have not evaluated the accuracy of translation services in our analysis.

In the work done in this chapter, the training and the test data were curated based on the languages, and it was quite natural for data across the languages to have a certain terminological overlap such as names of the crisis, locations, or people due to possible shared crises events. The scenario where languages and the crises events and/or event types are both discrete at the same time is another aspect of the problem, which we address in the next chapter.

While extracting the semantics from *BabelNet* (Section 5.2.2) we extracted them in English primarily. We found *BabelNet* (version 3.7) to be more enriched with English than with other languages. At the time of this work, it (BabelNet) recorded almost 17 million word senses in English, while the next highest number of word senses were noted to be in French with 7 million word senses*. Similarly in DBpedia, we found maximum instances per class (person, actor, athlete, politician, place etc.) in English†. Finding this bias towards *English* language, adding the semantics in English would have not only enabled us to extract maximum amount of labels/senses/properties but also helped in adding more concepts in one single language. Thus, also helping tackle the problem of data sparsity due to differences in the morphological forms of the languages. This resulted in an advantage gained by the semantic models over the purely statistical feature model.

For the cross-lingual classification, in this chapter, we performed 6 test cases. Extending such analysis to more languages will help in establishing the gains in classification performance observed by the semantic models over the baseline statistical features as statistically significant. As an alternative, multiple sets of train and test splits for each test case could also cater to the requirement of multiple iterations of experiments, which was not feasible in this particular study due to the limited overall size of the data. Although, we performed 10 iterations of 5-fold cross validation over the entire dataset and found SF+SemBN (*BabelNet Semantics*) better than rest of the models (particularly against the baseline with a statistically significant value of $p = 0.0192$, on a two-tailed t-test).

*BabelNet Statistics, http://live.babelnet.org/stats

†DBpedia Statistics, https://wiki.dbpedia.org/services-resources/datasets/dataset-statistics

The need for classification models being able to handle multiple languages is clear, since the language of information on social media during crises events varies significantly. Therefore, the ability of classification models to handle multi-lingual data is an advantage.

5.5 Summary

The work done in this chapter was aimed towards answering the research question *RQ3 - "To what extent could semantics improve crisis-relevancy classification of Tweets written in a new language?"*. We considered data from multiple crises events, and narrowed it down to data in three languages: *English*, *Italian*, and *Spanish*. We extracted statistical and semantic features for the data. We hypothesised that adding the semantic features could enahnce the similarity across data from different languages. Next, we created multiple classification models based on statistical and semantic features. The experiments for cross-lingual classification were designed in two ways: (i) training the model on crises data from a language and evaluating the model on crises data from a new language; (ii) training the model on crisis data from a language and evaluating the model on crises data from a new language but after translating the language of the test data to the language of the training data. This work explores the impact of semantic features in cross-lingual crisis data classification, which has not been explored in previous related works (Imran et al., 2016b; Li et al., 2018b; Lorini et al., 2019).

We were able to demonstrate, and thus answer research question *RQ3*, that models built on a combination of statistical and semantic features enhanced the classification accuracy when they were applied to crisis data from a new language. Also, the translation of the data to the same language enabled the classifiers to identify the crisis related information from a new language after translation.

The experiments conducted in this chapter have contributed towards answering research question *RQ3*. While answering research questions *RQ2* and *RQ3*, two distinct problems of crisis *types* and crisis data *language* have been explored respectively. However, in real world these two unique prob-

lems can co-occur, i.e., the classification model can encounter a previously unseen type of crisis and in a new language in the testing data. This forms part of our research questions *RQ4*, and we perform an in-depth study to answer this research question in the next chapter.

6

Classifying Crisis Relevancy Across Languages and Crisis Types

In the previous chapters we explored two discrete problems, (i) how the various classification models respond when they are trained on crises data in a certain language and applied to crises data in a new language; and (ii) how the various classification models respond when they are trained on data from a certain *type* of crisis and applied to data from a different *type* of crisis. So far, in the previous chapters

these two aspects of the crises data have been treated as distinct problems. However, these two aspects can co-occur, where the classification model is applied to a new *type* of crisis which contains data in a new language. In this chapter, we will address our last research question *RQ4*, where we determine what type of classification models are able to classify crises information relatedness when the *type* and *language* of the crisis event change at the same time.

6.1 Introduction

In Chapter 5, we explored how the classification models can respond to cross-lingual crises data. We saw that crisis events, across or within the same geographic locations, can result in multilingual data. Also, in Chapter 4 we saw that crises events can be of various *types*, depending on the nature of crisis event. In order to create classification models for crisis management/information systems, it is important to generate models that are adaptive to new crisis types and to data in new languages. So far, in the earlier chapters, we have shown how the inclusion of semantic features amplifies the classifier's accuracy in the scenario when the model is trained on certain types of crises events (e.g., floods) and evaluated on crises events of different types (e.g., earthquakes, fires). Similarly, we also showed the impact of semantic features inclusion and automated translation services in making classification models more language agnostic. What is yet to be explored is the performance of the classification models when these two problem scenarios exist together, i.e., when the data, on which the model is applied, is not only from a new crisis *type* but also in a new language. As discussed earlier in related works some of the works tried to address domain adaptation of crisis classification models on different crises events. Imran and colleagues (Imran et al., 2016b) considered crises data from two types of events *earthquakes* and *floods*, which was formed of data in more than one language. While the work did not exhibit a rigorous cross-lingual evaluation by limiting the analysis to only two crisis events in two languages, they did show that a classification model trained on data in *Italian* is more likely to

perform better when applied on crisis events in *Spanish* instead of *English*.

In this chapter, we aim to answer the fourth research question:

> RQ4 - *To what extent could semantics improve Tweets classification when the type of crisis event and language change?*

In this chapter, we deal with the problem when a classification model is strictly trained on certain types of crises events and in a particular language (for example data from earthquake events in *English*), and is evaluated on different types of crises events, in a different language (for example flood events in *Italian*). This is performed by creating a multi-lingual crises dataset by translating the data into 6 different languages, and then designing cross-crisis type classification on cross-lingual data sets. In these experiments, we analyse the impact of adding *semantic features* and translation of the data to the same language. The evaluation is conducted in two scenarios: (1) when the cross-crisis type data is not in the same language as of the training data; and (2) when the cross-crisis type data is in the same language as of the training data (bringing the data to same language via translation).

The contributions of the work done in this chapter can be summarised as follows:

- We generate hybrid statistical-semantic classification model by extracting semantics from two different knowledge bases: *DBpedia* and *BabelNet*.
- We use data from 26 different crises events, spanning across 7 types (floods, typhoons, earthquakes, shooting, explosion, bombing, and train crashes), and in 6 different languages (English, Spanish, Italian, German, Portuguese, and French), to classify crisis relatedness in a cross-crisis cross-lingual set up.
- Evaluate classifiers with multiple features, languages, and type of crises, resulting in a total of 1728 experiments.

- Show that data translation to the same language and then combining the DBpedia semantics outperforms the baseline statistical features by 16.42%, on average, in a cross-lingual cross-crisis classification scenario. While, adding the DBpedia semantics without translating to the same language outperforms the baseline statistical features, on average, by 11.24%.

6.2 Relevancy Identification Across Language and Crisis Types

As we mentioned earlier, we have the following aims: (i) analyse the performance of the classification models in classifying crisis-related tweets, when the type of crises events and the language of the training data, which the model is trained on, is different than those of the data the model is applied to (for example the model is trained on data from *floods* in English and applied to data from *earthquakes* in French), and (ii) analyse the impact of the machine translation and semantic features in alleviating the bias of the crises type and language due to the training data, and thus evaluating their impact on the performance of the classifiers.

The proposed approach for creating and evaluating the binary classification model comprises of the following phases, as also shown in Figure 6.1:

A. *Input Data and Preprocessing*: A binary label annotated dataset comprised of crisis events of multiples types is processed for alleviating training and evaluation bias towards a particular class or crisis types in certain languages. This is achieved by balancing the dataset across both the classes and then creating mono-lingual datasets in 6 languages, for all the crisis events. Thus, ensuring that all crisis events are covered in all the considered six languages.

B. *Training/Evaluation Sets Generation*: We segregate the datasets into training and evaluation sets in a way to evaluate the classification models in a cross-crisis-type and cross-lingual scenario.

C. *Feature Engineering*: Build the statistical and semantic features which are used for generating

the binary classification model.

D. *Model Selection and Training*: We train the classifier using the training data.

E. *Model Usage and Evaluation*: We evaluate the classification model on the held-out data (while segregating the training and evaluation data). Depending on the approach, if it involves bringing the evaluation data to the same language as of the training data, the language of the evaluation documents may be reconciled with the training language using machine translation.

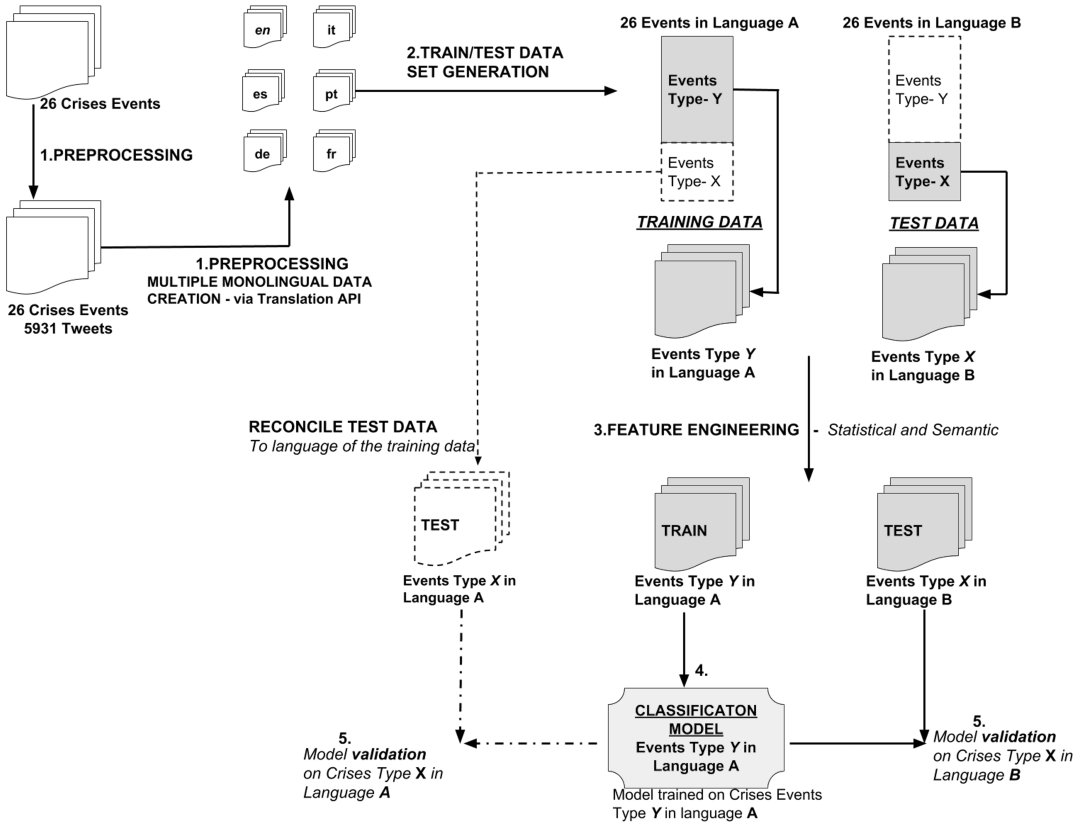

Figure 6.1: Pipeline for relevancy identification across language and crisis types

6.2.1 Input Data and Preprocessing

In order to train cross-lingual and cross-type binary classifiers that can identify *crisis-related* and *not related* documents, we need to have multiple mono-lingual and mono-crisis-type datasets so that data in a particular crisis-type and language can be used as the training data while the other crises-types in other languages can be used as the evaluation data.

Even though existing crisis-related Twitter datasets tend to provide tweets that are easily separable by their crisis-types (e.g., floods, fires, explosions, etc.), these datasets are usually composed of duplicates (e.g., retweets) and multilingual tweets that need to be taken care of before being used for training a cross-type and cross-language classification model. In particular, duplicate tweets may lead to an over-fitted model for certain types of tweets. Similarly, the presence of multilingual tweets may invalidate the cross-lingual setting that is needed for performing our cross-language experiment. Finally, annotated datasets may also be unbalanced. As a result, it is also important to enforce that a binary classifier has the same amount of positive and negative samples during its training phase for avoiding any kind of bias towards a specific class.

We identify duplicate tweets by matching the tweets, one by one, in pairs after removing user-handles (i.e., '@' mentions), special characters, and URLs. If the strings match, the new one is discarded. Similarly, different methods can be used of identifying and dealing with specific languages in tweets. In this work, we perform an 'identify and translate' language normalisation approach where we first identify the language of a tweet using automatic methods and then use machine translation tools for generating monolingual versions of crises-types datasets.

