Information Extraction from Microblogs during Disasters

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Kiran Zahra

aus

Pakistan

Promotionskommission

Prof. Dr. Ross Stuart Purves (Vorsitz)

Prof. Dr. Robert Weibel

Dr. Frank O. Ostermann

Zürich 2021

Faculty of Science
University of Zurich

Information Extraction from Microblogs during Disasters

Kiran Zahra
Geocomputation Unit
Department of Geography
University of Zurich
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

"We cannot stop natural disasters but we can arm ourselves with knowledge: so many lives wouldn't have to be lost if there was enough disaster preparedness."

- Petra Nemcova

ACKNOWLEDGMENTS

Completing my PhD at the Department of Geography, University of Zurich has been an amazing journey that I would have not been completed without the help and support of many people. I would like to pay my special regards to Prof. Dr. Ross S. Purves for supervising my thesis. His kindness, trust, support, professional competence, and scientific guidance made this journey possible and smooth. I also wish to express my deepest gratitude to Dr. Frank O. Ostermann for his scientific advice, support, patience, and encouragement from the beginning until the end of this research. My sincere thanks also go to Dr. Muhammad Imran for supporting a research collaboration between UZH and QCRI and giving me a great opportunity to work with him and get a chance to learn a lot in the field.

I am very grateful to the graduate school in Geography and UZH graduate campus for organizing various courses that helped me improve my written and oral scientific communication skills. I am also very thankful to the old and new members of the Geocomputation group for their support and feedback on various dry runs and discussions.

I would particularly like to thank Katja, Hoda, Raha, Meysam, Michelle, Lucia, and Duri for their help in babysitting Ashhad during various times throughout my PhD research. Their support made it possible for me to attend many conferences and courses. I would like to say a very big thanks to my mother for her prayers, kindness, and willingness to happily babysit Ashhad for several weeks on various occasions during her visits to Zurich.

And last but not the least, a big, fat, fluffy, and beautiful thanks to Ashhad, without his support, it was absolutely impossible for me to pursue my PhD research. I thank you Ashhad for bearing with me all the evenings when you wanted to play with me and I was so tired to play after work and all the nights when you wanted to speak with me and I slept.

Zurich, January 2021

SUMMARY

The excessive human intervention in the environment has accelerated the frequency and intensity of disasters triggered by natural hazards all around the globe. In the case of an emergency event, responders seek diverse and credible information sources to assess the situation on the ground. The advent of digital technology in the form of handheld devices, widespread internet connectivity, and the emergence of web 2.0 changed the way people used to communicate over a decade ago. Social media has substantially shaped information sharing as well as information gathering protocols particularly during mass emergency events. Moreover, advanced information extraction models trained on machine learning algorithms have provided a great opportunity to automatically extract information from a bulk of digital data. However, analysis of social media data has many challenges associated with it due to its informal and unstructured nature of communication.

The overall aim of this research was to address various challenges associated to extract information from short and noisy social media posts during mass emergency events. Firstly, I addressed word sense ambiguity problem. The literature states that collecting social media posts based on disaster-related keywords does not guarantee that the content contains information about the event. For this purpose, machine learning algorithms provide state-of-the-art solutions, however, these algorithms need training data to *learn* information extraction and classification task. To prepare training data for these algorithms costs time and other resources. To overcome this challenge, I used already prepared training data for the same kind of event and only swapped geographic locations in old training data with the geographic locations of new event. This simple approach substantially improved the precision and recall of detecting information from noise without preparing new training data for new disaster events. This experiment was performed using a supervised machine-learning algorithm Naïve Bayes.

Secondly, I addressed the challenge of analyzing credibility of shared information in social media posts. The literature states, eyewitness reports on social media are a credible source of information as users share their first-hand experience and personal observations. To understand the content of eyewitness reports, I designed a taxonomy that categorized different types of eyewitness reports posted during earthquakes, floods, hurricanes, and forest fires. Then I manually analysed the content of these eyewitness reports to extract various linguistic characteristics that were later used as domain-expert features to automatically categorize eyewitness reports from non-eyewitness reports using random

forest (another supervised machine-learning algorithm). The eyewitness reports classifier trained on domain expert features well performed as compared to the model trained on text features (i.e. most frequently occurring terms). I also concluded, however, that a balanced class distribution in the data also enhances the output of the classifier.

Finally, I addressed the challenge of missing context and semantics of shared information in social media posts that causes the loss of a huge amount of potentially useful information when posts are individually analysed for information. To address this challenge, I analysed social media threads for various disaster-related themes to extract information that adds context and semantics to the initial post. The results showed that disaster-related information can be categorized (in individual post as well as a thread) in six themes i.e. event reporting, location, time, intensity, casualty and damage reports, and help calls. Furthermore, I developed thematic lexicons for disaster-related themes that were used to develop an automated information extraction model. To compare the lexicon-based approach, I also developed word embedding models and stated that on average lexicon-based approach produced high quality results compared to complex word embedding models that require voluminous data to learn the vector space of each word. Furthermore, I used extractive text summarization technique to generate a summary of most important information shared in a thread.

There is a great potential for the methods developed in this research to apply on real-time social media feeds by emergency responders to extract credible and useful information during mass emergency events. These methods are independent of a particular type of social media (e.g. no tweet metadata fields have been used during analysis) and therefore are replicable and reproducible using a wider range of social media platforms.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

SYNOPSIS

1 INTRODUCTION

1.1 MOTIVATION

The omnipresent threat of natural disasters[1] is on the rise worldwide. From deadly earthquakes such as that demolishing community infrastructure in Albania [*Duni and Theodoulidis*, 2019], to hurricanes that caused damage to human life and property in the US [*Cerrai et al.*, 2020], to extreme monsoon flooding in India [*Gupta*, 2020] with many other natural disasters around the world have marked the year 2019. The "new normal" of wildfire season 2019 has witnessed the most devastating events of Australia (2019-20) [*Ward et al.*, 2020] and Amazon rainforest [*Escobar*, 2019] forest fire. The deadly summer heatwave in Europe[2] (2019) setting the temperatures to a record new high is a reminder of shocking impacts of climate change. Similarly, the year 2018 claimed thousands of lives and displaced a hundred thousand people due to many natural disasters[3]. These severe and extreme events link rapidly changing climate with accelerated frequency and magnitude of natural disasters [*Botzen and Van Den Bergh*, 2009]. Moreover, urbanization caused over occupancy of vulnerable areas that has also played its role in mounting the devastating impacts [*Pelling*, 2007].

Disaster response agencies (government and non-government organizations) are the backbone of a country's infrastructure to deal with catastrophic events. These agencies operate at different levels (local, regional, federal, international) and are responsible for different phases (response, recovery, mitigation, preparedness) of disaster management cycle [*Khan et al.*, 2008]. Emergency responders rigorously rely on different types of information that plays a key role to communicate between and within the organizations to mobilize various resources. For example, they seek reports about the event with its location to assess the magnitude of the disaster. Similarly, the reports about injured people and casualties are valuable to dispatch medical services. Moreover, emergency responders also seek for a continuous update of 'what is happening on the ground' known as situational awareness [*Karami et al.*, 2020].

1 An event triggered by a natural hazard

2 https://www.vox.com/world/2019/6/26/18744518/heat-wave-2019-europe-france-germany-spain

3 https://wtop.com/world/2018/11/10-of-the-deadliest-natural-disasters-in-2018/

Hazards, disasters, and other relevant concepts:

- Natural hazards are the elements of the physical environment driven by nature that are potentially harmful to mankind [*Burton*, 1993].
- Natural disasters are events triggered by natural hazards that cause disruption of the functioning of a community and/or cause damage. The impact of natural disasters depends on many factors such as vulnerability (the conditions that reduce people's ability to prepare for a disaster), capacity (the conditions that determine a community's ability to deal with a disaster), and risk (the probability of disruption of community infrastructure in case of a disaster).

According to a report[4] published by World Health Organization (WHO) in 2009 on information management and communication in disasters, the disaster-related information should meet certain standards such as accessibility, accountability, verifiability, relevance, timeliness, and sustainability. The quality of information highly depends on its source. One of the conventional sources to collect information during disasters is on the ground reporting by news media agencies and disaster responders. However, when the catastrophe of a disaster is huge, on the ground reporting becomes challenging and often cause time lapse due to inaccessibility of affected areas and lack of workforce.

Since the emergence of web 2.0, the role of the internet has changed from information provision into a system that supports communication and community building [*Tuten*, 2008]. The advancement in technology and cost-effective availability of internet in most parts of the world have provided new opportunities to create, access, and disseminate information [*Zeng et al.*, 2010] using various platforms that support social networking such as Facebook, YouTube, LinkedIn, and Twitter. These social media platforms are gaining immense popularity. According to an estimation, almost 3.6 billion[5] people used at least one type of social media platform in the year 2020. These users produce a huge amount of data on these platforms every single day. This data provides endless opportunities to analyse latest trends [*Asur et al.*, 2011], understand public mental health [*Gruebner et al.*, 2017], to know the sentiments of people [*Wan and Gao*, 2015], and collect information during crisis events as victims become active information collectors and distributors [*Spinsanti and Ostermann*, 2013].

One of the most commonly used social media platforms is Twitter with millions of daily active users worldwide. This is a microblogging platform (microblogs called tweets) with around 500 million[6] tweets posted every single day. The

4 https://reliefweb.int/sites/reliefweb.int/files/resources/753BA3EC98D0AE21852576A40078B90C-PAHO_CommGuide_ResponseTeams_dec09.pdf
5 https://bit.ly/3ivTaip
6 https://www.dsayce.com/social-media/tweets-day/

huge amount of tweets publicly posted on Twitter provides an opportunity to harvest this data to gather information during disasters. The reports about a disaster posted on Twitter include first-hand observations by users [*Bruns and Burgess*, 2012], news agency reports [*Wasike*, 2013], and retweets consist of repetitive information from mixed sources [*Mendoza et al.*, 2010]. Table 1.1 shows an example of publicly available tweets posted during different natural disasters that can potentially be a source of information for emergency responders.

Table 1.1: Example of publicly available tweets posted during different disasters (Accessed online at 11:00 3.8.2020)

Tweet	**Potential Information**
An earthquake with a preliminary magnitude of 5.3 hit eastern Japan on Thursday, May 17, 2018.	Report, magnitude, location, time and date of an earthquake
Bridge collapsed at Baksa Assam due to flood. Assam reels under flood as pandemic wears on. Around 29 lakh people from 26 districts have been affected by floods this year. As many as 108 have died. Several rivers and their tributaries were flowing above the danger level.	Damage and casualty report of a flood with its location
As daylight breaks in Hokkaido more reports coming in of collapsed walls from buildings with danger of collapsed homes and buildings due to compromised structures from the 6.7 earthquake or from aftershocks	Situational update of an earthquake with a damage report

Li and Rao 2010 highlight that people's response to disasters on Twitter is even faster than the response of organizations. Twitter users are like sensors [*Crooks et al.*, 2013] who share locational information in their tweets, which is very important to map the disaster region. Dugdale *et al.* 2012 analyse the role of Twitter during the Haiti earthquake 2010 and state that information shared by citizens during the event has proved to be very useful for saving human lives. The information content on Twitter often carries situational updates [*Verma et al.*, 2011], damage and casualty reports [*Terpstra et al.*, 2012], and cover all major aspects of disasters from multiple angles by multiple users [*Oh et al.*, 2010]. However, Dugdale *et al.* 2012 also report on many challenges associated with Twitter data, such as information overload, unstructured text, and the velocity of uploaded data. Figure 1.1 summarizes the opportunities and challenges social media provides in terms of an information source for crisis management in comparison to their needs.

The automatic extraction of information from text corpus has created enormous opportunities to analyse a huge amount of unstructured data i.e. social media posts, in virtually no time. Information extraction (IE) is defined as the process of identifying structured information from unstructured or semi-structured text

Figure 1.1: A high-level comparison between social media as an information source with crisis management needs. Source: F. O. Ostermann, adapted from personal communication, 13.08.2020

[*Jiang*, 2012]. The goal of information extraction is to identify and extract relevant pieces of information from a text corpus. For example, consider the following text:

Heard an earthquake happened somewhere in Japan hope yall are safe!!

A typical IE model will extract two pieces of information, i.e. i) event: earthquake, ii) location: Japan.