Similarly to the previous chapter, to get an idea of the languages used in the dataset, we use 3 different language detection APIs to determine the language of each document (tweet) : detectlanguage*,

*detectlanguage, detectlanguage.com

langdetect*, and TextBlob† and label the language of each tweet with what is agreed by at least 2 of the APIs. To this end, more than 30 languages are found in the dataset with English (en), Italian (it), Spanish (es), and Portuguese (pt) representing nearly 93% of the data.

Although in principle the manual translation of each tweet would lead to better translation accuracy, we decide to use automatic translation tools since they are more scalable than manual annotation. We create a multilingual dataset for 6 languages (Figure 2): English (en), Italian (it), Spanish (es), French (fr), German (de), and Portuguese (pt) by relying on the Google Translation API (Neural Machine Translation System), which was found to be most accurate over other automatic translation methods (Wu et al., 2016). Each tweet is translated to the rest of the 5 languages, one by one, if it is already in one of the 6 chosen languages. If the tweet is not in any of these 6 languages, then we translate it to all 6 languages. Following this, each annotated tweet is available in 6 different language (as shown in Fig. 6.2), and thus we create multiple mono-lingual and mono-crisis-type datsets.

6.2.2 Training & Evaluation Sets Generation

In the previous steps, we created mono-lingual datasets for 6 languages. Based on these mono-lingual datasets we can selectively generate the training and evaluation data. As we aim to have the training datasets be represented only in a certain language and of selected crises types, we first select a mono-lingual dataset (in any one of the languages) and then further select the crises events, which are particularly not of the types that we aim to evaluate the classifier on. To create the test data sets, on which we evaluate the model, we pick up another mono-lingual dataset in a different language than the training data, and specifically select those crises events which are of the type we want to evaluate the model on. These test data events do not occur in the training data as they were held out.

*langdetect, pypi.org/project/langdetect/
†TextBlob, textblob.readthedocs.io/en/dev/

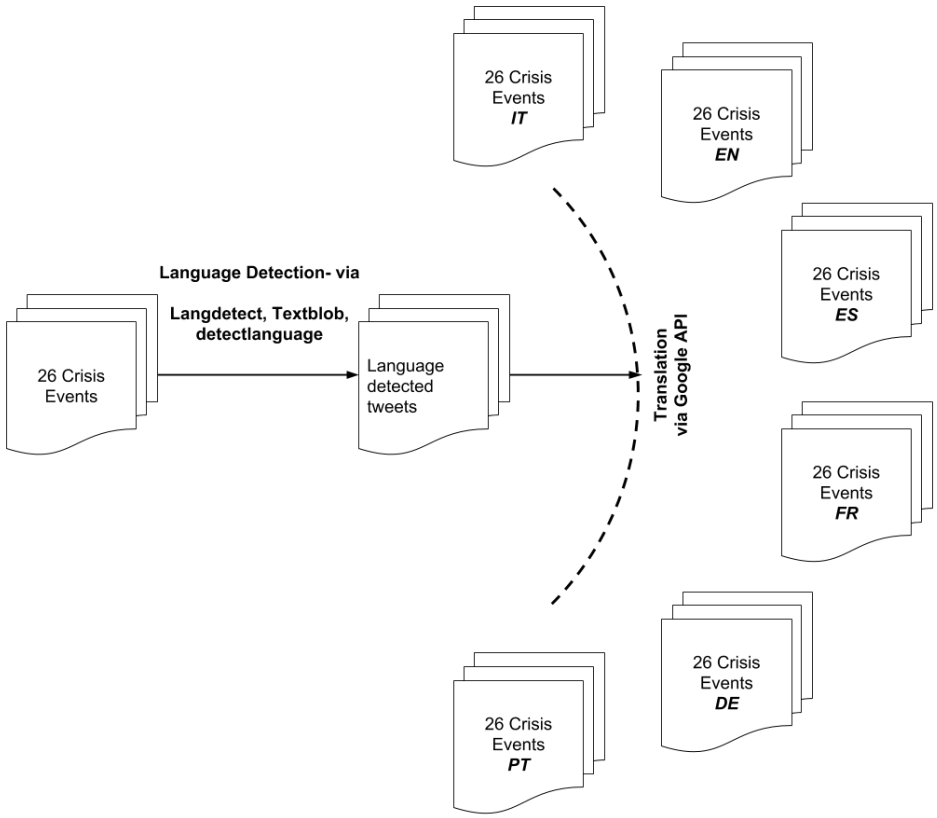

Figure 6.2: Multilingual dataset for crises events via translation

6.2.3 Features

As in the previous chapter, we generate two types of features for the binary classification task of classifying the tweets as crisis *related* and *not related*: *statistical features* and *semantic features*. To address the research question RQ4 in this chapter, we consider the *Statistical Features* as the baseline approach. The semantic features represent the named entities and associated semantic information extracted from the knowledge graphs.

Statistical Features

For every tweet we extract the same statistical features as earlier seen in the chapters 4 and 5, which can be referred to in the Section 4.2.2, under subsection Statistical Features.

In order to extract the statistical features for multiple languages, we chose spaCy[*] library to extract Part of Speech (POS) features. We tokenise the data to unigrams by regexp tokenizer in NLTK[†]. We use a dedicated list of words to filter out the stopwords[‡]. Furthermore, TF-IDF vector normalisation is applied over the unigrams to weigh the tokens in accordance with their relative importance within the dataset, and represent the data in the vector space. In the models, where we include the semantic features, the tokenisation, removal of stopwords, and TF-IDF normalisation is performed after semantic feature inclusion.

Semantic Features

Semantic features are aimed towards forming a more generic representation of crisis data information across languages and crisis types. The features are designed to broaden the context of documents and by making them less crisis-specific, thereby alleviating the issues of data scarcity in, otherwise,

[*] spaCy, www.spacy.io
[†] NLTK, www.nltk.org
[‡] Stop-words list, raw.githubusercontent.com/6/stopwords-json/master/stopwords-all.json

event specific vocabulary. As in the previous chapters 4 and 5, we extract the same semantic features, from BabelNet and DBpedia, and can be referred to in the Section 4.2.2, under subsection Semantic Features.

Given the multilingual nature of these knowledge bases, semantics are extracted in *English* regardless of the language of the post, thus bringing cross-lingual data closer contextually via the added semantic vocabulary. Generalization of semantics in one language also reduces potential data sparsity resulting from varying morphological forms of entities across languages. The semantic features will also bring the data from different types of crisis contextually closer. A conceptual representation of semantic expansion for an example is shown in the Figure 6.3.

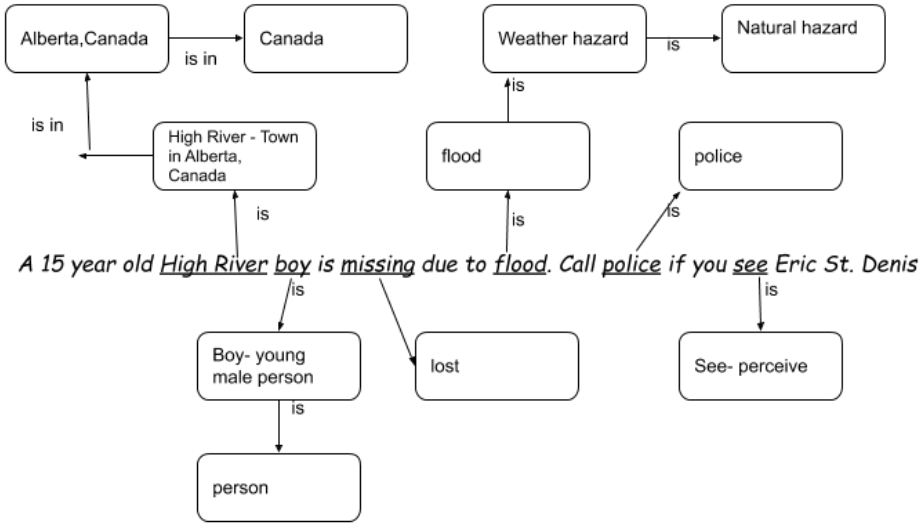

Figure 6.3: Conceptual representation of a semantically annotated post

Babelfy performs NER using the multilingual knowledge base BabelNet. BabelNet is structured

in a way where a common synset represents a certain entity/concept across all its variants in multiple languages. For example, the terms *police* and *policia* (in Spanish) are both represented by the same SynsetID in BabelNet with its English sense as *police*. So a same SynsetID and DBpedia URI is returned for the two terms in different languages. Further, we extract hypernyms and DBpedia properties to associate various entities across different languages. As an example, *guardie di sicurezza* ('secutiry guards' in Italian) and *police* share the same DBpedia subject - *security_guard*. Consequently, data from a wide range of crises events as well as languages get contextually aligned via semantic features.

Let us look at two tweets for an example, in Table 6.1, which originated in two different events and in two different languages. Post A originated in an earthquake event and is in English, while Post B originated during floods and is in Italian. From Table 6.1, we can see the two tweets gaining contextual similarity with the semantic features. We also see the similarity gained by the translation of the tweets mutually into each other's language.

6.2.4 Model Selection and Training

We need to train different models using the training datasets and engineered features, created in the previous steps, using a suitable supervised model. In the previous chapters, the appropriateness of SVM Linear Kernel was validated over RBF kernel, Polynomial kernel, and Logistic Regression. As a consequence, we opt for Support Vector Machine (SVM) with a Linear Kernel as the classification algorithm.

Multiple classification models can be constructed using different subsets of the generated features such as the statistical features and the semantic features. We design three different types of models as follows, and evaluate them separately:

- *SF*: This model uses only the statistical features and is the baseline.
- *SFSemBN*: The statistical features and semantic features from BabelNet are combined (entity

Table 6.1: Semantic expansion with BabelNet and DBpedia semantics

	Post A	Post B
Feature	*'#WorldNews! 15 feared dead and 100 people could be missing in #Guatemala after quake'*	*'Inondazioni in Sardegna, recuperato il cadavere di un poliziotto: almeno 10 tra morti e dispersi: E' morto uno d...'*
Babelfy Entities	*feared, dead, people, missing, quake*	*Inondazioni, recuperarto, cadavere, poliziotto, morti, dispersi, morto*
BabelNet Sense (English)	*fear, dead, citizenry, earthquake*	*floods, catch, dead body, police woman, dead, death, missing*
BabelNet Hypernyms (English)	*geological_phenomenon, natural disaster, group*	*Geological_phenomenon, natural disaster, hydrology, human_body, biological process*
DBpedia Properties	`dbr:Death`, `dbc:Communication`, `dbr:News`, `dbc:Geological_hazards`, `dbc:Seismology`, `dbr:Earthquake`	`dbc:Death`, `dbc:Geological_hazards`, `dbr:Death`, `dbr:flood`
Google Translation	To *it-'#Notizie dal mondo! 15 temuti morti e 100 persone potrebbero mancare a #Guatemala dopo il terremoto '*	To *en-'Floods in Sardinia, recovered the corpse of a policeman: at least 10 dead and missing: He died one d...'*

sense in English, and their hypernyms in English as well, as explained earlier).

- *SFSemDB*: This model combines statistical features with semantic features from DBpedia (labels in English, type, and other DBpedia properties).

6.2.5 Model Usage and Evaluation

To evaluate the models, we use the held-out data as mentioned in Section 6.2.2. To evaluate the performance of the models we determine the precision (P), recall (R) and F1-measure (F1). For the evaluation, two scenarios are considered: (i) the model is evaluated directly on the target document (from a different crisis type event and in a different language) by generating different features mentioned in the above section; (ii) in the second scenario, before generating the features of the target document (which is from a different crisis type event and in a different language), it is translated to the language of the corresponding training data using machine translation services. However, in both scenarios the test and the training data represent the data of different *types* of crisis.

For the first scenario, we use the same notations for the models as mentioned in the above section: *SF*, *SFSemBN*, and *SFSemDB*. For the second scenario, which uses machine translation on the test document before generating the features we use the following notations for the model:

- SF^T: The model uses only the statistical features but the test data is translated to the language of the training data.
- $SFSemBN^T$: The model is the same as *SFSemBN* but the test data is translated into the same language as the training language.
- $SFSemDB^T$: The same model as *SFSemDB* but the test data is translated to the same language as the training language.

6.2.6 Dataset

For this study as well, we use the CrisisLexT26 (Olteanu et al., 2015) dataset which comprises of annotated tweets from 26 different crises events. There are 1000 labelled tweets from each event categorised into the following labels: *'Related and Informative', 'Related but not Informative', 'Not Related' and 'Not Applicable'*. As shown in Table 6.2, we have broadly categorised the events into 11 *types* (the table also shows the language distribution across each crisis category). The categorisation approach is similar to the one used in Chapter 4, and is based on a broad understanding of the crisis events. For example, we treated *floods* and *typhoons* belonging to the same crisis type, since typhoons often result into floods.

As in previous chapters, since this work also focuses on the binary classification scenario, we merged the documents labeled as *Related and Informative* with the documents labeled as *Related but Not Informative* to form the *Related* class, and merged documents labeled as *Not Related* with documents labeled as *Not Applicable* to form the *Not Related* class. Following this, we remove the duplicate tweets using the method described earlier. To this end, there were 21378 unique tweets labeled as *Related* and 2965 unique tweets labeled as *Not Related*.