Téllez-Valero *et al.* 2005 introduced the idea of IE as a classification problem. They claim that multiple words in a relevant segment save the context and can extract interesting information. They applied their IE system on text from newspaper articles related to natural disasters and classified relevant information into various disaster-related information categories using the SVM machine-learning algorithm. Figure 1.2 shows a typical IE model.

Social media communication is different than the conventional ways of written communication such as emails and blog posts [*Aggarwal et al.*, 2012]. Tweeters often write short and to the point tweets using various slangs and abbreviations [*Khan et al.*, 2014], misspelt words [*Wu et al.*, 2010], and use colloquial expressions [*Abel et al.*, 2011]. Using disaster-related keywords to collect relevant tweets is not enough to retrieve the relevant information. Sakaki *et al.* 2013 state that keywords such as earthquakes and floods are sometimes used as metaphors on social media and require disambiguation to understand. Tweets are produced at high velocity and in bulk quantity [*Imran et al.*, 2013]. Depending on the scale of a disaster, availability of enough human annotators to filter out irrelevant information from the relevant information cannot be guaranteed. Many researchers have developed classifiers using machine learning algorithms to classify relevant and irrelevant information such as [*Hagras et al.*, 2017] and [*Stowe et al.*, 2018]. These classifiers use hand-annotated training dataset to learn the classification. These training datasets, however, require time as well as human and financial resources to be prepared.

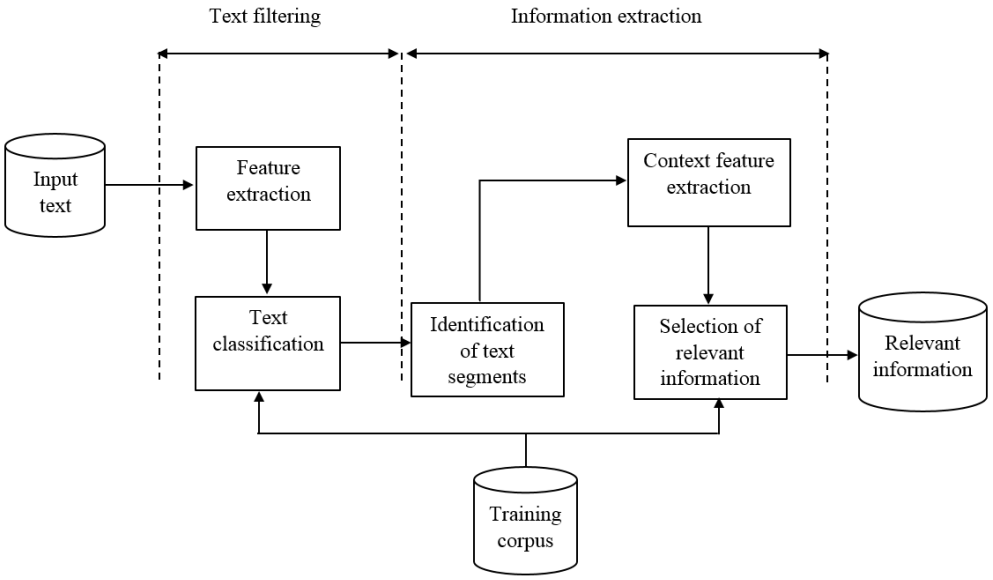


Figure 1.2: A typical IE architecture system capable of extracting relevant text segments. Adapted from *Téllez-Valero et al.* [2005]

Disaster response agencies and practitioners not only seek for new information sources but also for the high quality of shared information. Social media and other online sources have a known reputation for disseminating false information [*Del Vicario et al.*, 2016]. In crisis events, the purposeful spread of misinformation and also the accidental spread has been reported[7] [8]. However, results by [*Starbird et al.*, 2014] reveal that corrections to the misinformation emerge in the Twitter stream and they have distinct patterns (observable by plotting number of tweets in a particular time) than misinformation. They also claimed that these patterns could help detect misinformation automatically from a Twitter stream. Gupta *et al.* 2013 identified 10,350 unique tweets posted during Hurricane Sandy 2012 with fake images as compared to 5,767 unique tweets containing real hurricane destruction images. Zhao and Rosson 2009 argue that source credibility is one of the criteria used to determine the value of a piece of information and its trustworthiness. Truelove *et al.* 2015 state that first-hand reports (also known as eyewitness) are considered more credible than other sources of information on Twitter because these are personal experiences and/or observations of the users. They also developed an eyewitness reports taxonomy generalized on eyewitness reports published during various natural and anthropogenic disasters. Moreover, Fang *et al.* 2016 also identify different types of eyewitness reports from natural and human-induced disasters without considering the role of Spatio-temporal characteristics of natural disasters in identifying eyewitness reports. Doggett and Cantarero 2016 identified a set of linguistic features (such

7 http://gofwd.tumblr.com/post/34623466723/twitter-is-a-truth-machine
8 https://www.forbes.com/sites/kashmirhill/2012/10/30/hurricane-sandy-and-the-flood-of-social-media-misinformation/

as first-person pronouns, location markers) to classify between eyewitness and non-eyewitness reports and events. However, geolocation information in finding eyewitness events is central to their research despite the low availability of such data.

Most current research has focused on extracting relevant information at the level of individual tweets such as [*Gupta et al.*, 2013]; [*Spence et al.*, 2015] that is to say treating every single tweet as a potentially relevant piece of information, and classifying it as relevant or not without considering retweets or replies by other users. This might be problematic because tweet character limits constrain Tweeters use of conventional written communication style and somehow restrict them to be precise and to the point. As a result, information shared in tweets is often missing context and semantics. Abel *et al.* 2011 suggest exploring links and URLs posted in tweet content to add semantics to the shared information. Their methods focus on extracting facet-value by using semantic enrichment so that the tweets are discoverable for Twitter users and return results that are more meaningful during content exploration. boyd *et al.* 2010 explore Twitter conversations (known as threads) to analyse the characteristics of retweeting practice by users. However, their research ignores the potential of Twitter threads in enriching the semantics of information shared in the first tweet. Therefore, Twitter threads can be explored as a potential source of information to add context and meanings to the first tweet as well as to extract more in-depth information about the topic.

The above mentioned challenges show that the potential of using social media as an information source during natural disasters is usually undermined. A variety of automated methods are available to capture and store relevant information from social media feeds. However, these methods generalize all sorts of emergency events (human-induced or driven by natural phenomena) ignoring the particular information needs of emergency responders during a natural disaster. Therefore, this research aims to explore efficient information extraction methods from short, noisy, and unstructured text from social media during natural disasters. The objectives of this research are to:

- Analyze the performance of text classification algorithms to resolve keyword ambiguities to filter noise from information and to optimize tweet classification during natural disasters.
- Develop an eyewitness reports taxonomy particular to natural disasters and identify features that improve the performance of machine learning algorithms.
- Use the potential of social media threads as a source of information to add context and semantics to the information shared during natural disasters.

To meet the objectives of this research, I investigated the following research questions:

RQ1. How well does a machine-learning algorithm perform concerning text classification of information content for another event of the same nature, when the classifier was trained using data from a similar event but a different location?

RQ2. How can an eyewitness report's taxonomy categorize social media posts published during a natural disaster?

RQ3. How can the performance of machine learning algorithms improve with content-based linguistic features in identifying eyewitness reports?

RQ4. What information and semantics can the analysis of Twitter threads (as opposed to single tweets) add?

RQ5. What is the role of a lexicon-based approach in extracting relevant information from text corpus?

1.2 CONTRIBUTION OF THE THESIS

Figure 1.3 shows the core research interest of this project. Moreover, the contributions of this research are:

- The development of a new technique that is capable of improving the accuracy of text classification algorithm without investing time in preparing new training dataset.
- The designing of eyewitness reports taxonomy that is exclusive to natural disasters.
- Identifying a set of eyewitness features that improves the performance of machine learning algorithm in identifying eyewitness reports.
- Analysing the role of social media threads in adding the semantics and context of information shared in individual social media posts.

1.3 STRUCTURE OF THE THESIS

This dissertation consists of two complementary parts. Part I (Synopsis) provides a detailed overview of the research carried out in the scope of this thesis. Following the introduction, Chapter 2 (Background) provides a summary of the information necessary to understand the current state of research, and the research gaps that led to the conducted work. Chapter 3 (Methods) provides an overview of data collection steps used in the analysis, and the methodological approach. The main findings from Publications 1-3 are thematically presented in Chapter 4 (Results and interpretation). Chapter 5 (Discussion) discusses the various aspects of social media data that pose challenge in analysing those data as an information source. A summary of contributions and insights gained in the thesis, and an outlook of future research are given in Chapter 6 (Conclusion

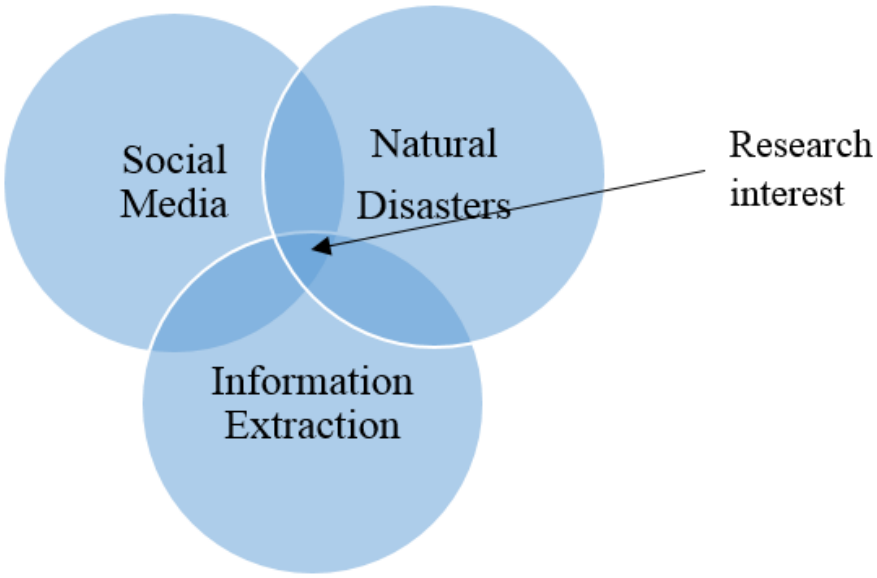


Figure 1.3: Core research interest of this project

and outlook). Part II (Publications) consists of the three research papers written over the course of this dissertation:

Publication I: Geographic variability of Twitter usage characteristics during disaster events.

Zahra, K., Ostermann, F. O., & Purves, R. S. (2017). Geographic variability of Twitter usage characteristics during disaster events. *Geo-spatial information science*, 20(3), 231-240.

PhD candidate's contributions: Coordinated and developed research ideas in collaboration with co-authors. Performed all the experiments for analysis of the data. Wrote the draft of the paper and incorporated several rounds of feedback from the co-authors.

Publication II: Automatic identification of eyewitness messages on twitter during disasters.

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on Twitter during disasters. *Information processing and management*, 57(1), 102107.

PhD candidate's contributions: Initiated the collaboration with QCRI and developed research idea with co-authors. Conducted manual analysis of data for taxonomy and characteristics. Developed machine learning models with other co-authors. Wrote the draft manuscript and incorporated co-authors' feedback.

Publication III: Towards an automated information extraction model from Twitter threads during disasters.

Zahra, K., Das, R. D., Ostermann, F. O., and Purves, R. S. Towards an automated information extraction model from Twitter threads during disasters (submitted).

PhD candidate's contributions: Developed research idea in collaboration with co-authors. Gathered and processed the data. Developed methodology in collaboration with co-authors. Performed parts of the analysis. Wrote the draft manuscript and incorporated co-authors' feedback.

2

BACKGROUND

In this chapter, I present the background on relevant concepts of natural disasters and the disaster management cycle. Moreover, I will discuss the information needs of emergency responders during crisis events. Besides, I will explore various methods developed to extract information from social media and the challenges associated with the analysis of unstructured and informal microblogs for information extraction.