Next, to avoid the bias between the *Related* and *Not Related* classes, we balance the data by under sampling the majority class and matching the number of Related tweets with with the Not Related ones via a random selection process, across each crisis event. This result is a final dataset with overall size of 5931 tweets (2966 Related and 2965 Not Related). And as described earlier, we perform language normalisation to create 6 versions of the monolingual datasets, using Google Translation API, of the following languages: *English* (en), *Italian* (it), *Spanish* (es), *Portuguese* (pt), *German* (de), and *French* (fr).

It is important to mention, that for language reconciliation tasks for the models SF^T, $SFSemBN^T$, and $SFSemDB^T$, to avoid a repetition of task, we do not re-translate the data from a given language to

the other. Instead, as we have already created 6 mono-lingual datasets following the translation of the entire dataset to all 6 languages as language normalisation task, we reuse the corresponding translated monolingual datasets to generate the test data for the relevant held-out crises-type events and consider it as a translated version of the test data. As a whole, we had 26 crises events, across 10 crises types, and in 6 versions of monolingual datasets. For the experiments, we chose the following crises types: *floods/typhoons*, *earthquakes*, *train crashes*, and *bombing/explosion/shooting*.

6.3 Experiments

In the following sections we will provide details of the experimental set up where we create and validate multiple classification models based on statistical features, semantic features, and translation of the data.

6.3.1 Experimental Setup

As these evaluations are about the cross-crisis types and cross language, we select the following crises types and events for the experiments:

- We train the classification models on the rest of crisis event *types* except *Bombing/Shooting/Explosion* and evaluate the model on *LAX*, *BOB*, and *WTX*. All the test events are *Bombing/Shooting/Explosion* type of crises.
- We train the classification models on the rest of crisis event *types* except *train crash* and evaluate the model on *SPT* and *LAM*. All the test events are *train crash* type of crises.
- We train the classification models on the rest of crisis event *types* except *floods* and *typhoons* and evaluate the model on typhoons- *TPY*, *TYP* and floods- *ALB*, *QFL*, *CFL*, *PHF*, and *SAR*; all the test events are *flood/typhoon* type of crises.

Table 6.2: Event types and original language distribution (en:*English*, it:*Italian*, es:*Spanish*)

Event Type	Event Instances	Event Type	Event Instances
Wildfire/Bushfire	Colorado Wildfire (CWF), Australian Bushfire (ABF) *en-99.1%, it-0%, es- 0.1%, other-1.6%*	Haze	Singapore (SGR) *en-97.47%, it-0%, es- 0%, other-2.53%*
Earthquake	Costa Rica (COS), Italian (ITL), Bohol (BOL), Guatemala (GAU) *en-43.6%, it-18.6%, es-30.9%, other-6.9%*	Helicopter crash	Glasgow (GLW) *en-99.89%, it-0%, es-0.11%, other-0%*
Flood/Typhoon	*Typhoon-* Yolanda (TPY), Pablo (TYP) *Flood-* Colorado (CFL), Queensland (QFL), Alberta (ALB), Philippines (PHF), Sardinia (SAR) *en-82%, it-12.7%, es-1.1%, other-4.2%*	Building collapse	Savar Building (SVR) *en-86.9%, it-0.82%, es-5.19%, other-7.1%*
Terror/Shooting/ Explosion	Los Angeles (LAX), Boston Bomb (BOB), West Texas (WTX) *en-95.1%, it-0.1%, es-2.1%, other-2.7%*	Location Fire	Brazil Pub (BRZ), Venezuela Refinery (VNZ) en-20.3%, it-0.1%, es-45.8%, other-33.9%
Train crash	Spain Train (SPT), Lac Megantic (LAM) en-47.9%, it-0.1%, es-28%, other-24%	Meteor	Russia (RUS) en-87.56%, it-0.64%, es-2.56%, other-9.24%

- We train the classification models on the rest of crisis event *types* except *earthquakes* and evaluate the model on earthquake- *GAU*, *ITL*, *BOL*, and *COS*. All the test events are *earthquake* type of crises.

Further, the evaluations are designed in two ways, as defined in Section 6.2.5:

A. Train and test in cross-lingual set up (i.e., the language of the training and the test data are different). These would be carried out using the models: *SF*, *SFSemBN*, and *SFSemDB*

B. Test data language reconciled with training data (i.e., the test data is brought to the same language as that of the training data). These would be carried out using the models: SF^T, $SFSemBN^T$, and $SFSemDB^T$.

It is important to remind ourselves, here, that all the crises events are available in all the 6 monolingual datasets, i.e., in all 6 different languages. For each model in cross-lingual evaluation (*SF*, *SFSemBN*, and *SFSemDB*), whenever the training data is in a certain language, the test data can be in other 5 languages. This counts to 30 cross-lingual evaluations for each test event. As there are 16 events, this makes it 480 evaluation cases across each model. Given that there are 3 models (*SF*, *SFSemBN*, and *SFSemDB*), we have a total of 1440 cross-lingual evaluation experiments. For the models where the test data is reconciled with the language of the training data, i.e., SF^T, $SFSemBN^T$, and $SFSemDB^T$, there are 6 evaluation cases for each test event in each model as both the training and test data are available in 6 languages. For the 16 events, it makes 96 evaluation cases in each translation model. Given that there are 3 translation models, we have 288 such evaluation cases. Thus, overall there are 1728 unique evaluation experiments performed in the entire analysis.

We now describe the results of each of the above scenario.

6.3.2 Results: Train and test in cross-lingual set ups

In this section we measure the performance of the two semantic models *SFSemBN* and *SFSemDB*, over the baseline model *SF*. Figure 6.4 shows the violin plots comparing an overall distribution across 480 observations each in *SF*, *SFSemBN*, and *SFSemDB*. From the plots in Fig. 6.4 (which shows the violin plots of the F_1 score distribution across each model) and from Table 6.3 we can see that SFSemDB performs outrightly better than the baseline SF, with an increased overall mean F_1 score and a reduced deviation. While SFSemBN also shows an overall increment in the mean F_1 score, a lesser standard deviation makes *SFSemDB* more consistent. *SF* has an average F_1 score of 0.556 with a standard deviation of 0.07, *SFSemBN* has an average F_1 score of 0.566 with a standard deviation of 0.07, and *SFSemDB* has an average F_1 score of 0.610 with a standard deviation of 0.06. The performance of *SFSemDB* when compared to baseline *SF* was found to be statistically significant (via 2 sample t-test) with a p-value<0.001. While the *SFSemBN* had a higher mean than baseline SF, it was not found to be statistically significant with a p-value=0.289.

6.3.3 Results: Test data language reconciled with training data

In this section we measure the performance of the translation models, where the language of the test data is reconciled with the language of the training data, i.e., SF^T, $SFSemBN^T$, and $SFSemDB^T$ over the baseline model *SF*. From Figure 6.5 (which shows the violin plots of the F_1 score distribution across each model) and Table 6.4, we observe that when the test data is reconciled to the same language as that of the training data, the average F_1 score increases (with and without the semantic features). Addition of semantic features reduces the deviation in the performance (as can be visualised from the violin plots of the F_1 scores in Figure 6.5). Highest mean F_1 score of 0.638 and lowest deviation of 0.058 is observed in the $SFSemDB^T$, with a more consistent distribution in comparison to the other models and also found to be statistically significant over the baseline with a p-value<0.001. SF^T has an average

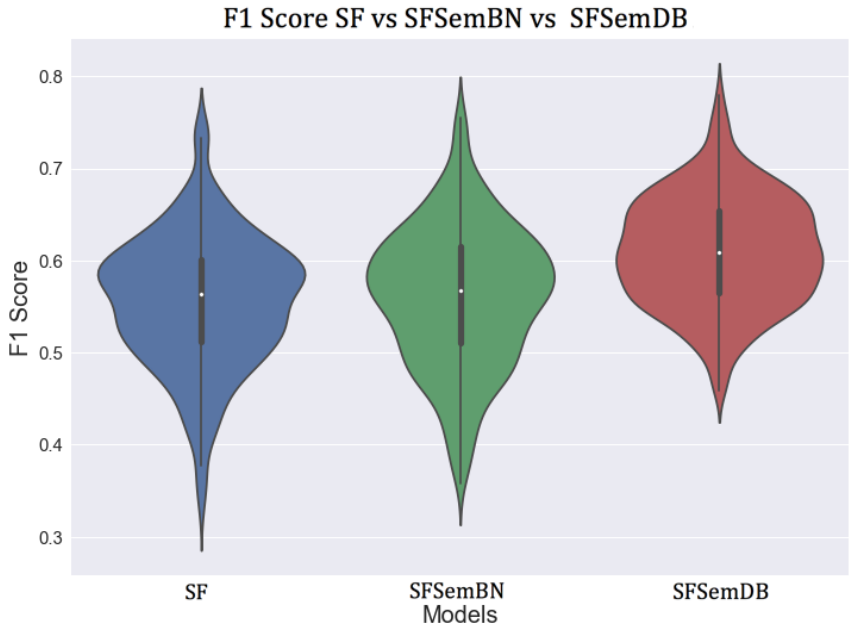

Figure 6.4: Violin plots: F1 score distribution across SF, SFSemBN and SFSemDB

F_1 score of 0.626 with a standard deviation of 0.07 (also being statistically significant over the baseline). $SFSemBN^T$ has an average F_1 score of 0.620 with a standard deviation of 0.07, while being statistically significant over the baseline with a p-value<0.001.

6.3.4 Results: Overall Performance Across All Models

Figure 6.4 and Figure 6.5 show the violin plots for all the models, and by also considering the Tables 6.3 and 6.4, we can say that while both semantic models enhance the performance of the classifier, the best performance is achieved with the combination of the *DBpedia semantics* and the translation in the $SFSemDB^T$ model. $SFSemDB^T$ shows an average F_1 score of 0.638 and an average gain of 16.42% over the baseline *SF* model. If we do not take translation to same language into consideration, *SFSemDB* is the best performing model with an average F_1 score of 0.606 and an average gain (across all the test

Table 6.3: Average overall performance and average performance across crises types, for the models:*SF*, *SFSemBN*, and *SFSemDB*

	SF			SFSemBN			SFSemDB		
	P	R	F_1	P	R	F_1	P	R	F_1
Floods/Typhoons									
AVG.	0.618	0.583	0.551	0.666	0.607	0.567	0.684	0.648	0.628
Earthquakes									
AVG.	0.556	0.551	0.529	0.589	0.561	0.519	0.622	0.608	0.584
Bombing/Explosion/Shooting									
AVG.	0.598	0.586	0.571	0.626	0.613	0.601	0.607	0.602	0.598
Train Crash									
AVG.	0.644	0.618	0.608	0.613	0.607	0.602	0.603	0.592	0.583
Overall									
AVG.	0.602	0.580	0.556	0.633	0.597	0.566	0.644	0.623	0.606
STD.	0.08	0.06	0.07	0.07	0.05	0.07	0.06	0.06	0.06
p-value						0.289			<0.001

events) of nearly 11% over the baseline. We also observe that SF^T model also shows a substantial and statistically significant improvement over the baseline, with an average F_1 score of 0.626. We can see that both, translation and addition of *DBpedia semantics* help in overcoming the over fitting of the models to a specific language or crisis types occurring in the training data.

We also analysed the performance of the classification models across different languages, i.e., when the training and the test data were in same or different languages as shown in Figures 6.6, 6.8, and 6.7. It is to be noted that when languages are the same, they indicate the case of translation models across all the test events in that particular language, i.e., SF^T, $SFSemBN^T$, and $SFSemDB^T$. In SF^T, German (*de*) had the highest average F_1 score of 0.653 (with a standard deviation of 0.085) and Italian (*it*)

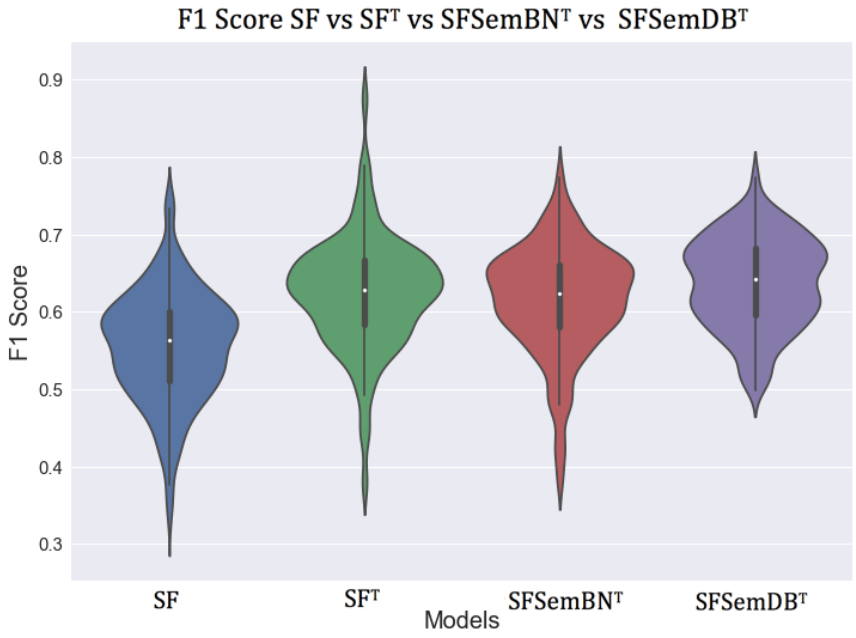

Figure 6.5: Violin plots: F1 score distribution across SF, SF^T, SFSemBN^T and SFSemDB^T

with the lowest average F_1 score of 0.606 (standard deviation of 0.09). In the *SFSemDB*^{T}, French (*fr*) showed the highest average F_1 score of 0.65 (standard deviation 0.07) and Italian (*it*) with the lowest F_1 score of 0.62 (standard deviation of 0.05).

If we observe and compare the bar graphs in the Figure 6.6 and 6.7, we see a definite improvement in the train-test language combination in the *SFSemDB* models (including translation model), in comparison to the corresponing case in the *SF* model.

6.4 Discussion

In this chapter, we aimed at answering the fourth research question *RQ4* by generating hybrid models that use statistical and semantic features to classify the crises data as *related* and *not related*, and are to some extent language as well as crisis type agnostic at the same time. As compared to the work

Table 6.4: Average overall performance and average performance across crises types, for the models:*SF*^{T}, *SFSemBN*^{T}, and *SFSemDB*^{T}

	SF			SF^T			SFSemBN^T			SFSemDB^T		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Floods/Typhoons												
avg.	0.618	0.583	0.551	0.698	0.66	0.643	0.707	0.663	0.644	0.711	0.683	0.672
Earthquakes												
avg.	0.556	0.551	0.529	0.604	0.589	0.569	0.623	0.597	0.567	0.638	0.625	0.608
Bombing/Explosion/Shooting												
avg.	0.598	0.586	0.571	0.631	0.627	0.623	0.638	0.631	0.626	0.609	0.605	0.602
Train Crash												
avg.	0.644	0.618	0.608	0.708	0.691	0.685	0.644	0.635	0.630	0.625	0.613	0.604
Overall												
avg.	0.602	0.580	0.556	0.663	0.640	0.626	0.665	0.637	0.620	0.663	0.645	0.638
std.	0.08	0.06	0.07	0.08	0.06	0.07	0.07	0.06	0.07	0.066	0.059	0.058
p-value						<0.001			<0.001			<0.001

done in the previous chapter, we expand the languages to 6 languages via machine translation APIs and scale the experiments to cross-crisis types simultaneously. It is a challenging task to get a large scale annotated data which spans across several languages and several crises event types. Hence, we simulated the multilingual crises data scenario by recreating multilingual versions of different crises events via translation of the original data. We translated from the original data (for each crisis event) to 6 different languages using Google Translation API, i.e., English (en), Portuguese (pt), Italian (it), German (de), Spanish (es), and French (fr). For the experiments, relying on the translation service was not a time costly process, as Google Cloud allows translation of a maximum of 10 million characters per 100 seconds per project*.

*Google Cloud quotas, https://cloud.google.com/translate/quotas

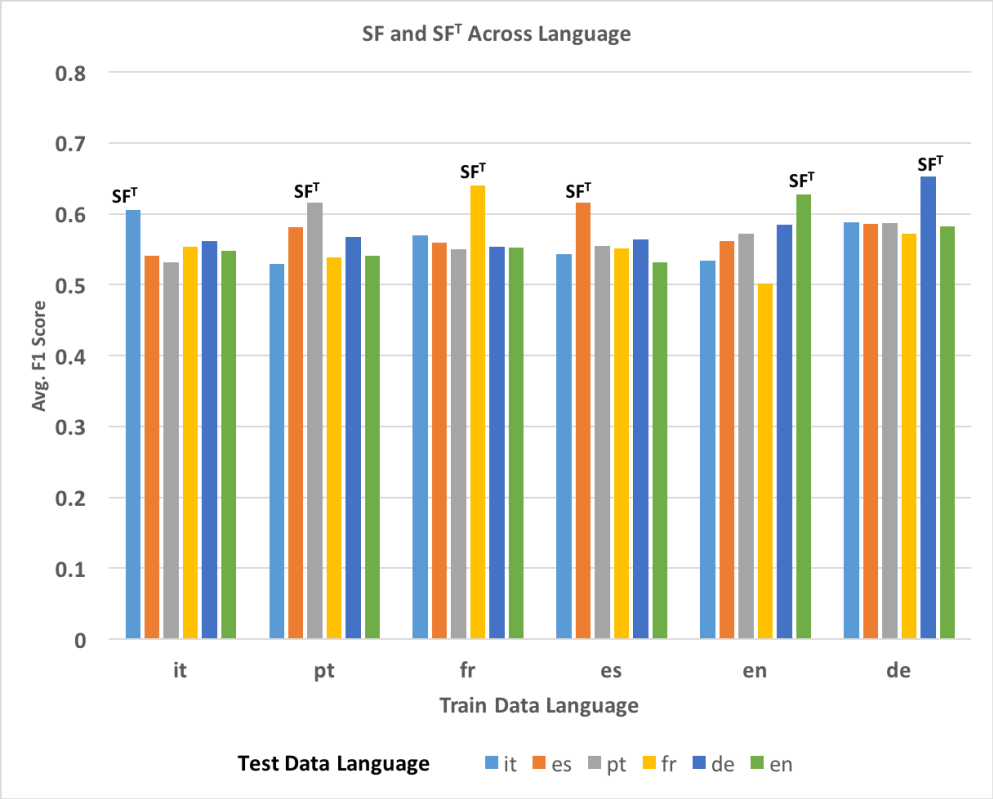

Figure 6.6: SF and SFT across languages

Much like NLP tools and semantic expansion via knowledge bases, machine translation also does not guarantee complete accuracy and can have different levels of accuracy in translations between different languages or might not be even available for a lot of non-European or low-resourced languages. But in order to simulate the cross-lingual cross-crisis scenario we considered it as an appropriate way to determine the feasibility of such methods in such problems. Some of the statistical features are, however, language independent. We did observe that both translation and semantic features (particularly DBpedia semantics) enhances the performance. The translation of the data brings the data to the same language, which catalysis the alignment of the vocabulary in the same language (entities, parts of speech, etc). The semantic features align the context across different crises types. DBpedia semantics

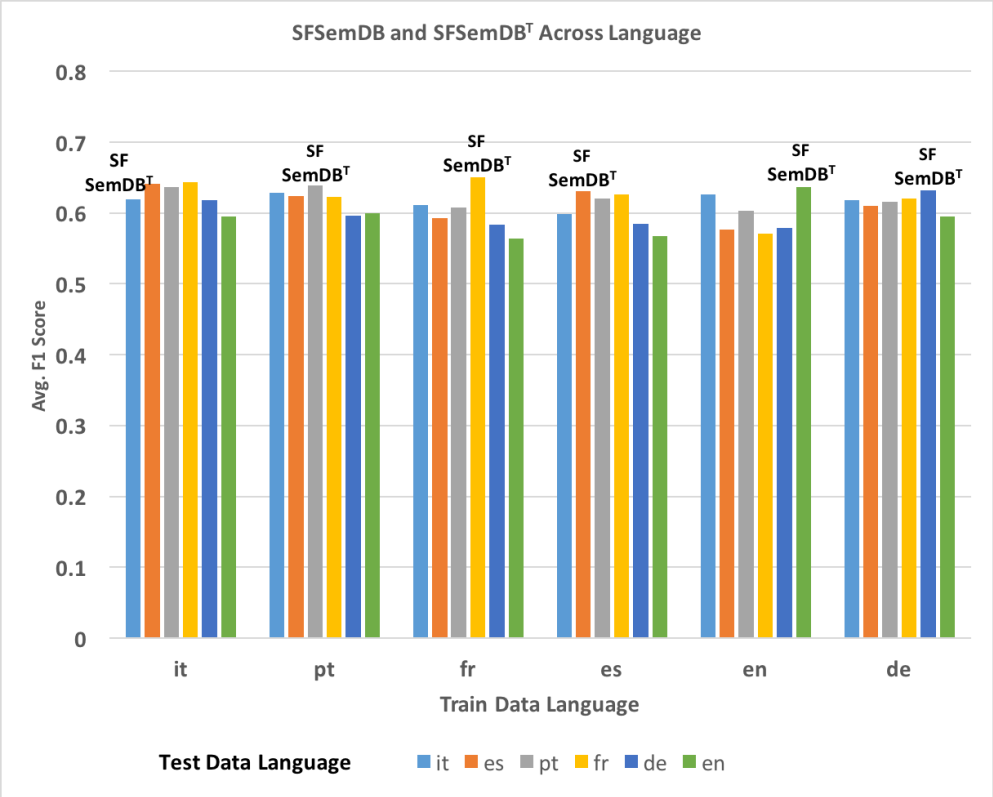

Figure 6.7: SFSemDB and SFSemDB^T across languages

show more impact than the BabelNet semantics, a possible reason is that DBpedia features include higher number of properties which connect the entities at a deeper level. It is to be noted, that once the test and the training data is brought into a same language, the problem fundamentally converts into a cross-crisis classification, which is the problem explored in Chapter 4 (we did not consider the aspect of language in Chapter 4). If the translation is not viable, then the *SFSemDB* turns out to be the best performing feature model for cross-lingual cross-crisis classification.

Figure 6.8: SFSemBN and SFSemBN^T across languages

6.5 Summary

The work done in this chapter is aimed towards answering the research question *RQ4 - "To what extent could semantics improve Tweets classification when the type of crisis event and language change?"*. In this chapter, we took a broader and more realistic aspect of the problem where the incoming crises oriented data might vary in terms of language and the crisis type. A divergence in the language and the nature of the crisis event can impact the validity of any crisis-relevancy classification model. We created different models based on statistical features, translation of the data, and addition of the semantic features. We were able to show that both translation and addition of semantics help in addressing the problem. If translation is not viable, then combining the statistical features with DBpedia features results in

the best performing model. With translation, it is statistical features with DBpedia features extracted from the translated data that performs the best on such a problem.

7

Discussion and Future Work

In this thesis, we have investigated different aspects of the crisis data classification problem. We investigated the impact of semantic features in cross-crisis and cross-lingual crisis classification. We also explored how automated machine translation could complement in building a language agnostic crisis data classifiers. Throughout the experiments we followed a general methodology as defined in Section 1.3 and built hybrid classification models, based on *statistical* and *semantic* features, to classify the data as crisis *related* and *not related*. The overall research scope of this thesis was explored via four research

questions, as seen in Section 1.2. Evaluations performed across various experimental settings, while addressing the research questions, broadly demonstrated that adding the semantics is an effective approach over statistical feature approaches, to develop crisis relatedness classification systems which are applicable to not only new types of crisis events but also in new languages. In the following sections of this chapter, we will discuss the challenges, limitations, and potential future directions we have identified in the course of this thesis.

7.1 Semantic Extraction

Throughout our experimental settings, we had a scenario where we enrich the data by adding the extracted semantics via knowledge bases. We observe that the hybrid semantic features models (with the combination of statistical features) generally outperformed the non-semantic feature models. This was a general observation in the experiments addressing different research questions. However, semantic extraction poses its own challenges. These challenges often pertain to the way knowledge bases are built or the extent to which semantics need to be expanded. For example, if we are using BabelNet to extract hypernyms of associated entities/concepts in a text, then there is a possibility that among the hypernyms we end up extracting a very broad/abstract entity as we noticed in some of the cases in Chapters 3 and 4. The knowledge bases are not always strictly adhering to the hierarchy of concepts because of various automated approaches adopted to create them (since it is nearly impossible to manually curate knowledge graphs representing millions of entities). Such scenarios highlight the need to determining the abstractness of any concerned semantic (concept). As an attempt to address this, in Chapter 3 we created a hierarchy of concepts extracted from BabelNet and analysed the ranking of levels of informative concepts by plotting the Information Gain score of the concepts against their hierarchy level. Further, based on such a hierarchy the abstract concepts were filtered out, which showed a minor improvement in the performance of the classifier. Though the improvements in the classifier's

performance were not large, they demonstrated the potential value of using concept filtering based on abstractness, and the need for future research to explore and improve this approach further.

Similarly, extracting semantics via knowledge bases such as DBpedia has another set of challenges. We retrieve the semantics of entities which are annotated by Named Entity Recognition services (NER). Firstly, the NER services can sometimes be inaccurate and link with a wrong entity. Secondly, expanding the semantics through a knowledge base such as DBpedia can sometimes lead to irrelevant and completely out of domain concepts, which can add to noise in the data and confuse the classifier. For example, in the given tweet - '*Scary Super Typhoon in Philliphines is 236 mph. It's roughly the top speed of Formula 1 cars.*', we can comprehend the context of the text as being related to a crisis situation. However, the NER service will return all the annotated entities, and in this case it returns a link of the phrase *Formula 1* with the corresponding entity *Formula One* in the knowledge graph. However, while expanding the semantics for the entity *Formular One* through various properties such as *type* and *subject*, the context of the overall text gets drifted towards concepts related to *Formula One*, as it is linked with a number of concepts from that domain (Formula One car racing event). When we expand the semantics, it is not trivial to establish which are the relevant semantics and which are the ones that can potentially contribute to noise. Constructing domain-specific relevancy of a knowledge graph is explored in some of the works (Lalithsena et al., 2016; Lalithsena et al., 2017; Perozzi et al., 2014). Some of these works have been explored from a recommender system perspective in the movies or books domain. Generating domain specific knowledge graphs is an extensive research area on its own. While, in this thesis our focus was on building classification models to identify crisis related information from social media and enhancing the applicability of such systems across crisis types and languages, refining the type of semantics (via knowledge graphs) within the premises of crisis situations is a potential next step as a future course of the work proposed in this thesis.