2.1 NATURAL DISASTERS AND THE DISASTER MANAGEMENT CYCLE

Natural disasters are complex phenomena capable of disrupting human life and causing widespread damage and environmental losses. These disasters have the power to affect our society and our way of life substantially and consistently not only in developing countries but also in developed countries, such as Japan, Italy and the US [*Alexander*, 1993]. However, the developing countries suffer more because of low resilience [*De Zeeuw et al.*, 2011] and high vulnerability. Natural disasters can be geophysical (e.g. earthquakes), hydrological (e.g. floods), climatological (e.g. forest fires), meteorological (e.g. storms), biological (e.g. diseases), or a combination of many of these types [*Bryant et al.*, 2005]. These disasters are different and have distinct temporal and spatial characteristics. For example, earthquakes last from few seconds to several minutes as compared to floods, which usually spread over a long period ranging from a few hours to days, weeks or months. On the other hand, droughts develop over weeks to months or even years. In terms of spatial characteristics, depending on the intensity, earthquakes affect a smaller region compared to floods, forest fires, and droughts that usually spread over a vast region. These spatial and temporal characteristics shape the need for information during disaster management [*Aubrecht et al.*, 2013].

The disaster management cycle (Figure 2.1) covers a broad range of activities designed to contain the impact of disaster events by reducing its adverse effects on humans. When a disaster hits, the response is the first phase that starts immediately during or after the impact of the disaster to meet the basic needs of the victims. Depending on the aftermath of the disaster, the response phase can last from a few days to several months.

The cycle then enters the recovery phase, where settlements and infrastructure are reconstructed and redeveloped. This is followed by mitigation, where meas-

ures are taken to minimize the effects before the impact of a future disaster. Finally, in the preparedness phase, disaster response agencies are prepared with the means to help the community to cope with anticipated impacts.

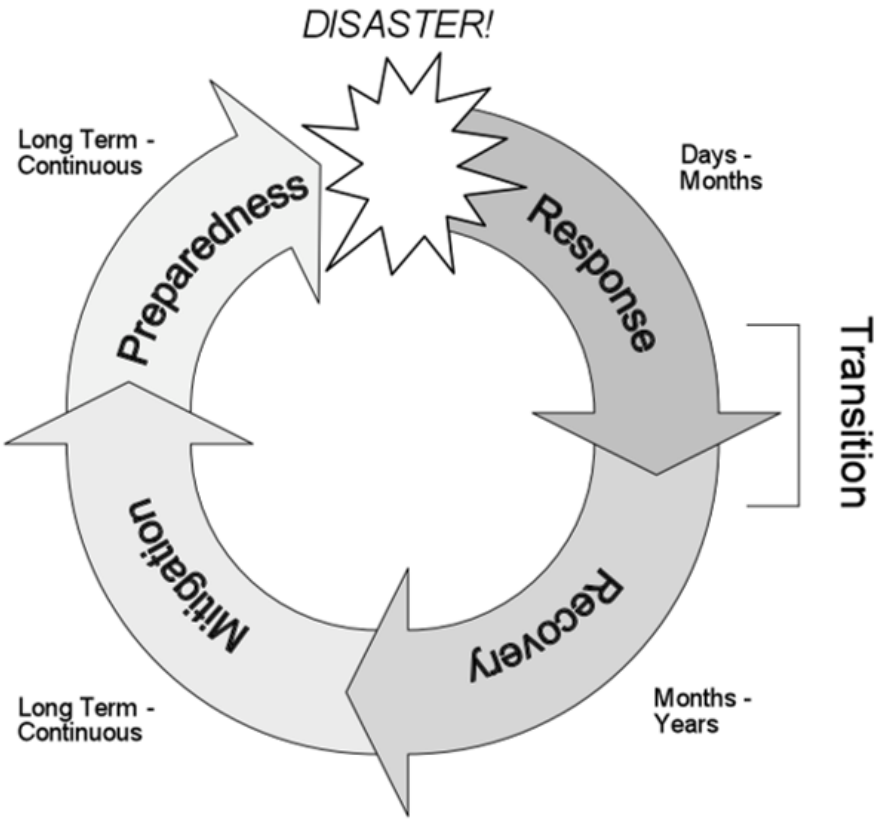


Figure 2.1: Disaster management cycle. Source: *Haddow and Bullock* [2004] as cited in *Howden* [2009]

2.2 INFORMATION NEEDS DURING DISASTERS

When a catastrophe such as a tsunami or an earthquake happens, there is an urgent need for different kinds of information to be shared among the affected population, their relatives and friends, disaster responders, and various government and non-government organizations. This information need can broadly categorize as follows:

- What happened? (Report about the disaster event).
- Where did it happen? (Location of the disaster event).
- What is the impact of the event? (Damage and casualties caused by the disaster).
- Are there any help calls? (People in need of help in the disaster-hit region).

- What are the latest updates for all of the above? (Situational awareness).

Emergency responders are undoubtedly the first point of interaction in case of a disaster. From the phases of mitigation and preparedness to the phases of response and recovery, they are highly dependent on information coming from different sources. Information is the most valuable commodity particularly during the response phase to make critical decisions to help the victims.

In the past two decades, Hurricane Katrina (2005) is one of the worst natural disasters in the US history[1] causing more than 1800 casualties[2] and massive damage to the infrastructure is a relatively recent example of lack of coordination and information exchange between various disaster response organizations. Thompson *et al.* 2006 state that one of the reasons for such a massive loss of life was due to inability of disaster responders to validate important information and act timely. Cooper and Block 2007 explain the unpreparedness of the Department of Homeland Security to deal with natural disasters in the US during Hurricane Katrina in the following words:

> *"The Department of Homeland security's obsession with terrorist attacks had undermined the nation's readiness for natural disasters and ironically had made the country more vulnerable to calamity, not less."*

They further describe that affected city and state officials were unprepared to process real-time information to take immediate evacuation decisions that result in the cost of human life. Federal government's information-gathering agencies also complained that due to the lack of information processing they had an incomplete and inaccurate picture of the ruined city and thus were unable to help. The tragic outcome of Hurricane Katrina demonstrated the need for a strong emergency management system during natural disasters [*Bullock et al.*, 2017]. The backbone of an emergency management system is the framework of an accurate and timely flow of information between several responding units.

Laituri and Kodrich 2008 highlight the key contributions of online disaster response community such as writing blogs, uploading pictures, and information dissemination. They also claim that collection and dissemination of this information are usually faster than government organizations however, establishing online disaster response communities require well-developed protocols for data collection, its storage, analysis, and dissemination. Moreover, local needs and infrastructures for disaster response vary in different geographic regions [*Shaw*, 2014], therefore, online disaster response communities is not a well-established concept worldwide.

Transparent and reliable information helps practitioners to correctly assess the damage caused by a disaster and evaluate the needs of a community. The main

1 https://www.nationalgeographic.com/environment/natural-disasters/reference/hurricane-katrina.html

2 https://www.livescience.com/22522-hurricane-katrina-facts.html

source of such information for disaster response agencies is on the ground reports collected by disaster responders or mainstream news media. Journalists and disaster responders use their resources to gather information or personally visit affected areas that cause time-lapse and require many resources. For more than a decade, people have been turning to new online sources such as social networks and microblogging platforms to learn what is happening during or immediately after an event to get the information [*Cobo et al.*, 2015].

2.3 WEB 2.0 AND SOCIAL MEDIA

Web 2.0, the second phase in the Web's evolution (after Web 1.0), has led to the development of web-based communities and applications [*Yin et al.*, 2009]. These advancements in communication technology have further developed many possibilities for how people send and receive information. Social media are one of the outcomes of these advancements. For example, people not only use social media to seek information about trending topics [*Asur et al.*, 2011] or situational updates during an emergency event [*Vieweg et al.*, 2010] but also as the means to disseminate information [*Liu et al.*, 2016] to people in their social circle or the public at large. Social media have become rapidly popular, particularly among youth and professional individuals [*Alabi*, 2013]. According to the Global digital overview report[3] published in 2020, more than 3.8 billion people are the active social media users with Facebook and YouTube ranking at the top.

Goodchild 2007 refers to citizens as sensors, capable of observing and sharing information that can potentially be very useful during natural disasters. Poser and Dransch 2010 also state that social media users can help disaster responders by providing useful information observed by their senses, such as; sight sense can help to observe information about high water levels in case of a flood. Similarly, hearing sense can help detection of creaking sound for earthquake intensity estimation, and smell sense can help observe forest fire in the vicinity. Toriumi *et al.* 2013 analysed information dissemination behaviour of social media users after the great east Japan earthquake in 2011. They concluded that social media users adapt their behaviour to diffuse important information rapidly during a disaster.

2.4 MICROBLOGGING PLATFORM – TWITTER

Twitter is a well-known microblogging platform with millions[4] of daily active users worldwide. Twitter provides its users with an opportunity to share their views and opinions with their friends, followers, and the public at large. Tweets are microblogs of up to 280 characters (140 until 2017) usually characterized as unstructured text [*Das and Kumar*, 2013]. Due to a limited number of characters,

3 https://datareportal.com/reports/digital-2020-global-digital-overview
4 https://www.omnicoreagency.com/twitter-statistics/

tweets are often informal, noisy, brief, contain misspellings, and grammatical mistakes [*Imran et al.*, 2016]. Twitter officially supports 34 languages[5], however, an analysis of 62 million tweets collected over a four week time reveals that those tweets are written in more than 100 languages, at times multiple languages in one tweet [*Hong et al.*, 2011]. Tweeters can mention a particular user using *@username* anywhere in tweet content and use URLs to share more information. They can also reply to a particular user by using *@username* exclusively at the beginning of the tweet. Hashtags (represented by a symbol and followed by a keyword) used to tag a tweet to a particular topic [*Chang*, 2010]. Hashtag keywords are usually comprised of more than one word concatenated together without space.

Tweets are generated in bulk (on average 6,000 tweets per second[6]) at a high velocity [*Rajadesingan and Liu*, 2014] making it impossible to manually collect and analyse every tweet. To address this problem, researchers have developed many (semi-) automated tools to collect tweets from the Twitter Application Programming Interface (API) and analyse the content. **TweetTracker** [*Kumar et al.*, 2011], a tool that was designed to support Humanitarian Aid and Disaster Relief (HADR) operations for monitoring and analysing tweets during a Cholera outbreak in Haiti. This tool was able to collect tweets in near real-time (unlike **Twinder** [*Tao et al.*, 2012] another tool to collect tweets that does not support working with streaming API) based on keywords used in tweet content and hashtags. The main disadvantage of TweetTracker was that it was unable to analyse tweet content for its relevance. To overcome this limitation, *Abel et al.* [2012] developed **Twitcident** that was capable of filtering relevant tweet content. However, their approach was based on rule-based decision making that was not robust to adapt to different events. Another tool **Tweedr** developed by Ashktorab *et al.* 2014 used machine learning algorithms to detect actionable information from Twitter stream. However, this tool required human annotated data to train those algorithms for every disaster similar to **AIDR** [*Imran et al.*, 2014].

Terpstra *et al.* 2012 demonstrated the use of analysing the tweets as an information source posted during a hurricane called 'Pukkelpop storm' that struck Belgium in 2011 during a music festival. This storm caused five deaths and left many severely injured[7]. They collected tweets based on storm-related keywords such as 'pukkelpop, ppop, and pp11' in Dutch using Twitcident *Abel et al.* [2012] that has a graphical function to display the number of collected tweets identified by the search query per minute. Figure 2.2a shows the tweets three hours before and five hours after the incident started.

5 https://developer.twitter.com/en/docs/twitter-for-websites/twitter-for-websites-supported-languages/overview
6 https://bit.ly/2PN9gIi
7 https://www.bbc.com/news/world-europe-14586001

Terpstra *et al.* 2012 state that sudden peaks in collected data are a strong indication of an event. They also observed a small peak encircled in red appeared almost half an hour before the actual storm hit the location. Their analysis reveals that those tweets were about the bad weather conditions in the festival area. These tweets could be considered as an unofficial *warning* before the storm. They also applied Twitcident's topic-filters of damage and casualty reports to analyse the tweet content. Figure 2.2b and 2.2c show a higher number of tweets related to damage and casualty reports during the storm. Their findings support the use of Twitter as a data source during disaster events.