7.2 Multiple Crisis Type Data

Our research scope has revolved around developing classification models applicable across distinct crisis events and of distinct *types*. We managed to utilise a dataset which was spread across multiple crisis events. It is challenging to create datasets showcasing a huge representation of diverse types of crisis events. Additionally, it is also not trivial to manage a substantial volume of data (*related* and *not related* tweets) across each event and each crisis type. In our study, the data which we used was not uniformly distributed in two aspects: (a) number of tweets across each crisis type; (b) number of crisis events across each crisis type. Although, we did manage to create balanced data sets for individual events in most of our experiments, which enabled us to train the classification models with a relatively mitigated bias. In order to build systems that are applicable to unseen crisis events, training them on a wide range of situations will boost their ability to be adaptive.

While, it is imperative to learn from a diverse set of crisis situations, it is also important to ensure that a diverse range of information is also fed to the classification models to learn from. In our work, the crisis events were regarded as belonging to a certain *type* based on how the event was identified by the official agencies (e.g., typhoon, earthquake, flood). What we did not analyse about the data was whether or not different types of crisis events were generating different type of content. There is a possibility that certain events might generate similar content (e.g., typhoons and floods). Therefore, in terms of training the classification system which can identify crisis related information from a diverse content, we can think of analysing the content similarity across the data as a future step. Thus, being selective with the nature of content being used for training and testing. One of the possible methods to induct this into the methodology, in future, is to use cosine similarity between tweets of different types and to determine a threshold value based on which criteria can be established while curating training and testing data.

7.3 Multilingual Crisis Data

Language forms a very critical aspect of crisis data classification problem. The classification systems are valuable when they are responsive to the content in a new language. In our research scope, we kept this aspect of the problem as one of our core research questions. In order to build up crisis data classification systems, we took the data from different crisis events which resulted in a multilingual data source. However, this did not yield a large scale multilingual data set equally (or significantly) distributed across all the found languages in the data, instead it was skewed in its distribution across various languages. In Chapter 5, we experimented with the data originating in three languages, while in Chapter 6 we curated a multilingual data source by using the automated machine translation service. Finding or curating a large scale multilingual crisis data evenly distributed across several languages is a challenging task. Firstly, not every crisis event that happens across the globe might come to notice to be able to collect data. Secondly, there might not be a sufficient volume of data getting generated online in certain geographical locations, thus effecting the amount of the data in a language prominent in that area. Thus, to simulate the multilingual scenario we decided to rely on automatic machine translation systems. Machine translation systems are not absolutely efficient in translation and there is a possibility of an incorrect translation or not a completely accurate translation. Nevertheless, automatic machine translation certainly helps in developing a proof of concept for developing crisis data classification models.

Different languages might have lexical or syntactic similarities due to common roots in the language evolution tree. Considering these relationship between the languages, as a future work, we can perform an in-depth exploration of the connection between the languages based on lexical similarities of the data in different languages and how the classifiers behave across different languages. In Chapter 5, we tried to determine the ranked order correlation metric of informative features between the data originating from two different languages. However, it can certainly be extended to an elaborative

study to establish the similarities between the languages, and determine whether or not translating the data necessarily works well in each cross-lingual scenario.

7.4 Experiment Results

Across the different experiments conducted in this thesis, we analysed how semantic features extracted from knowledge graphs can be exploited to generate machine learning based classifiers to identify crisis related information in social media data. We used two different knowledge graphs; *DBpedia* and *BabelNet* for extracting the semantic features. We observe that while both the types of semantic features (i.e. *DBpedia semantics* and *BabelNet semantics*) show improvement over the baseline in many test cases, the *BabelNet semantics* were not consistent. *DBpedia semantics* show a consistent improvement across the test cases, in general, throughout all the research questions. One of the possible explanation is that we use a higher number of properties from DBpedia to extract additional contextual information in comparison to BabelNet where we only extract hypernyms. We began our exploration, addressing research question *RQ1*, with initial experiments on classifying crisis related data on new crisis events, in Chapter 3. We trained our classifiers on random crisis events, predominantly in English, and used an unseen crisis event as the test data. Here, the crisis events in the training and test data were only segregated on the criteria of distinct events and not on *type* of crisis or language the event represented. In Chapter 4, addressing research question *RQ2*, when the crisis *type* of the test event is not seen in the training data, we found *DBpedia semantics* as the best and most consistent feature set up. In Chapter 5, addressing the research question *RQ3*, when the language of the crisis data is not seen in the training data, we tested two scenarios: (i) keeping the test data as it is and adding the semantics, (ii) translating the test data to the language of training data and then adding the semantics. In both cases, the semantic features show a better performance over the actual baseline (statistical feature model without the translation). *DBpedia semantics* had shown a consistent performance when the

data was cross-lingual and had not been translated. We also found the Spearman's Rank Order Correlation between ranked informative features (based on IG) being improved across all pairs of cross-lingual data while using *DBpedia semantics*. Spearman's Rank Order Correlation between ranked informative features was a way to determine that by adding the semantics, how the actual data gets effected in terms of valuable information. The addition of semantics enabled the cross-lingual data to reflect more similarity with respect to the informative information/features across the languages. In Chapter 6, we combined the two unique research problems of cross-crisis type and cross-lingual data, which is more likely to occur in real crisis situations, where a new type of crisis event can reflect data in multiple languages. We created six monolingual versions of the dataset (in six different languages) by using automated machine translation service. This enabled us to create unique experiment cases of selectively choosing training and test data in certain crisis types and in a certain language. We observed that in this scenario (experiments for addressing research question *RQ4*), the most consistent improvement was exhibited by the *DBpedia semantic* models, with and without translation.

Another potential aspect of crisis information identification problem could be to identify the temporal trends of semantics or topics across crisis events. Analysing the temporal trends could help in determining if different topics, within crisis events, exhibit a pattern in their life span during crisis. A possible approach, as a future work, could be to create the topic clusters and visualise them in temporal order, thus being able to analyse the gain or loss in traction of different topics. Such an analysis could aid in fine tuning the classification systems to focus on the content which is more likely to be relevant at a certain point of time, as the life span of a crisis event progresses. This could be a potential area for future research.

It should be stated that the approaches explored in this thesis rely on natural language processing tools, knowledge bases, and translation services (if opting for translation based models). These tools are not always absolutely accurate, particularly on social media data where the text often does not comply thoroughly with the grammatical and lexical standards. However, the scientific studies conducted

in this research thesis, by defining the scope of experiments spread across different research questions, are meant to explore the potential methods that can be adopted to tackle a genuine challenge faced by global communities, i.e., of identifying relevant information when it matters the most- during crises!

8

Conclusion

The broad research objective of this thesis was to explore classification strategies for identifying crisis related information from social media data. We focused on Twitter, as a use case, to collect the data for this study. A wider research question investigated in this thesis was:

"To what extent could semantics improve crisis relatedness classification of Twitter data?"

To this end, we directed our research exploration to the following four research questions:

- RQ1 - *How could the addition of semantics improve the binary classification of Tweets with regards to their relevancy to crises?*
- RQ2 - *To what extent could semantics improve Tweets classification for new types of crisis events?*
- RQ3 - *To what extent could semantics improve crisis-relevancy classification of Tweets written in a new language?*
- RQ4 - *To what extent could semantics improve Tweets classification when the type of crisis event and language change?*

We addressed the above research questions individually in Chapters 3, 4, 5, and 6 respectively. We hypothesised that enriching the tweets with semantics extracted through entity extraction and knowledge graphs can be an effective approach to deal with diverse crisis data across crisis types and languages. Semantics extracted via knowledge graphs can homogenise the context in the data by establishing the relationships which exist between various concepts. In this chapter, we will summarise our findings from individual chapters that addressed each research question.

8.1 Classifying Crisis Data - A Hybrid Statistical Semantic Approach

In Chapter 3, we focused on addressing the first research question,

- RQ1 - *How could the addition of semantics improve the binary classification of Tweets with regards to their relevancy to crises?*

To address this research question we considered the crisis events, from CrisisLexT26 dataset, which were in English. We considered 9 crisis events for the analysis. To extract the *semantic features* we used NER service *Babelfy* to link entities in the tweets and *BabelNet* to extract hypernyms for linked entities. To extract the *statistical features*, quantified linguistic and structural properties of text were

computed. To classify the tweet as crisis *related* or *not related*, a binary classification approach was adopted. We used SVM (support vector machine) with Linear Kernel as the classification algorithm. To determine how the classifier responds to unseen crisis events, the classifier was trained on 8 out of 9 crisis events, and the left out event was treated as the testing data. Our analysis showed that when the classifier is applied to unseen crisis events, the semantic features enhance the accuracy of the binary classification. We also observed that semantic enrichment, sometimes, results in the inclusion of very abstract concepts. To address this, we proposed a filtering mechanism of abstract concepts based on the hierarchy of concepts in *BabelNet* and information gain score of informative features. The hierarchy from BabelNet was created using hypernym-hyponym relationship of the concepts, which allowed us to iterate through the relationship tree of concepts in the data. The filtering approach showed some improvement over the *semantic feature* model.

The main conclusion from the work conducted in this chapter was that semantic features do enhance the classifier's performance when it is applied to an unseen crisis event. We did not take into consideration the *type* of crisis events in the training or the testing data. This formed part of the second research question *RQ2*, addressed in Chapter 4.

8.2 Classifying Crisis Information Relevancy Across Crisis Types

In Chapter 4, we focused on addressing second research question,

- RQ2 - *To what extent could semantics improve Tweets classification for new types of crisis events?*

To address this research question, we explored a specific scenario where the classification model was trained on data from certain *types* of crisis events, and applied to data from new *types* of crisis events. For instance, we analysed how will the model perform when it is trained on data from crisis events other than *earthquakes*, and applied to *earthquake* type crisis events. We observed that when the classifier is applied to new *types* of crisis events (i.e., the classifier has not seen the testing *type* events in the

training data), the accuracy drops on average by around 17% (see *SF* model average F_1 scores in Table 4.5 and 4.6), in the *statistical features* model (SF). However, when we include semantic features, the classification accuracy of the model on unseen crisis *types* increases by +7.2% in F1 in comparison to non-semantic models. We noticed that semantic features, particularly *DBpedia semantics*, enhanced the classifier's adaptability to identify crisis related information from unseen crisis types. The inclusion of semantic features made the vocabulary of crisis events more broader and less event specific. This increased the scope of broader concepts becoming discriminative/informative, which are likely to exist in unseen crisis events as well.

8.3 Classifying Crisis Information Relevancy Across Multiple Languages

In Chapter 5, we focused on addressing third research question,

- RQ3 - *To what extent could semantics improve crisis-relevancy classification of Tweets written in a new language?*

Crisis data is often multilingual, not only across diverse crisis events from different geographic locations, but it could also be multilingual within the same event. Hence, language forms a crucial factor of crisis relevancy classification models, so that they are adaptive to crisis data in new languages. It is neither feasible to train a model from scratch in a new language in real time nor is it feasible to build a model trained on all languages. We conducted the study to determine how the classifier would perform when the model is trained on crisis events in a certain language, and applied to crisis events in a new language. Other than the statistical features model, we tried two approaches; adding the semantic features, and translating the test data from its original language to the language of training data. We considered all the events from CrisisLexT26 dataset and narrowed down our analysis to three languages (English, Italian, and Spanish), which the original CrisisLexT26 data set primarily existed in. We hypothesised that semantic features can aid in enhancing the morphological (vocabulary), along

with contextual, similarity across data from different languages. We investigated two scenarios: (i) when the model is trained on crisis data from a certain language and evaluated on data from a new language; (ii) the model is trained on crisis data from a certain language and evaluated on crisis data from a different language but only after translating the test data from its original language to the language of the training data.

Our findings in this chapter demonstrated that a combination of statistical and semantic features enhances the performance (average F_1 score) of classifier by 8.26%-9.07%, in comparison to the traditional statistical models, when dealing with cross-lingual classification. Also, translating the data to the same language improves the classifier's performance in identifying crisis related information from crisis events in a new language.