(a) Spikes in the number of tweets/min help event detection

(b) Damage reports

(c) Casualty reports

Figure 2.2: Twitter activity captured and analysed using Twitcident during Pukkelpop storm 2011 incident. Source: *Terpstra et al.* [2012]

Another in-depth study [*Alam et al.*, 2018] of Twitter content shared during three hurricanes in 2017 in the US shows Twitter not only produces a huge volume of tweets but also disseminate valuable information. They collected tweets from Twitter streaming API during these hurricanes based on event-specific keywords and hashtags. Then they analyse the proportion of relevant and irrelevant content in the tweets (Figure 2.3).

In this case, the relevant category is a combined version of all the information categories that are important for disaster responders. On average, the proportion of relevant messages is above 50 percent during all three hurricanes. These

results also support the usability of Twitter content posted during natural disasters.

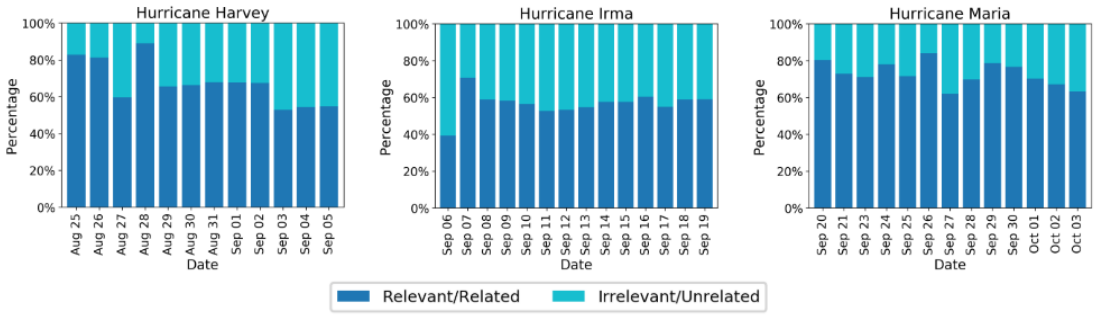


Figure 2.3: Daily proportion of relevant and irrelevant tweets for three hurricanes. Source: *Alam et al.* [2018]

2.4.1 *Challenges related to processing microblogs*

Twitter has established a sound reputation among researchers for over more than a decade as an opportunity to extract information during crisis events [*Truong et al.*, 2014] ; [*Cobo et al.*, 2015]. However, there are many challenges associated with the processing of microblogs [*Bifet and Frank*, 2010]. Figure 2.4 shows broadly categorized challenges. In the following sections, I will discuss these challenges in detail.

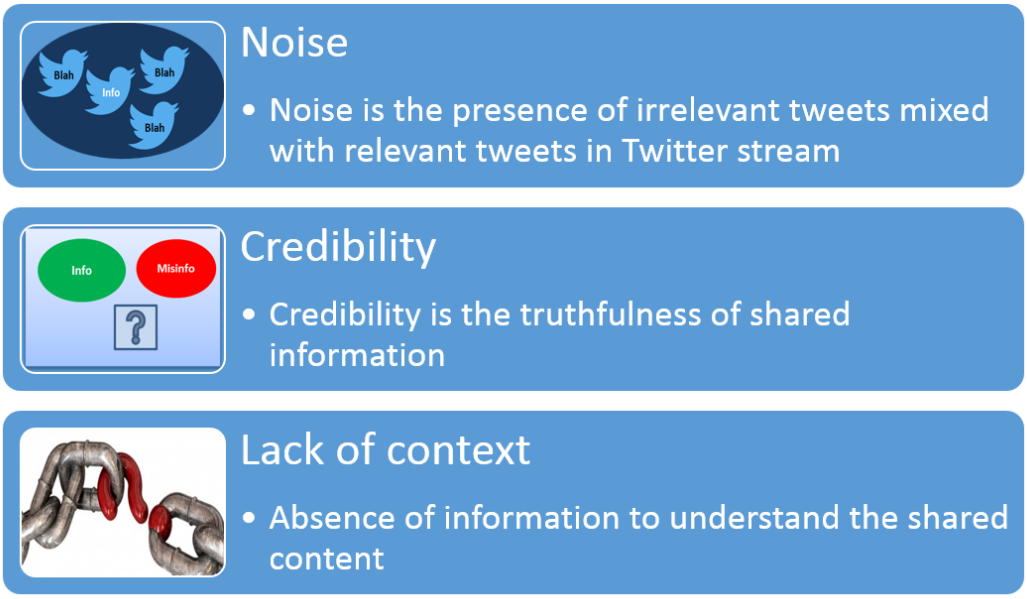


Figure 2.4: Challenges associated with the processing of microblogs

2.4.2 *Noise*

The free form of writing tweets indirectly becomes a source of noise in tweets that makes it a challenging task to extract relevant information. When the tweets are collected based on particular keywords, a large amount of noise is also a part of the data [*Sakaki et al.*, 2010]. This is because of language ambiguities such as polysemy [*Naseem and Musial*, 2019] and the use of metaphors [*Ghosh et al.*, 2015]. Spam is another type of noise that includes advertisements, automatic weather and radio updates [*Petrović et al.*, 2010]. Spammers use trending topic keywords and hashtags to inject their messages into Twitter stream to get visibility [*Rowe et al.*, 2012]. These spam messages constitute a big amount of data produced by Twitter.

Researchers developed various methods using machine learning algorithms to automate the process of filtering irrelevant information. Verma *et al.* 2011, Truong *et al.* 2014, and Cobo *et al.* 2015 used Naïve Bayes, a supervised machine learning algorithm, trained on different set of features such as unigrams, bi-grams, tweet content and metadata features and achieved an accuracy of above 75 percent for detecting the relevant information. Similarly, Musaev and Hou 2016 used Support-Vector machine algorithm using bag-of-words (BoW) model and word2vec word embeddings and achieved over 90 percent accuracy in determining the relevant class.

Although machine learning algorithms show promising results in classifying tweets as relevant or not relevant; however, they require a well-prepared training data for high performance. Crowdsourcing is one of the possible solutions to prepare training data as demonstrated by Imran *et al.* 2014. However, this solution comes with its limitations such as cost, and lack of quality and control. Verma *et al.* 2011 evaluated the performance of various features in machine learning algorithms to predict the right class. They state that algorithms trained on one event perform well on the same kind of events. Their results reveal the accuracy of over 80 percent on categorizing tweets that contribute to situational awareness. However, their research ignores the role of geographic features and toponyms in reporting the disaster events.

2.4.3 *Information quality – credibility*

The quality of shared information is very important particularly when the information is used for decision-making during emergency events. Goodchild and Glennon 2010 highlight that information quality of user generated content (UGC) is a major concern because there is no quality assurance as in the case of official information collection and dissemination sources. Senaratne *et al.* 2016 claim that the quality of UGC can be assessed by analysing the credibility of the contributor and shared content.

The concept of the credibility of information has received significant attention since the spread of the internet as a new and unique source of information in the late 1990s. Credibility can be considered as the measure of the believability of information [*Castillo et al.*, 2011], or accuracy, objectivity, and fairness of information [*Hilligoss and Rieh*, 2008]. The increasing reputation of Twitter as a real-time source of information [*Osborne et al.*, 2012] has attracted researchers to study the credibility of social media content as it has a known reputation for disseminating false information [*Vicario et al.*, 2016]. Truelove *et al.* 2014 stated eyewitness reports (personal observations of users) are considered a credible source of information on social media. When social media users share first-hand reports it is likely to be more credible than non-eyewitness reports or retweets. Therefore, to assess the credibility of social media content, many researchers have analysed the data for eyewitness reports.

Takahashi *et al.* 2015 observed the tweeting behaviour of users before, during, and after the Typhoon Haiyan in 2011 and state that 4.9 percent of posted tweets were personal/first-hand observations. The reason for such a low number was the disruption of power supply, phone lines, and internet access immediately after the hurricane hit the area. Reuter *et al.* 2017 surveyed to analyse the use of social media during emergencies. A relatively low number of users (5 percent) from their sample use social media to share information, however, 38 percent out of those shared reports were eyewitness or first-hand reports. Truelove *et al.* 2014 designed and tested a generalized taxonomy of eyewitness reports for many types of disasters including natural and human-induced. However, natural disasters have distinct spatial and temporal characteristics than from human-induced disasters. Therefore, a taxonomy of eyewitness reports exclusive to natural disasters can better help disaster responders to filter information during a disaster event.

Many researchers have developed machine-learning classifiers to detect eyewitness reports from non-eyewitness reports using social media data. Table 2.1 shows a non-exhaustive list of these methods with the features used to train the models. In most of the cases, content-based features are comprised of the length of a tweet (character and word), presence of URLs and emoticons, and user-based features are comprised of the number of friends, followers, verified account, and the number of statuses. However, these methods are mainly dependent on Twitter data structure for various features from tweet metadata or user metadata.

2.4.4 *Lack of context and semantics – Twitter threads and text summarization*

Tweeters can write a limited number of characters in a tweet, restricting them to share to the point information. Although Twitter allows 280 characters, however only about 1 percent of tweets hit this limit, only 12 percent are longer than

Table 2.1: A non-exhaustive list of research with features used to determine eyewitness reports

Research	Features
Castillo et al. [2013]	Content/message-based features, user/source-based features, topic-based features, propagation-based features
Gupta and Kumaraguru [2012]	Content/message-based features, user/source-based features
Gupta et al. [2014]	Content/message-based features, user/source-based features, Metadata-based features, Network-based features, URLs and links
Alrubaian et al. [2018]	Content/message-based features, user/source-based features, hybrid features
Abbasi and Liu [2013]	User/source-based features
Kang et al. [2012]	Content/message-based features, user/source-based features, hybrid features

140 characters, and only 5 percent are longer than 190 characters[8]. This practice indicates that tweet content conveys limited information and is often missing the context and semantics of shared content. Current research in the field of information extraction from Twitter during natural disasters has focused on classifying tweets as *information* or *not information* by looking at individual tweets separately [*Zahra et al.*, 2017]. This practice discards a big amount of data that can potentially be a relevant piece of information.

Naseem and Musial 2019 use deep intelligent contextual embeddings to add context, semantics, syntax, and sentiment knowledge of words to determine the sentiment of a tweet. However, their method is limited to sentiment analysis and does not address information extraction possibilities. Romero and Becker 2019 propose a hybrid semantic enrichment framework to classify event-related tweets. Their framework consists of named entity recognition, external document enrichment, and semantic enrichment using linked open data cloud. They extract features from tweet content and external documents to enrich the semantics of a tweet to determine its relevance to a particular event. However, their approach ignores the possibility of enriching information shared in one tweet from another tweet.

It is a common observation that some Twitter posts constitute a conversation [*Purohit et al.*, 2013] despite having a character limit. Twitter users tend to reply or comment on a tweet to form a conversation (i.e. threads). These threads can

8 https://www.axios.com/a-year-after-tweets-doubled-in-size--brevity-still-rules-610efb0f-7799-4874-8d65-a0f3e807b310.html

be a potential source to add context, meaning, and missing information to the initiating tweet or to extract more information about the topic. Buntain and Golbeck 2017 analyse Twitter threads to identify fake news propagation by using content, user, structure, and temporal features. However, their research does not take into account the possibility of enriching the content of one tweet from the other tweets in a thread.

Twitter data is often too voluminous for manual summary and analysis even after discarding irrelevant tweets. When a Twitter thread is used to extract information, the amount of data can be excessive. To address this challenge, text summarization is an efficient tool to minimize a document's size and at the same time retain key information. To create text summaries using various algorithms, there are two main approaches: abstractive text summarization [*Moawad and Aref*, 2012] and extractive text summarization [*Ledeneva et al.*, 2008].

Abstractive text summarization is defined as "the task of generating a short and concise summary that captures the salient ideas of the source text" [*Liu et al.*, 2018]. This implies that new phrases and sentences may be used in the resulting description that are not part of the original text but are suggested by the algorithm to convey the information effectively. Whereas, extractive text summaries "produce a set of most significant sentences from a document, exactly as they appear" [*Ferreira et al.*, 2013]. Extractive summaries, therefore, only include phrases that occur in the source text. In order to extract valuable information posted to Twitter status during sports events, Nichols *et al.* 2012 used extractive text summarization. In order to provide an event summary, they also concatenated multiple text summaries.

To summarize, a variety of automated methods is available with various research gaps, such as the requirement of well-prepared training data to eliminate noise, a generalized eyewitness reports taxonomy for all types of disasters, and lack of context and semantics in short social media posts. Therefore, to address these research gaps, the overall aim of this research is to develop such methods that are capable of extracting relevant, credible, and useful information timely from social media posts shared during natural disasters.

3

METHODS

This research focuses on extracting relevant and reliable information from unstructured, informal, and short social media posts published during different natural disasters. Due to the number of tweets produced at a given time, text classification models using machine-learning algorithms have been developed to automate the information extraction process in this research. Information is a relative term having different meanings in different contexts [*Floridi*, 2010]. Therefore, I define the term information for this research in three different contexts. Figure 3.1 shows the conceptual framework of this research, defines the term information and its context, and shows the link between them.

The first context is a very basic level of information extraction where *information* and *not information* classes are broadly defined. In this context, *information* is any social media post that is about an earthquake event with its location. The *not information* class contains all other posts including different types of noise (see section 2.4.2). The second context is about extracting credible information from the tweets that are about a disaster event. Therefore, the *information* is defined as an eyewitness report. Any social media post that reports first-hand personal observations about a disaster event is considered as an *information* tweet. On the same note, a non-eyewitness report is categorized as *not information*. The third context is about extracting various themes of disaster-related information from a social media thread where the initiating post is an eyewitness report posted during/about a disaster event to add more information. In this context, several disaster-related themes such as event report, location, intensity are defined as *information*. Similarly, rest of the content in the threads is *not information*.

3.1 DATA COLLECTION

Twitter provides access to collect real-time tweets via streaming API using many parameters (such as language, location, track etc.) that meet the query criteria. For this research, track and language parameters were explicitly defined using default values for the rest of the parameters. For the track parameter, a list of natural disaster related keywords (Table 3.1) is compiled based on common words in English used to describe those events. The keywords in this parameter are space-sensitive but not case-sensitive. For the language parameter, English was used to collect the tweets. After running a script in R for more than 36 months from April 2016 until July 2019 from all over the world, around 90 Million tweets were collected including retweets. There are some gaps in the

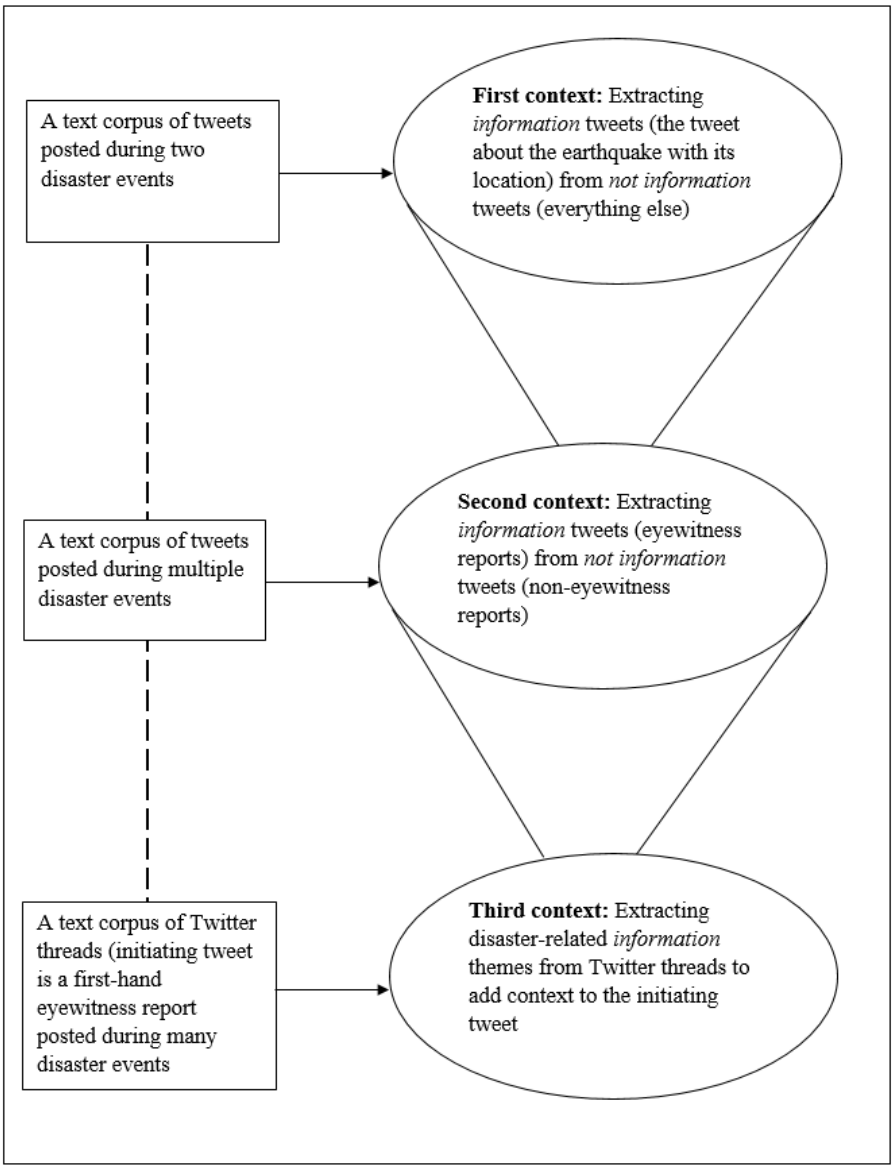


Figure 3.1: A conceptual framework of this research

data due to technical reasons, such as disrupted connection with Twitter API, and rebooting of the system for server maintenance. Moreover, depending on the query and Twitter activity, a subset of actual tweets from Twitter streaming API is the part of this tweet corpus.

3.2 INFORMATION AND NOT INFORMATION CLASSIFICATION MODEL

For the first context in figure 3.1, I selected a time window of 24 hours from 8:57 UTC 24.8.2016 to 8:57 UTC 25.8.2016 to retrieve a subset from the main tweet corpus. This subset was selected because of two major earthquakes that struck Italy (6.2 magnitude) and Myanmar (6.8 magnitude) during this time. The main tweet corpus consists of all the tweets containing any natural disaster keyword

Table 3.1: List of keywords used to query Twitter streaming API

Tsunami	flood	Earthquake
Landslide	earth quake	fore shock
fore-shock	after shock	after-shock
landslide	land slide	Avalanche
rockfall	rock fall	mud slide
mudslide	earth slip	earth-slip
hurricane	cloud burst	heavy rainfall
extensive rainfall	heavy rain	extensive rain
rain storm	forest fire	inundation
overflow	flash-flood	volcanic eruption
volcanoes	volcano	lava

and posted from any part of the world. Therefore, the first query retrieved only those tweets that have the term *earthquake* in its content posted during the selected time window. It resulted in 282,177 tweets (234,620 about Italy and 47,557 about Myanmar earthquake). I manually annotated a subset of data to prepare test and training dataset (details in Table 3.2). I defined two categories to classify the data: *Information* and *Not information*. The *information* class contains tweets that are about the earthquake event with its location, whereas, the *not information* class contains everything else that does not meet *information* class definition. During the manual annotation, both *information* and *not information* classes were balanced in the dataset, however, that is not the case when analysing the full data. The classes were highly imbalanced. I addressed the class imbalance problem in the next information extraction task.

Table 3.2: Manually annotated tweets

Dataset	**Information Italy**	**Not Information**	**Information Myanmar**
Training	350	350	350
Test	150	150	150

To develop the classification model, a commonly used text classification machine learning algorithm Naïve Bayes [*Li et al.*, 2018] was used to classify the tweets based on frequency of terms. The training and test data ratio used to develop classification model was 7:3. This means that 70 percent of annotated data was used to train the model and 30 percent data was used to test the performance of the model.

I trained and tested the classifier on three different cases and six different scenarios to analyse the role of a geographic features in determining the right class.

For the first case, the classifier was trained on 350 *information* tweets from Italy earthquake and 350 tweets from *not information* class. Then the classifier was tested on 300 tweets i.e. 150 *information* tweets from Italy earthquake and 150 tweets from *not information* class. Then for the second scenario, I replaced independent geographic entity feature i.e. Italy with Myanmar keeping rest of the content same in 350 *information* tweets. The following is an example of replacing the geographic entity of Italy with Myanmar.

The original tweet is:

Italy earthquake: 'At least 20 dead including two children' as 'apocalyptic' 6.2 magnitude leaves towns in ruins: https://t.co/Z1p6zegb1r

The tweet after replacing geographic location:

Myanmar earthquake: 'At least 20 dead including two children' as 'apocalyptic' 6.2 magnitude leaves towns in ruins: https://t.co/Z1p6zegb1r

This data was used to train the classifier again and tested on Myanmar test data i.e. 150 *information* tweets from Myanmar and 150 tweets from *not information* class.

For the second case, the classifier was trained on 350 *information* tweets from Myanmar earthquake and 350 from *not information* class. Then the classifier was tested on 300 tweets i.e. 150 *information* tweets from Myanmar and 150 from *not information* class. Similarly, for the second scenario, I replaced independent geographic entity feature i.e. Myanmar with Italy keeping rest of the content same, trained the classifier again, and tested on Italy test data. The following is an example of replacing the geographic entity of Myanmar with Italy.

The original tweet is:

5.2 earthquake!!! All of Myanmar is now awake!

Tweet after replacing geographic location:

5.2 earthquake!!! All of Italy is now awake!

For the third case, the *information* class in training dataset consists of 175 *information* tweets from Italy and 175 *information* tweets from Myanmar. However, 350 *not information* class tweets remain the same. This classifier was first tested on the Italy test data and then on the Myanmar test data to evaluate its performance. Figure 3.2 shows the overall workflow of the classifier with three cases and six scenarios.

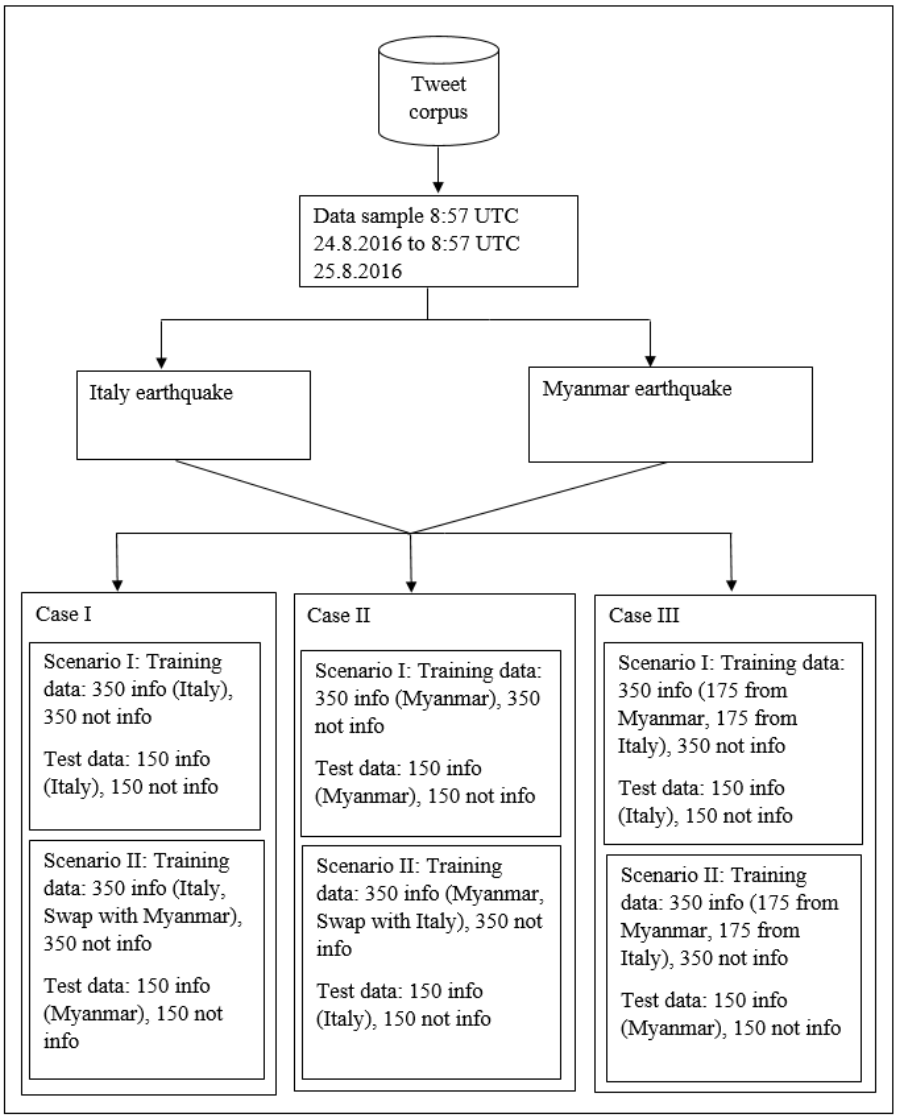


Figure 3.2: Flow chart of *Information* and *not information* classifier

3.3 EYEWITNESS AND NON-EYEWITNESS CLASSIFICATION MODEL

The analysis of the results of the classifier (section 3.2) revealed that not every tweet that is about a disaster with its location is a relevant piece of information for disaster responders because of redundancy and the concern of credibility of shared information. Therefore, the second context in figure 3.1 is about extracting credible information i.e. eyewitness reports. To develop this classifier, two datasets were retrieved from main tweet corpus. The first dataset was retrieved from 1 to 28 August, 2017 for manual analysis of tweets for three natural disasters i.e. earthquake, hurricane, and flood. This timespan was chosen because there was a major Hurricane (i.e. Hurricane Harvey) and a flood event (i.e. Indian flood) occurred with many small earthquakes. The second dataset was retrieved from the complete Twitter data corpus (July 2016 to May 2018,

excluding the first dataset's period to avoid duplication) to produce more annotated data via crowdsourcing for four natural disasters i.e. earthquake, flood, hurricane, and forest fire. Each sample consists 2000 randomly selected tweets from each disaster type excluding retweets as well as duplicate tweets.