8.4 Classifying Crisis Relevancy Across Languages and Crisis Types

In Chapter 6, we focused on addressing the last research question,

- RQ4 - *To what extent could semantics improve Tweets classification when the type of crisis event and language change?*

In Chapters 4 and 5, we focused on two discrete problems of varying types of crisis events and multilingual data across crises. In this chapter, we considered the situation when both of these problems, of varying crisis type and crisis data language, occur at the same time. In real world scenarios, this is more likely to be the case where a new type of crisis event occurs and the incoming data is in an entirely new language than what the classifier has been trained on. To explore this scenario, we considered the data from 26 crisis events, from CrisisLexT26 dataset. This data spanned across 7 types of crisis (floods, typhoons, earthquakes, shooting, explosion, bombing, and train crashes). We created 6 mono-lingual versions of the dataset in 6 languages by translating the data using automated machine translation service.

To answer the research question, we created two experimental scenarios: (i) evaluate the crisis relevancy classification model's performance on tweets, when the type of crisis events and the language of tweets in the training data are different to what the model is tested on (for instance, we train the model on tweets from *earthquake* type of events in *English* and apply the model on tweets from *flood* type of events in *Spanish*); (ii) we evaluate the same scenario as the previous one, but the test data is brought into the same language as that of the training data (we do this by referring to the test event in the mono-lingual dataset, in the same language as the training data is in). In the two scenarios mentioned above, we evaluate statistical and semantic feature models, with and without translation. We performed a total of 1728 experiments across different combinations of languages and crisis types in training and test datasets. We were able to show that translation of the test data to the same language, as of training data, and then enriching with *DBpedia* semantics outperforms the baseline model of *statistical* features, on average, by 16.42% (compare average F_1 score of SFSemDB$^{\mathrm{T}}$ model in Table 6.4 with average F_1 score of SF model in Table 6.3) in a cross-lingual cross-crisis classification. Whereas, when translation to the same language is not performed, then *DBpedia* features outperforms the baseline model of *statistical* features, on average, by 11.24% (compare average F_1 score of SFSemDB model with average F_1 score of SF model in Table 6.3).

A

Information Gain vs Hierarchy Level

In Chapter 3, Section 3.2.2 discusses *Semantic Filtering Features* by filtering out concepts based on hierarchy generated from BabelNet. Figures 3.3 and 3.4, show plotting of semantic features, for training data corresponding to Singapore Haze and Australia Bushfire, between *Information Gain* score and *levels* indicating the depth in the hierarchy generated using *hypernym-hyponym* relationship in BabelNet knowledge graph. A similar analysis was conducted for the training data across all the test events. The following graphs show the plotting for informative features against their depth in the hi-

erarchy. To be noted that crisis event name in each graph is indicative of the fact that the analysis is performed on the training data corresponding to the mentioned crisis event (which is the test data for that particular case).

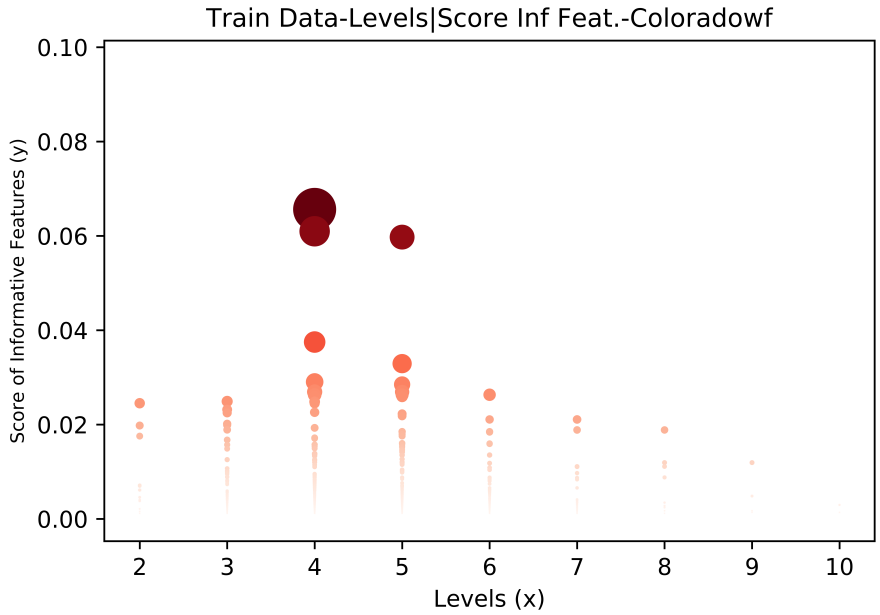

Figure A.1: Information Gain/Level:Training Data-Colorado Wildfire

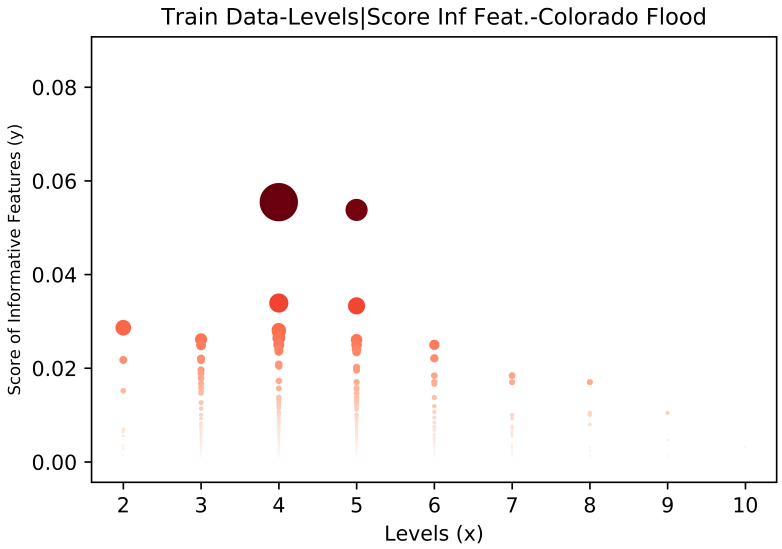

Figure A.2: Information Gain/Level:Training Data-Colorado Flood

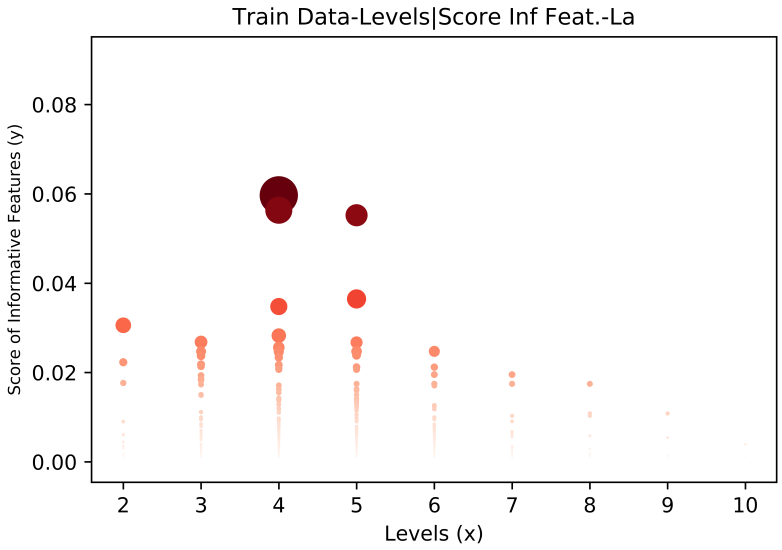

Figure A.3: Information Gain/Level:Training Data-LA Shooting

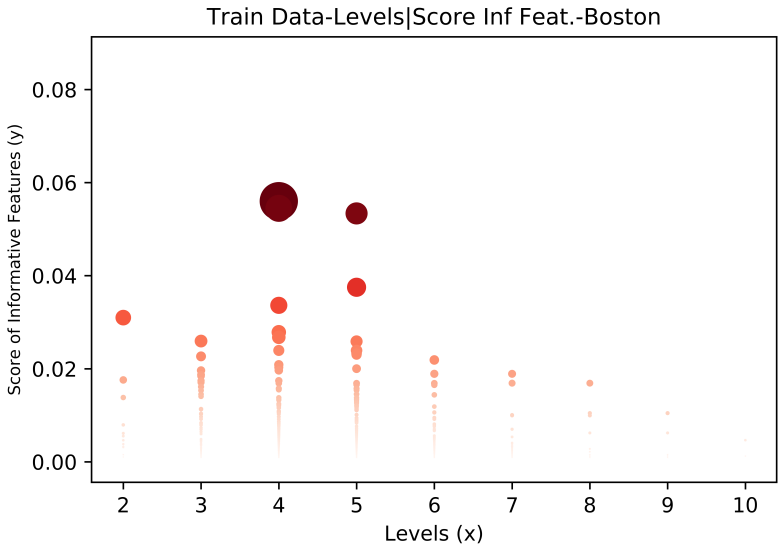

Figure A.4: Information Gain/Level:Training Data-Boston Bombing

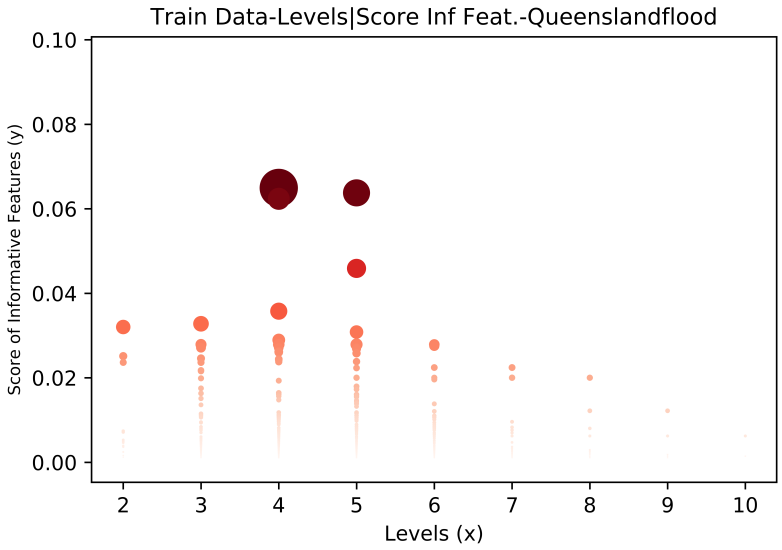

Figure A.5: Information Gain/Level:Training Data-Queensland Flood

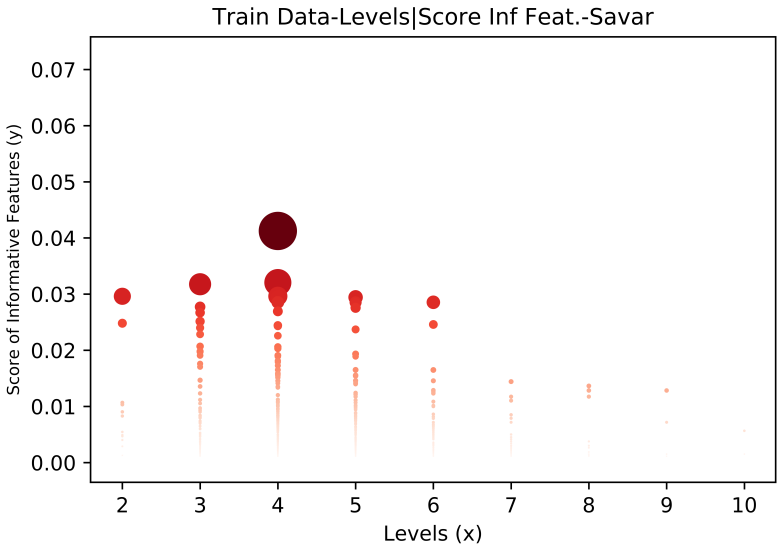

Figure A.6: Information Gain/Level:Training Data-Savar Building Crash

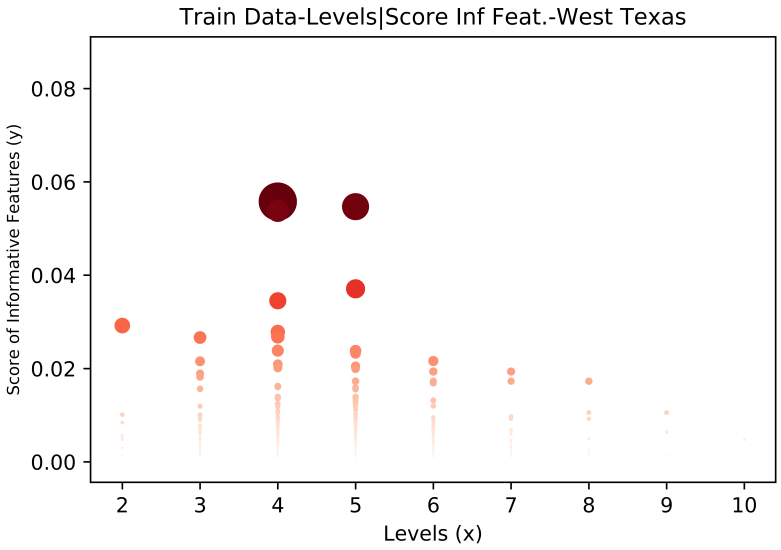

Figure A.7: Information Gain/Level:Training Data- West Texas Explosion

References

[1] Abel, F., Celik, I., Houben, G.-J., & Siehndel, P. (2011). Leveraging the semantics of tweets for adaptive faceted search on Twitter. In *International Semantic Web Conference* (pp. 1–17).: Springer.