3.3.1 *Manual Analysis*

I performed a manual analysis to determine different types of eyewitness reports on the first data sample that comprised 2000 tweets each from earthquake, flood, and hurricane disasters. This analysis was performed in the following three steps:

- **Identify tweet source:** In the first step, it was determined by reading tweet content if a tweet is posted by an eyewitness or not. Tweets were categorized into three classes. Table 3.3 shows the classes, definitions, and examples of each class. The highlighted text in the example are the characteristics that helped identify the tweet in a particular class such as in case of eyewitness class **our office, we, and here** are personalized location markers and first person pronouns and in case of non-eyewitness class **The Mainichi, URL** is the source of the information shared in the tweet. Where it was not possible to identify any eyewitness or non eyewitness characteristics, the tweet was categorized as don't know class. The following sections provide more details on eyewitness characteristics.
- **Identify eyewitness types:** In the second step, only eyewitness tweets were examined to determine if they could further categorize in different types of eyewitness reports. After analysing the tweets, I identified three types of eyewitness reports i.e. direct eyewitness, indirect eyewitness, and vulnerable direct eyewitness. Details are provided in section 4.2.1.
- **Identify eyewitness report's characteristics:** In the third step, various linguistic characteristics (called domain-expert features) were identified from the tweet content of eyewitness reports. This was done by reading and understanding the content of each eyewitness report and then extracting various linguistic features. This step was performed on both data samples and the same characteristics were identified. Tables 3.4, 3.5, and 3.6 show domain-expert features with examples.

3.3.2 *Crowdsourcing*

The types of eyewitness reports identified in manual analysis were then used to collect more data labelled by crowdworkers using a paid crowdsourcing platform i.e. figure-eight (Appen since 2019) on the second data sample. This sample consists of 2000 tweets each for four natural disasters i.e. earthquakes, floods, hurricanes, and forest fires. A job was designed on Figure-eight where

Table 3.3: Eyewitness classes with definitions and examples

Class	Definition	Example
Eyewitness	A message that is posted by an eyewitness who has personally observed the phenomena by any of his senses and reported	*Knee-deep flood outside **our office**. **We** are basically stranded **here** but still have to work work work.*
Non-eyewitness	A message that is posted by anyone else other than an eyewitness	*More landslide, flood warnings as heavy rain lashes wide area of Japan - **The Mainichi https://t.co/9fZqfZFhON***
Don't know	A message for which it is not possible to determine any of the above categories	*My state, everybody. Drought, fire, flood, hail and tornadoes are possible in the same space at the same time. You'd think they cancel out but they just encourage each other.*

the crowdworkers were required to read the tweets and then categorize them according to the developed eyewitness reports taxonomy.

Figure-eight provides various quality control measures to provide as accurate as possible results. Table 3.7 shows these quality control settings used for this experiment. The group of crowdworkers that belong to level two category were allowed to do this task. They are a relatively small group of more experienced people. At the beginning of the experiment, eight test questions were developed to train the crowd which were later increased to 21 questions using trusted judgements from the crowd. The minimum accuracy to pass the test questions and qualify for the job was 80 percent. The crowd was shown five tweets on one page and were supposed to spend at least 50 seconds in reading and understanding the tweets before moving on to next set of tweets.

3.3.3 *Domain-expert features engineering*

The domain-expert features described in tables 3.4, 3.5, and 3.6 are operationalized into features to learn automatic classifiers. Using dictionaries and thesauri, a list of terms was created for characteristics 2, 4, 8, and 12 from table 3.4 and characteristic 1 from table 3.6. For instance, for words related to perceptual senses, such as hearing, seeing, and smelling were used. Similarly, to indicate the severity of an event the terms such as extreme, small, hazardous, and large

Table 3.4: Direct eyewitness characteristics

No.	Characteristic	Examples
(1)	Reporting small details of surroundings	window shaking, water in basement
(2)	Words indicating perceptual senses	seeing, hearing, feeling
(3)	Reporting impact of disaster	raining, school canceled, flight delayed
(4)	Words indicating intensity of disaster	intense, strong, dangerous, big
(5)	First person pronouns and possessive adjectives	I, we, me
(6)	Personalized location markers	my office, our area
(7)	Exclamation and question marks	! , ?
(8)	Expletives	wtf, omg, s**t
(9)	Mention of a routine activity	sleeping, watching a movie
(10)	Time indicating words	now, at the moment, just
(11)	Short tweet length	one or two words
(12)	Caution and advice for others	watch out, be careful
(13)	Mention of disaster locations	area and street name, directions

Table 3.5: Indirect eyewitness characteristics

No.	Characteristic	Examples
(1)	Mention of locations or people the author knows	mom, dad, hometown
(2)	First person possessive adjective	my, our
(3)	Expressing emotions	thoughts, worry, relief
(4)	Reporting safety, damage, missing	missing, safe

were used as a basis and was then expanded with synonyms, whereas the ex-

Table 3.6: Vulnerable direct eyewitness characteristics

No.	Characteristic	Examples
(1)	Warnings and alerts about expected disasters	flash flood warnings
(2)	Associating warnings with current weather situation	flash flood alert with rain
(3)	Expressing emotions	hate, disgust, anger, scare

Table 3.7: Quality control measures

Measures	Setting
Group of crowd works	Level 2
Test questions	8+13
Test questions accuracy	80 percent
Time spent on each judgement	50 seconds

pletives were based on Wiktionary[1] enriched by various slangs[2]. For caution and advice, words and their synonyms were searched in the dictionary.

For characteristic 9 in table 3.4, a list of daily routine activities[3] was developed. The characteristics 3 and 10 in table 3.4 were operationalized by collecting the words directly from the message content. In Table 3.4, the remaining features (5, 7, and 11), i.e. first-person pronouns and adjectives, exclamation and question marks, and short message length, were simple to enforce by inserting the corresponding terms (I, me, ...), characters (exclamation and question marks), and counting the words in the message.

There were a few characteristics in tables that were overlapping with each other. Therefore, only one characteristic was operationalized and the other one was dropped. These are characteristic 6 in table 3.4 and characteristic 2 in table 3.5 with characteristic 5 in table 3.4. Similarly, characteristic 2 in table 3.6 was overlapping with characteristic 3 in table 3.5.

A list of terms derived from indirect eyewitness reports was developed for characteristic 3 and 4 in table 3.5. Stop Words[4] were removed. The number of occurrences of matching terms (only uni-grams) for each of the operationalized characteristics was calculated and in the case of short message length, binary absence or presence was considered.

1 https://en.wiktionary.org/wiki/Category:English_swear_words
2 https://www.speakconfidentenglish.com/english-internet-slang/
3 https://www.vocabulary.cl/Lists/Daily_Routines.htm
4 http://www.lextek.com/manuals/onix/stopwords1.html

The first characteristic *reporting small details of the vicinity* in table 3.4 and the second characteristic *associating warnings with current weather situations* in table 3.6 were too abstract to operationalize and therefore were discarded. Similarly, the last characteristic *mentions of disaster locations* in table 3.4 and first in table 3.5 were repetitive and also discarded because tweets are often too short and informal that location identification from microblogs is another aspect of extensive research. In addition to domain-expert features, the other kind of features used were based on the bag-of-words (BoW) model. In particular, the features of uni-grams and bi-grams were derived from the textual content of tweets and their TF-IDF scores were used to train models for machine learning algorithms.

3.3.4 *Supervised machine learning models*

Among different machine learning algorithms, Random Forest is considered best for the classification of textual data [*Xu et al.*, 2012]. Therefore, to automate the process of identifying eyewitness reports from the Twitter stream, I also used Random Forest to develop the models. Two types of features were used to train the model. First, textual features (uni-grams and bi-gram) that were automatically extracted from the tweet content based on their TF-IDF scores. Second, the domain-expert features that were identified in eyewitness reports during manual analysis. For each event type, four different models were trained to compare the results. Following is the detail:

i. A model trained using textual features (baseline): In this case, the first 50 most frequently occurring terms were used based on their TF-IDF score to train a BoW model.

ii. A model trained using domain-expert features: In this case, the classification model was trained only on domain-expert features identified from tweet content during manual analysis.

iii. A model trained using text and domain expert features: In this case, both text-based and domain-expert features were used to train the model.

iv. A model trained using text and domain expert features with class balancing: In this case, a class balancing technique SMOTE [*Chawla et al.*, 2002] was used to balance the classes and then the model was trained on both textual and domain-expert features.

These models were trained and tested on annotated data collected by crowdsourcing. The models performance was evaluated using the cross-validation (10-fold) technique and presented using standard performance evaluation metrics such as precision, recall, and F-measure in chapter 4. Figure 3.3 shows the complete methodology in the form of a flowchart.

3.4 CONTEXT AND ADDITIONAL INFORMATION EXTRACTION MODEL

The analysis of the results of the classifier (section 3.3) revealed that although eyewitness reports are a credible source of information for disaster responders, they are comparatively few in number (table 4.7 and 4.12). A further analysis of the content revealed that these tweets are often missing context to the information shared in the tweet. Therefore, I analysed Twitter threads for identifying additional information and context to the information shared in the initiating tweet (i.e. the third context in figure 3.1).

Twitter streaming API does not support downloading of Twitter threads due to technical design; therefore, data for this experiment were manually collected. First, Twitter was searched for first-hand eyewitness tweets posted during an earthquake. The search string was particularly comprised of eyewitness features as described in [*Zahra et al.*, 2020] such as *I just felt an earthquake*. These query strings returned various publicly available tweets posted during earthquake events. Before selection, every tweet was manually analysed on the following criteria:

- Tweet content should be about an earthquake event.
- Tweet must have at least one reply to form a thread.

First 200 tweets meeting the above-mentioned criteria were selected. The number of tweets in selected threads range from two to several hundred tweets depending on the social circle of the tweeter and popularity of the tweet. Therefore, a threshold value of maximum 10 tweets (in chronological order) in a thread was applied to keep the content of the thread comprehensible for crowdworkers. Two annotators were given the paid job to manually collect the user name, time, and tweet content from all the tweets in 200 threads. The annotators clicked on each URL to open the thread and then copied the required information in an Excel sheet. Table 3.8 summarises the properties of the 200 threads collected.

Table 3.8: Summary values for the 200 threads

Per thread	**Average**	**Median**	**Total**
Number of tweets	6.9	7	1380
Number of unique users	6.44	7	1288
Length of thread (characters)	409.185	402.5	81837
Length of thread (tokens)	70.62	67	14124

3.4.1 *Thematic taxonomy*

To extract disaster-related information from Twitter threads, it is important to know the types of information Tweeters share during disasters and classify those types into appropriate categories. After conducting literature research [*Imran et al.*, 2014]; [*Tapia et al.*, 2011]; [*Ashktorab et al.*, 2014] and discussions with several experts, relevant information during the disaster were categorized into six themes. Table 3.9 shows the themes, their definitions, and examples. The time theme comprised of two sub-themes i.e. relative and absolute time stamps.

Table 3.9: Thematic classification with definitions and examples of relevant content (the example tweet is a fictitious example)

Theme	Definition	Example
Event reporting	Report about the event	I just **felt an earthquake** in California at 12:00. . . shook the whole building. I need help. . . .one building collapsed.
Location	The location where the event occurs	I just felt an earthquake in **California** at 12:00. . . shook the whole building. I need help. . . .one building collapsed.
Time	The time when the event happened (includes absolute as well as relative timestamps)	I **just** felt an earthquake in California **at 12:00**. . . shook the whole building. I need help. . . .one building collapsed.
Intensity	The intensity of the event	I just felt an earthquake in California at 12:00. . . **shook the whole building**. I need help. . . .one building collapsed.
Casualty and damage reports	Includes reports where people are reporting about casualties and damage caused by the event	I just felt an earthquake in California at 12:00. . . shook the whole building. I need help. . . .**one building collapsed**.
Help calls	It includes reports where people are asking for help	I just felt an earthquake in California at 12:00. . . shook the whole building. **I need help**. . . .one building collapsed.

3.4.2 *Crowdsourcing*

The crowdsourcing platform, figure-eight (section 3.3.2) was used to collect the following information from the threads:

- Which themes are present in the initiating tweet?
- Does the thread add additional information to the initiating tweet?
- Which themes are present in the thread?

The crowdworkers were asked to read the first tweet and choose which disaster-related themes are present in the tweet (a check-box based response). Then they were asked to read the whole thread and respond if the whole thread adds more information to what is shared in the first tweet (a binary radio button based response). In case of a positive response, they were presented again with the list of themes to choose which themes were present in the thread (again a check-box based response). In the case of a negative response, the crowd workers were redirected to the next thread.

For the quality control measures shown in Table 3.7, I again chose level two crowdworkers for this experiment. The test questions were eight and the minimum accuracy to qualify for the job was 50 percent. The time spent on reading one thread was again 50 seconds.

3.4.3 *Thematic lexicons*

For four of the disaster-related information themes i.e. event reporting, intensity, casualty and damage reports, and help calls and one sub-theme time (relative) thematic lexicons were developed following a two-step process. Firstly, a list of seed words [*Chen and Skiena*, 2014] was developed by manually searching in the thread corpus for three themes (event reporting, casualty and damage reports, and help calls) and one sub-theme (time (relative)) and combining the other general terms in English that are used to describe these phenomena. These seed words are the basic terms in a lexicon by which the lexicons are further developed. For intensity theme, Modified Mercalli intensity scale [*Wood and Neumann*, 1931] was used to extract seed words. Secondly, I used WordNet that is a large lexical electronic database for English [*Fellbaum*, 2010] to collect seed words definitions. Two annotators, one native English speaker, read and understood the definitions and based on a mutual agreement discard the irrelevant ones. The selected definitions were then used to retrieve the set of synonyms (called synsets) from WordNet and together with seed words were used to populate the lexicons. Table 3.10 shows various characteristics of all thematic lexicons.

Table 3.10: Thematic lexicon characteristics

Theme	Seed words	Total synset definitions	Selected synset definitions	Retrieved synsets	Total number of words in the lexicon
Event reporting	4	61	21	61	65
Time (relative)	11	20	7	18	29
Intensity	56	371	105	393	449
Casualty and damage reports	9	227	80	225	234
Help calls	4	81	21	101	105

3.4.4 *Information extraction models*

Four information extraction models were developed to summarize the content of a thread that is categorized as a positive case to add information to the initial tweet. I used extractive text summarization technique [*Jain et al.*, 2017] to extract the summary of relevant disaster-related information from the whole thread and to discard irrelevant information. Before developing the models, preprocessing steps were performed to develop Twitter threads corpus. First, each tweet was segmented into individual sentences. Then special symbols from the sentences were removed and were concatenated to form a corpus. Then each sentence was tokenized into words. Finally, all stop words were removed such as 'a', 'the', 'is' using stop words lexicons from the NLTK library in Python [*Loper and Bird*, 2002].

The first information extraction model was trained on GloVe that is an unsupervised learning algorithm for obtaining vector representations of words [*Pennington et al.*, 2014]. This algorithm was trained on a corpus of six billion words comprised of Wikipedia 2014 articles and Gigaword 5 dataset. The second model was trained on word2vec, using word embeddings published on CrisisNLP . These embeddings are trained on 52 million Twitter messages posted during various disasters [*Imran et al.*, 2016]. After creating the word embeddings from both algorithms, a cosine similarity matrix was generated. Then this similarity matrix was used to create a graph for each thread, where a node in the graph represented a sentence and the edge between two nodes represented the similarity value. Following that, a TextRank (a variant of PageRank) algorithm

[*Mihalcea and Rarau*, 2004] was applied to the cosine similarity graph to rank the sentences in a given thread.

The third information extraction model was developed as a baseline using a bag-of-words (BoW) approach. In this model, the term frequencies of each token in the text corpus were calculated. Then the term frequencies of all tokens present in a sentence were combined to assign an aggregate score to each sentence, which was then used to rank the sentences in a given thread based on the highest score.

Before developing the fourth information extraction model i.e. thematic bag-of-words (TBoW), an additional preprocessing step was performed on the text corpus known as lemmatization. Lemmatization is the process of finding the normalized form of a word [*Plisson et al.*, 2004]. To develop the model, the thematic tokens from each sentence were extracted by looking up a given token against five lexicons, e.g., event reporting, time (relative), intensity, casualty and damage reports, and help calls. For spatial and temporal (absolute) thematic tokens, a pre-trained neural network-based Named Entity Recognition (NER) model was developed. To boost the performance of the NER, particularly for retrieving the location entities, many spatial rules such as location names being proper nouns or common nouns appearing after spatial prepositions, e.g., at, near, or to were also developed.

The TBoW model is a modification of the BoW model where more weight is assigned to the thematic terms. Therefore, for every token in a sentence, its term frequency was assigned as its respective weight. If the token is also a thematic token (related to one of the six themes), then the weight is assigned by relative thematic magnitude. Here it is assumed that during a disaster, the spatial and temporal aspects, as well as help calls are more important for emergency responders. Therefore, a weight of 10 is assigned to the terms that belong to three themes (location, time (absolute, relative), and help calls), and a weight of five is assigned to the terms that belong to the other three themes (event reporting, intensity, and casualty and damage reports). All non-thematic tokens receive a weight value of one. In many cases, a thematic token appeared more than once in a thread. In this case, all the tokens (even repetitive ones) were considered every time and assigned respective weights. Each weight was then normalized using the maximum weight in a given sentence. Following this, all thematic entities in a sentence were combined and computed an aggregate score for the given sentence. Since the TBoW was based on frequencies, sentences with more thematic tokens in a given thread received a higher score.

To generate text summaries from all four models, the top 30 percent highly ranked/scored sentences were selected from each thread according to the respective model. Figure 3.4 shows the overall methodology of developing threads corpus, performed analysis, and four information extraction models.