[2] Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 305–308).: ACM.

[3] Acar, A. & Muraki, Y. (2011). Twitter for crisis communication: lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities*, 7(3), 392–402.

[4] Agarwal, P., Vaithiyanathan, R., Sharma, S., & Shroff, G. (2012). Catching the long-tail: Extracting local news events from Twitter. In *Sixth International AAAI Conference on Weblogs and Social Media*.

[5] Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

[6] Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M., & Soroa, A. (2009). A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 19–27).: Association for Computational Linguistics.

[7] Ahmad, K., Cheng, D., & Almas, Y. (2007). Multi-lingual sentiment analysis of financial news streams. In *1st International Workshop on Grid Technology for Financial Modeling and Simulation*, volume 26 (pp. 001).: SISSA Medialab.

[8] Alam, F., Joty, S., & Imran, M. (2018). 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)* (pp. 1077–1087).

[9] Allan, J. (2002). *Topic Detection and Tracking: Event-based Information Organization*, volume 12. Springer Science & Business Media.

[10] ALRashdi, R. & O'Keefe, S. (2019). Deep learning and word embeddings for tweet classification for crisis response. *arXiv preprint arXiv:1903.11024*.

[11] Alsaedi, N., Burnap, P., & Rana, O. (2016a). Sensing real-world events using Arabic Twitter posts. In *Tenth International AAAI Conference on Web and Social Media*.

[12] Alsaedi, N., Burnap, P., & Rana, O. (2016b). Sensing real-world events using social media data and a classification-clustering framework. In *2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 216–223).: IEEE.

[13] Araujo, M., Reis, J., Pereira, A., & Benevenuto, F. (2016). An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing* (pp. 1140–1145).: ACM.

[14] Ashktorab, Z., Brown, C., Nandi, M., & Culotta, A. (2014). Tweedr: Mining Twitter to inform disaster response. In *ISCRAM*.

[15] Atefeh, F. & Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence*, 31(1), 132–164.

[16] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. G. (2007). Dbpedia: a nucleus for a web of open data. In *ISWC'07/ASWC'07 Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference* (pp. 722–735).

[17] Baeza-Yates, R., Ribeiro, B. d. A. N., et al. (2011). *Modern information retrieval*. New York: ACM Press; Harlow, England: Addison-Wesley,.

[18] Balahur, A. & Turchi, M. (2014). Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1), 56–75.

[19] Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. In *Fifth international AAAI conference on weblogs and social media*.

[20] Blanford, J. I., Bernhardt, J., Savelyev, A., Wong-Parodi, G., Carleton, A. M., Titley, D. W., & MacEachren, A. M. (2014). Tweeting and tornadoes. In *ISCRAM*.

[21] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.

[22] Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 440–447).

[23] Bontcheva, K. & Rout, D. (2014). Making sense of social media streams through semantics: a survey. *Semantic Web*, 5(5), 373–403.

[24] Brants, T., Chen, F., & Farahat, A. (2003). A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* (pp. 330–337).: ACM.

[25] Bruns, A., Burgess, J., Crawford, K., & Shaw, F. (2011). Crisis communication on Twitter in the 2011 south east queensland floods: Report addressing the use of social media services during national disasters, with a focus on the queensland police force's use of Twitter'.

[26] Bunescu, R. & Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *11th conference of the European Chapter of the Association for Computational Linguistics*.

[27] Burel, G., Saif, H., & Alani, H. (2017a). Semantic wide and deep learning for detecting crisis-information categories on social media. In *International Semantic Web Conference* (pp. 138–155).: Springer.

[28] Burel, G., Saif, H., Fernandez, M., & Alani, H. (2017b). On semantics and deep learning for event detection in crisis situations.

[29] Can, E. F., Ezen-Can, A., & Can, F. (2018). Multilingual sentiment analysis: An rnn-based framework for limited data. *arXiv preprint arXiv:1806.04511*.

[30] Caragea, C., McNeese, N., Jaiswal, A., Traylor, G., Kim, H.-W., Mitra, P., Wu, D., Tapia, A. H., Giles, L., Jansen, B. J., et al. (2011). Classifying text messages for the haiti earthquake. In *Proceedings of the 8th international conference on information systems for crisis response and management (ISCRAM2011)*: Citeseer.

[31] Carmel, D., Roitman, H., & Zwerdling, N. (2009). Enhancing cluster labeling using Wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval* (pp. 139–146).: ACM.

[32] Carvin, A. (2012). *Distant witness: Social media, the Arab Spring and a journalism revolution*. CUNY Journalism Press.

[33] Castillo, C. (2016). *Big crisis data: social media in disasters and time-critical situations*. Cambridge University Press.

[34] Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications*, 40(1), 200–210.

[35] Cobo, A., Parra, D., & Navón, J. (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).: ACM.

[36] Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

[37] Corley, C. D., Dowling, C., Rose, S. J., & McKenzie, T. (2013). Social sensor analytics: Measuring phenomenology at scale. In *2013 IEEE International Conference on Intelligence and Security Informatics* (pp. 61–66).: IEEE.

[38] Cresci, S., Tesconi, M., Cimino, A., & Dell'Orletta, F. (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* (pp. 1195–1200).: ACM.

[39] Cristianini, N., Shawe-Taylor, J., et al. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

[40] Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

[41] Dai, W., Xue, G.-R., Yang, Q., & Yu, Y. (2007). Transferring naive bayes classifiers for text classification. In *AAAI*, volume 7 (pp. 540–545).

[42] Daiber, J., Jakob, M., Hokamp, C., & Mendes, P. N. (2013). Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems* (pp. 121–124).

[43] Dashtipour, K., Poria, S., Hussain, A., Cambria, E., Hawalah, A. Y., Gelbukh, A., & Zhou, Q. (2016). Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8(4), 757–771.

[44] De Choudhury, M., Diakopoulos, N., & Naaman, M. (2012). Unfolding the event landscape on Twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work* (pp. 241–244).: ACM.

[45] De Longueville, B., Smith, R. S., & Luraschi, G. (2009). Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires. In *Proceedings of the 2009 international workshop on location based social networks* (pp. 73–80).: ACM.

[46] Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th International Conference on Data Engineering Workshop* (pp. 507–512).: IEEE.

[47] Denef, S., Bayerl, P. S., & Kaptein, N. A. (2013). Social media and the police: tweeting practices of british police forces during the august 2011 riots. In *proceedings of the SIGCHI conference on human factors in computing systems* (pp. 3471–3480).: ACM.

[48] Deriu, J., Lucchi, A., De Luca, V., Severyn, A., Müller, S., Cieliebak, M., Hofmann, T., & Jaggi, M. (2017). Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the 26th international conference on world wide web* (pp. 1045–1052).: International World Wide Web Conferences Steering Committee.

[49] Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2451–2460).: ACM.

[50] Dou, W., Wang, X., Skau, D., Ribarsky, W., & Zhou, M. X. (2012). Leadline: Interactive visual analysis of text data through event identification and exploration. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 93–102).: IEEE.

[51] Duek, S. & Markovitch, S. (2018). Automatic generation of language-independent features for cross-lingual classification. *arXiv preprint arXiv:1802.04028*.

[52] Elloumi, S., Jaoua, A., Ferjani, F., Semmar, N., Besançon, R., Al-Jaam, J., & Hammami, H. (2013). General learning approach for event extraction: Case of management change event. *Journal of Information Science*, 39(2), 211–224.

[53] Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics* (pp. 363–370).: Association for Computational Linguistics.

[54] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, (pp. 1189–1232).

[55] Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3), 10–14.

[56] Geisser, S. (1974). A predictive approach to the random effect model. *Biometrika*, 61(1), 101–107.

[57] Genc, Y., Sakamoto, Y., & Nickerson, J. V. (2011). Discovering context: classifying tweets through a semantic transform based on Wikipedia. In *International conference on foundations of augmented cognition* (pp. 484–492).: Springer.

[58] Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3), 291–304.

[59] Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228–5235.

[60] Guo, Q., Diaz, F., & Yom-Tov, E. (2013). Updating users about time critical events. In *European Conference on Information Retrieval* (pp. 483–494).: Springer.

[61] Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157–1182.

[62] Hamerly, G. & Elkan, C. (2002). Alternatives to the k-means algorithm that find better clusterings. In *Proceedings of the eleventh international conference on Information and knowledge management* (pp. 600–607).: ACM.

[63] Heverin, T. & Zach, L. (2010). *Microblogging for crisis communication: examination of Twitter use in response to a 2009 violent crisis in the Seattle-Tacoma, Washington, area*. ISCRAM.

[64] Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., & Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 782–792).

[65] Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 289–296).: Morgan Kaufmann Publishers Inc.

[66] Hu, B., Lu, Z., Li, H., & Chen, Q. (2014). Convolutional neural network architectures for matching natural language sentences. In *Advances in neural information processing systems* (pp. 2042–2050).

[67] Hu, J., Fang, L., Cao, Y., Zeng, H.-J., Li, H., Yang, Q., & Chen, Z. (2008). Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 179–186).: ACM.

[68] Hu, X., Sun, N., Zhang, C., & Chua, T.-S. (2009). Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 919–928).: ACM.

[69] Hua, T., Chen, F., Zhao, L., Lu, C.-T., & Ramakrishnan, N. (2013). Sted: Semi-supervised targeted event detection. *KDD'13*, (pp. 11–14).

[70] Hughes, A. L. & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International journal of emergency management*, 6(3-4), 248–260.

[71] Hughes, A. L., St Denis, L. A., Palen, L., & Anderson, K. M. (2014). Online public communications by police & fire services during the 2012 Hurricane Sandy. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 1505–1514).: ACM.

[72] Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys (CSUR)*, 47(4), 67.

[73] Imran, M., Castillo, C., Lucas, J., Meier, P., & Rogstadius, J. (2014a). Coordinating human and machine intelligence to classify microblog communications in crises. In *ISCRAM*.

[74] Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014b). AIDR: Artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 159–162).: ACM.

[75] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013a). Extracting information nuggets from disaster-related messages in social media. In *Iscram*.

[76] Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013b). Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1021–1024).: ACM.

[77] Imran, M., Mitra, P., & Castillo, C. (2016a). Twitter as a lifeline: Human-annotated Twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*.

[78] Imran, M., Mitra, P., & Srivastava, J. (2016b). Cross-language domain adaptation for classifying crisis-related short messages. *arXiv preprint arXiv:1602.05388*.

[79] Jadhav, A. S., Purohit, H., Kapanipathi, P., Anantharam, P., Ranabahu, A. H., Nguyen, V., Mendes, P. N., Smith, A. G., Cooney, M., & Sheth, A. P. (2010). Twitris 2.0: Semantically empowered system for understanding perceptions from social data.

[80] Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

[81] Kanayama, H., Nasukawa, T., & Watanabe, H. (2004). Deeper sentiment analysis using machine translation technology. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

[82] Kanhabua, N. & Nejdl, W. (2013). Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd International Conference on World Wide Web* (pp. 1335–1342).: ACM.

[83] Karimi, S., Yin, J., & Paris, C. (2013). Classifying microblogs for disasters. In *Proceedings of the 18th Australasian Document Computing Symposium* (pp. 26–33).: ACM.

[84] Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1746–1751).

[85] Kumar, S., Morstatter, F., Zafarani, R., & Liu, H. (2013). Whom should i follow?: identifying relevant users during crises. In *Proceedings of the 24th ACM conference on hypertext and social media* (pp. 139–147).: ACM.

[86] Lafferty, J., McCallum, A., & Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[87] Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

[88] Lalithsena, S., Kapanipathi, P., & Sheth, A. (2016). Harnessing relationships for domain-specific subgraph extraction: A recommendation use case. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 706–715).: IEEE.

[89] Lalithsena, S., Perera, S., Kapanipathi, P., & Sheth, A. (2017). Domain-specific hierarchical subgraph extraction: A recommendation use case. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 666–675).: IEEE.

[90] Leavitt, A. & Clark, J. A. (2014). Upvoting Hurricane Sandy: event-based news production processes on a social news site. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 1495–1504).: ACM.

[91] Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., & Bizer, C. (2015). Dbpedia – a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 167–195.

[92] Li, C., Sun, A., & Datta, A. (2012a). Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 155–164).: ACM.

[93] Li, H., Caragea, D., & Caragea, C. (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 (ISCRAM 2017)*.

[94] Li, H., Caragea, D., Caragea, C., & Herndon, N. (2018a). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, 26(1), 16–27.

[95] Li, H., Caragea, D., Li, X., & Caragea, C. (2018b). Comparison of word embeddings and sentence encodings as generalized representations for crisis tweet classification tasks. *en. In: New Zealand*, (pp.13).

[96] Li, H., Guevara, N., Herndon, N., Caragea, D., Neppalli, K., Caragea, C., Squicciarini, A. C., & Tapia, A. H. (2015). Twitter mining for disaster response: A domain adaptation approach. In *ISCRAM*.

[97] Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C.-C. (2012b). Tedas: A Twitter-based event detection and analysis system. In *2012 IEEE 28th International Conference on Data Engineering* (pp. 1273–1276).: IEEE.

[98] Lo, S. L., Cambria, E., Chiong, R., & Cornforth, D. (2016). A multilingual semi-supervised approach in deriving singlish sentic patterns for polarity detection. *Knowledge-Based Systems*, 105, 236–247.

[99] Lorini, V., Castillo, C., Dottori, F., Kalas, M., Nappo, D., & Salamon, P. (2019). Integrating social media into a pan-european flood awareness system: A multilingual approach. *arXiv preprint arXiv:1904.10876*.

[100] Ludwig, T., Reuter, C., & Pipek, V. (2015). Social haystack: Dynamic quality assessment of citizen-generated content during emergencies. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 22(4), 17.

[101] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

[102] Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 227–236).: ACM.

[103] Mathioudakis, M. & Koudas, N. (2010). Twittermonitor: trend detection over the Twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1155–1158).: ACM.

[104] Matykiewicz, P. & Pestian, J. (2012). Effect of small sample size on text categorization with support vector machines. In *Proceedings of the 2012 workshop on biomedical natural language processing* (pp. 193–201).: Association for Computational Linguistics.

[105] Melville, P., Chenthamarakshan, V., Lawrence, R. D., Powell, J., Mugisha, M., Sapra, S., Anandan, R., & Assefa, S. (2013). Amplifying the voice of youth in africa via text analytics. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1204–1212).: ACM.

[106] Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems* (pp. 1–8).

[107] Metaxas, P. & Mustafaraj, E. (2013). The rise and the fall of a citizen reporter. In *Proceedings of the 5th Annual ACM Web Science Conference* (pp. 248–257).: ACM.

[108] Mihalcea, R., Banea, C., & Wiebe, J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th annual meeting of the association of computational linguistics* (pp. 976–983).

[109] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119).

[110] Minka, T. & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence* (pp. 352–359).: Morgan Kaufmann Publishers Inc.

[111] Minka, T. P. (2001). Expectation propagation for approximate bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence* (pp. 362–369).: Morgan Kaufmann Publishers Inc.

[112] Moro, A., Raganato, A., & Navigli, R. (2014). Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2(1), 231–244.

[113] Munro, R. & Manning, C. D. (2012). Short message communications: users, topics, and in-language processing. In *Proceedings of the 2nd ACM Symposium on Computing for Development* (pp.4).: ACM.

[114] Muthukrishnan, S. et al. (2005). Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2), 117–236.

[115] Naaman, M., Boase, J., & Lai, C.-H. (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 189–192).: ACM.

[116] Navigli, R. & Ponzetto, S. P. (2010). Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 216–225).

[117] Navigli, R. & Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217–250.

[118] Nguyen, D. T., Al Mannai, K. A., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. In *Eleventh International AAAI Conference on Web and Social Media*.

[119] Nguyen, T. H. & Grishman, R. (2015). Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2 (pp. 365–371).

[120] Olteanu, A., Castillo, C., Diaz, F., & Vieweg, S. (2014). Crisislex: A lexicon for collecting and filtering microblogged communications in crises. In *ICWSM*.

[121] Olteanu, A., Vieweg, S., & Castillo, C. (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* (pp. 994–1009).: ACM.

[122] Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., & Ounis, I. (2012). Bieber no more: First story detection using Twitter and Wikipedia. In *Sigir 2012 workshop on time-aware information access*.

[123] Paulheim, H. & Fürnkranz, J. (2012). Unsupervised feature generation from linked open data. In *International Conference on Web Intelligence, Mining, and Semantics (WIMS'12)* (pp.31).

[124] Peddinti, V. M. K. & Chintalapoodi, P. (2011). Domain adaptation in sentiment analysis of Twitter. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

[125] Pedrood, B. & Purohit, H. (2018). Mining help intent on Twitter during disasters via transfer learning with sparse coding. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation* (pp. 141–153).: Springer.

[126] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

[127] Perozzi, B., Akoglu, L., Iglesias Sánchez, P., & Müller, E. (2014). Focused clustering and outlier detection in large attributed graphs. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1346–1355).: ACM.

[128] Petrović, S., Osborne, M., & Lavrenko, V. (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* (pp. 181–189).: Association for Computational Linguistics.

[129] Phuvipadawat, S. & Murata, T. (2010). Breaking news detection and tracking in Twitter. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 3 (pp. 120–123).: IEEE.

[130] Ponzetto, S. P. & Strube, M. (2007). Deriving a large scale taxonomy from Wikipedia. In *AAAI*, volume 7 (pp. 1440–1445).

[131] Popescu, A.-M. & Pennacchiotti, M. (2010). Detecting controversial events from Twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1873–1876).: ACM.

[132] Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.

[133] Power, R., Robinson, B., Colton, J., & Cameron, M. (2014). Emergency situation awareness: Twitter case studies. In *International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries* (pp. 218–231).: Springer.

[134] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[135] Purohit, H., Castillo, C., Diaz, F., Sheth, A., & Meier, P. (2014). Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday*, 19(1).

[136] Qu, Y., Huang, C., Zhang, P., & Zhang, J. (2011). Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 25–34).: ACM.

[137] Ratinov, L., Roth, D., Downey, D., & Anderson, M. (2011). Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 1375–1384).: Association for Computational Linguistics.

[138] Rebele, T., Suchanek, F., Hoffart, J., Biega, J., Kuzey, E., & Weikum, G. (2016). Yago: A multilingual knowledge base from Wikipedia, Wordnet, and Geonames. In *International Semantic Web Conference* (pp. 177–185).: Springer.

[139] Reuter, C., Marx, A., & Pipek, V. (2011). Social software as an infrastructure for crisis management–a case study about current practice and potential usage. In *Proceedings of the 8th International ISCRAM Conference* (pp. 1–10).

[140] Reuter, C., Marx, A., & Pipek, V. (2012). Crisis management 2.0: Towards a systematization of social software use in crisis situations. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM)*, 4(1), 1–16.

[141] Ristoski, P. & Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web semantics: science, services and agents on the World Wide Web*, 36, 1–22.

[142] Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 1524–1534).: Association for Computational Linguistics.

[143] Ritter, A., Etzioni, O., Clark, S., et al. (2012). Open domain event extraction from Twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1104–1112).: ACM.

[144] Robinson, B., Power, R., & Cameron, M. (2013). A sensitive Twitter earthquake detector. In *Proceedings of the 22nd international conference on world wide web* (pp. 999–1002).: ACM.

[145] Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., & Laredo, J. A. (2013). Crisistracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, 57(5), 4–1.

[146] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851–860).: ACM.

[147] Sayyadi, H., Hurst, M., & Maykov, A. (2009). Event detection and tracking in social streams. In *Third International AAAI Conference on Weblogs and Social Media*.

[148] Schulz, A., Guckelsberger, C., & Janssen, F. (2017). Semantic abstraction for generalization of tweet classification: An evaluation of incident-related tweets. *Semantic Web*, 8(3), 353–372.

[149] Schulz, A., Ristoski, P., & Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. In *Extended semantic web conference* (pp. 22–33).: Springer.

[150] Schulz, A., Schmidt, B., & Strufe, T. (2015). Small-scale incident detection based on microposts. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media* (pp. 3–12).: ACM.

[151] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1–47.

[152] Shaw, F., Burgess, J., Crawford, K., & Bruns, A. (2013). Sharing news, making sense, saying thanks: Patterns of talk on Twitter during the queensland floods. *Australian Journal of Communication*, 40(1), 23–40.

[153] Siolas, G. & d'Alché Buc, F. (2000). Support vector machines based on a semantic kernel for text categorization. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, volume 5 (pp. 205–209).: IEEE.

[154] Song, Y., Wang, H., Wang, Z., Li, H., & Chen, W. (2011). Short text conceptualization using a probabilistic knowledgebase. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[155] Starbird, K., Muzny, G., & Palen, L. (2012). Learning from the crowd: collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions. In *Proceedings of 9th International Conference on Information Systems for Crisis Response and Management, ISCRAM* (pp. 1–10).

[156] Starbird, K., Palen, L., Hughes, A. L., & Vieweg, S. (2010). Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 241–250).: ACM.

[157] Stowe, K., Anderson, J., Palmer, M., Palen, L., & Anderson, K. M. (2018). Improving classification of Twitter behavior during hurricane events. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media* (pp. 67–75).

[158] Stowe, K., Paul, M. J., Palmer, M., Palen, L., & Anderson, K. (2016). Identifying and categorizing disaster-related tweets. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media* (pp. 1–6).

[159] Strötgen, J. & Gertz, M. (2013). Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2), 269–298.

[160] Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web* (pp. 697–706).: ACM.

[161] Sutton, C. & McCallum, A. (2006). *An introduction to conditional random fields for relational learning*, volume 2. Introduction to statistical relational learning. MIT Press.

[162] Sutton, C., McCallum, A., et al. (2012). An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4), 267–373.

[163] Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009). Adapting naive bayes to domain adaptation for sentiment analysis. In *European Conference on Information Retrieval* (pp. 337–349).: Springer.

[164] Tang, J., Wang, X., Gao, H., Hu, X., & Liu, H. (2012). Enriching short text representation in microblog for clustering. *Frontiers of Computer Science*, 6(1), 88–101.

[165] Tao, K., Abel, F., Hauff, C., & Houben, G.-J. (2012). What makes a tweet relevant for a topic. *Making Sense of Microposts (# MSM2012)*, (pp. 49–56).

[166] Thomson, R., Ito, N., Suda, H., Lin, F., Liu, Y., Hayasaka, R., Isochi, R., & Wang, Z. (2012). Trusting tweets: The fukushima disaster and information source credibility on Twitter. In *Proceedings of the 9th International ISCRAM Conference* (pp. 1–10).: Vancouver: Simon Fraser University.

[167] Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1* (pp. 173–180).: Association for Computational Linguistics.

[168] Toutanova, K. & Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13* (pp. 63–70).: Association for Computational Linguistics.

[169] Trim, C. (2013). The art of tokenization. *IBM Developerworks*.

[170] Truelove, M., Vasardani, M., & Winter, S. (2015). Towards credibility of micro-blogs: characterising witness accounts. *GeoJournal*, 80(3), 339–359.

[171] Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (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* (pp. 1079–1088).: ACM.

[172] Vieweg, S. E. (2012). *Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications*. PhD thesis, University of Colorado at Boulder.

[173] Wang, C., Song, Y., Li, H., Zhang, M., & Han, J. (2016a). Text classification with heterogeneous information network kernels. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[174] Wang, C., Song, Y., Roth, D., Zhang, M., & Han, J. (2016b). World knowledge as indirect supervision for document clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 11(2), 13.

[175] Wang, F., Wang, Z., Li, Z., & Wen, J.-R. (2014). Concept-based short text classification and ranking. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 1069–1078).: ACM.

[176] Weng, J. & Lee, B.-S. (2011). Event detection in Twitter. In *Fifth international AAAI conference on weblogs and social media*.

[177] Wick, M., Kanani, P., & Pocock, A. (2016). Minimally-constrained multilingual embeddings via artificial code-switching. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[178] Wu, W., Li, H., Wang, H., & Zhu, K. (2011). Towards a probabilistic taxonomy of many concepts. *Microsoft Res. Redmond, WA, USA, Tech. Rep. MSR-TR-2011-25*.

[179] Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 481–492).: ACM.

[180] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

[181] Wukich, C. & Mergel, I. (2015). Closing the citizen-government communication gap: Content, audience, and network analysis of government tweets. *Journal of Homeland Security and Emergency Management*, 12(3), 707–735.

[182] Xiao, M. & Guo, Y. (2014). Semi-supervised matrix completion for cross-lingual text classification. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[183] Xu, J., Peng, W., Guanhua, T., Bo, X., Jun, Z., Fangyuan, W., Hongwei, H., et al. (2015). Short text clustering via convolutional neural networks.

[184] Yin, J., Karimi, S., Lampert, A., Cameron, M., Robinson, B., & Power, R. (2015). Using social media to enhance emergency situation awareness. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[185] Yin, J., Karimi, S., Robinson, B., & Cameron, M. (2012). Esa: emergency situation awareness via microbloggers. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 2701–2703).: ACM.

[186] Zaki, M. J., Meira Jr, W., & Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.

[187] Zhang, S. & Vucetic, S. (2016). Semi-supervised discovery of informative tweets during the emerging disasters. *arXiv preprint arXiv:1610.03750*.

[188] Zhang, Z., Gentile, A. L., & Ciravegna, F. (2011). Harnessing different knowledge sources to measure semantic relatedness under a uniform model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 991–1002).: Association for Computational Linguistics.

[189] Zhou, X., Wan, X., & Xiao, J. (2016). Cross-lingual sentiment classification with bilingual document representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1403–1412).

[190] Zhou, Y., Nie, L., Rouhani-Kalleh, O., Vasile, F., & Gaffney, S. (2010). Resolving surface forms to Wikipedia topics. In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 1335–1343).: Association for Computational Linguistics.

[191] Zielinski, A., Bügel, U., Middleton, L., Middleton, S., Tokarchuk, L., Watson, K., & Chaves, F. (2012). Multilingual analysis of Twitter news in support of mass emergency events. In *EGU General Assembly Conference Abstracts*, volume 14 (pp. 8085).: Citeseer.