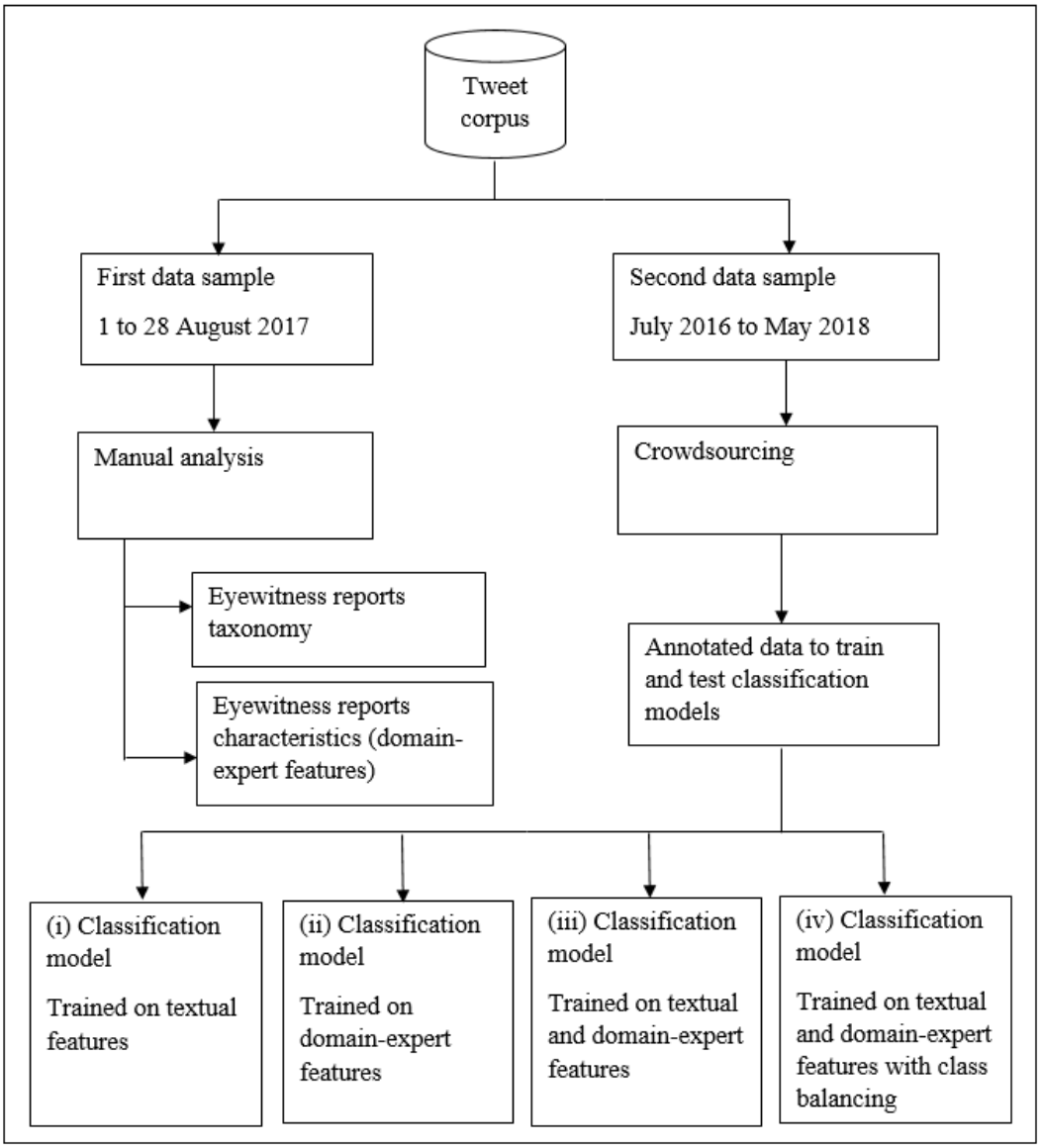


Figure 3.3: Flow chart of eyewitness reports classifier

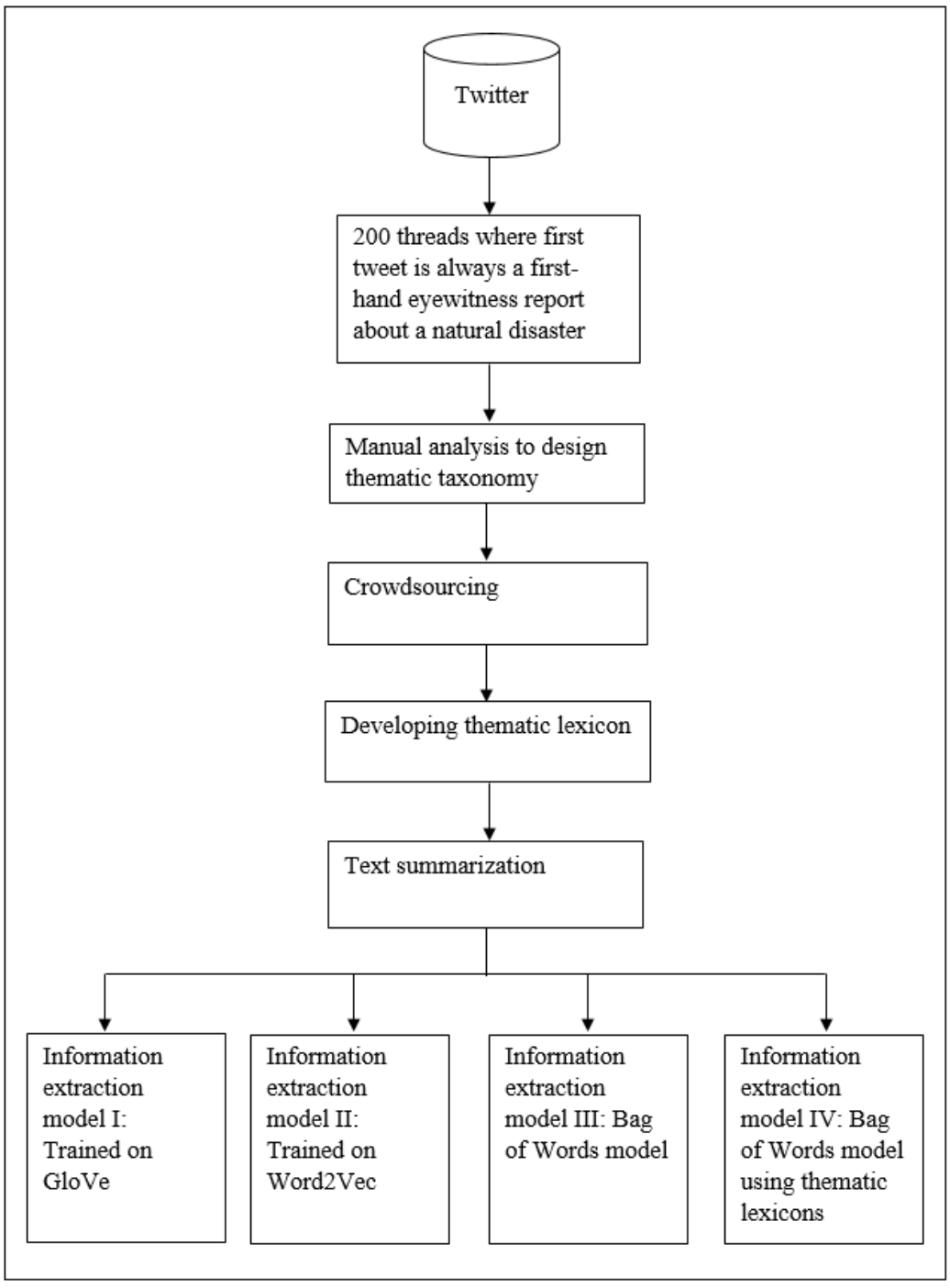


Figure 3.4: Flow chart of information extraction from Twitter threads

4

RESULTS AND INTERPRETATION

This chapter summarizes the main outcomes of this research. Section 4.1 presents the confusion matrix of *information* and *not information* classifier. The following section 4.2 presents the results of eyewitness and non-eyewitness classifier with manual analysis and crowdsourcing experiment results in detail. The final section 4.3 elaborates the results of thematic information classifier from social media threads and text summarization.

4.1 INFORMATION AND NOT INFORMATION CLASSIFICATION MODEL

The test data prepared from Italy that comprised 150 tweets from *information* class and 150 tweets from *not information* class was used for the first case to assess the output of the classifier (Table 4.1). For tweets categorized as containing information, the precision of the classifier was very high i.e. 98 percent, meaning that almost all tweets classified using this method contain information, while a 93 percent recall indicates that a small number of tweets have been wrongly discarded. The output decreased somewhat for the test data from Myanmar (150 *information* tweets and 150 *not information* tweets) but remained relatively high when the classifier was run in a different geographical area by substituting only geographical features in the training data (Table 4.2).

I repeated the same experiment for the second case by switching Italy with Myanmar this time. For the first scenario, the test data (150 *information* tweets, 150 *not information* tweets) was used from Myanmar earthquake data. The results (Table 4.3) for Myanmar show 99 percent precision with Myanmar data-trained classifier. For the second scenario, where test data was from Italy (Table 4.4), although the classifier was trained on Myanmar data with swapped geographic locations, 94 percent precision was very good for Italian earthquake reporting tweets.

Table 4.1: Confusion Matrix for Italy (Case I, Scenario I)

	Actual Class			
Predicted Class	Class	Information	Not Information	Precision
	Information	147	3	98%
	Not Information	11	139	92.7%
	Recall	93%	97.9%	

Table 4.2: Confusion Matrix for Myanmar (Case I, Scenario II)

	Actual Class			
Predicted Class	Class	Information	Not Information	Precision
	Information	133	17	88.7%
	Not Information	12	138	92%
	Recall	91.7%	89%	

Table 4.3: Confusion Matrix for Myanmar (Case II, Scenario I)

	Actual Class			
Predicted Class	Class	Information	Not Information	Precision
	Information	149	1	99.3%
	Not Information	11	139	92.7%
	Recall	93.1%	99.3%	

Table 4.4: Confusion Matrix for Italy (Case II, Scenario II)

	Actual Class			
Predicted Class	Class	Information	Not Information	Precision
	Information	141	9	94%
	Not Information	37	113	75.3%
	Recall	79.2%	92.6%	

Table 4.5: Confusion Matrix for Italy (Case III, Scenario I)

	Actual Class			
Predicted Class	Class	Information	Not Information	Precision
	Information	146	4	97.3%
	Not Information	37	113	75.3%
	Recall	79.8%	96.6%	

Table 4.6: Confusion Matrix for Myanmar (Case III, Scenario II)

	Actual Class			
Predicted Class	Class	Information	Not Information	Precision
	Information	146	4	97.3%
	Not Information	36	114	76%
	Recall	80.2%	96.6%	

The third case experiment was done to show the performance of the classifier when training data was split in half between Italy and Myanmar and tested on original test datasets without swapping geographic locations. The results (table 4.5 and (table 4.6) again indicate very high precision of 97 percent for both Italy and Myanmar.

These results provide an insight into the role of geographic location in determining the right class with improved precision and recall. Although the classification algorithm Naïve Bayes used for this model works on the bag-of-words approach (i.e. it ignores the context and relationship between individual terms), however, swapping geographic locations in training data adds familiar terms that help the classifier in determining the correct class with better precision and recall. This outcome has major consequences, as it shows that training data from other regions can enable the classifier to extract information from tweets posted during a new disaster event in another region with improved accuracy.

4.2 EYEWITNESS AND NON-EYEWITNESS CLASSIFICATION MODEL

This section describes manual analysis and crowdsourcing results with the classification models capable of identifying eyewitness and non-eyewitness tweets automatically.

4.2.1 *Manual Analysis results*

Two annotators manually analyzed every tweet in the sample following the Identify tweet source annotation task guidelines described in section 3.3.1. The outcome of this manual analysis is shown in Table 4.7.

Table 4.7: Frequency of eyewitness, non-eyewitness and don't know cases

Event type	**Sampled tweets**	**Eyewitness**	**Non-eyewitness**	**Don't know**
Floods	2,000	148	113	1,739
Earthquakes	2,000	367	321	1,312
Hurricanes	2,000	296	100	1,604

The number of tweets that eyewitnesses have shared is very low. A total of 148, 367, and 296 tweets were found for floods, earthquakes, and hurricanes, respectively, as posted by an eyewitness. It was not possible to ascertain with adequate reliability for a large number of tweets whether they were shared by eyewitnesses, i.e. the don't know instances in the last column of table 4.7. Moreover, the tweets belong to eyewitness class were manually analysed further for different types of eyewitness reports. The results of this analysis are shown

in table 4.8. In particular, there are three categories of eyewitnesses: (i) direct eyewitness, (ii) indirect eyewitness, and (iii) vulnerable direct eyewitness.

Table 4.8: Frequency of different types of eyewitness reports

Event type	**Direct eyewitness**	**Indirect eyewitness**	**Vulnerable direct eyewitness**
Floods	62	2	84
Earthquakes	354	13	0
Hurricanes	95	16	185

The direct eyewitness category is where people share first-hand knowledge and information of an incident or its impacts. There are numerous ways in which information on incidents can be provided through direct eyewitness reports. Some examples of these reports taken from manually analyzed data from all three forms of disasters are shown in Table 4.9.

The indirect eyewitness category includes direct eyewitness reports from the family/friends or people from the social circle of the person who shares the report. These can be disaster reports received via video or voice communication. There was a small number of tweets in the dataset from this group, however, this is an important and potentially useful category, since this information can play an important role in identifying missing people in a disaster-hit region. Table 4.10 shows various examples of indirect eyewitness reports posted during flood, earthquake, and hurricane events from the data.

The vulnerable direct eyewitness category includes incident reports where the user is present in a high-risk region, for which disaster warnings have been issued. In such eyewitness reports, users share information about an upcoming disaster. These reports were only found in the hurricane and flood dataset due to their predictable nature. Table 4.11 shows some examples of vulnerable direct eyewitness reports from the manual analysis annotation.

4.2.2 *Crowdsourcing results for the second data sample*

An audit was performed on the outcomes of crowdsourcing experiment to determine the inter-rater agreement with the crowdsourced annotation. From each dataset, 50 sample messages were picked and two persons annotated them again. Later on, the results were compared with the crowd annotation, the results showed 90 percent agreement for floods, 92 percent for earthquakes, 82 percent for hurricanes, and 94 percent for wildfires dataset. This high agreement shows that the data annotated using the crowdsourcing platform was of high quality.

Table 4.9: Direct eyewitness reports from manual analysis

Floods direct eyewitness reports
I almost died driving home from work because it started to downpour and flood on the freeway and lightning and its 99 f**king degrees out
No one even notified me that this flood in our area has reached almost 3 feet. but atleast i was able to reach home safely.
Earthquakes direct eyewitnesses reports
Most intense earthquake i've experienced in japan so far...that is
Just felt the house shaking in Tokyo. Been awhile since I felt an earthquake. I hope it wasn't a bad one anywhere on the island.
Hurricanes direct eyewitness reports
Please pray for us right now, the winds and rain is heavy and the hurricane hasn't even hit us yet. #hurricaneharvey2017
This hurricane ain't no joke, the rain and winds are heavy right now. #hurricaneharvey2017

The results from crowdsourcing experiment are shown in Table 4.12. As compared to the manual analysis results in table 4.8, the number of direct eyewitness, indirect eyewitness, and vulnerable direct eyewitness classes have substantially increased with highest earthquake direct eyewitness instances (1557 of 2,000). An in-depth analysis of earthquake data revealed that as soon as an earthquake is felt in a region, many more 'felt reports' are generated about the same event and the Twitter streaming API captures those reports in a sequence because of the keyword match. When the subset was collected through random sampling, some of those sequences became the part of the data. Therefore, earthquake direct eyewitness reports are huge in crowdsourced dataset.

4.2.3 *Classification Models*

The labelled data obtained from crowdsourcing was used to train and test machine learning classifiers. The indirect and vulnerable eyewitness classes were, however, small in number (Table 4.12). In addition, the manual classification revealed that these classes can be difficult to discern as they share some common eyewitness characteristics. Therefore, all three eyewitness types i.e. direct, indirect, and vulnerable were combined into one class, namely eyewitness. Consequently, the classification task consists of three classes: eyewitness, non-eyewitness, and don't know class.

The results of eyewitness reports classification for floods, hurricanes, earthquakes and wildfires, are shown respectively in Tables 4.13, 4.14, 4.15 and 4.16. The last columns of the tables show class distribution. After applying a class

Table 4.10: Indirect eyewitness reports from manual analysis

Floods indirect eyewitness reports
Some days in Thailand has been insane, there has been massive flood on the road to the city (only have image on my dad's phone)
The hsm school and my uncles house are right behind eachother and they were ruined in the flash flood)):
Earthquakes indirect eyewitnesses reports
F*cking hell...my wife and kids are in Tokyo and they're in the middle of an earthquake Jesus Murphy just how crap can one day get?
Was Facetiming my brother in Tokyo when an earthquake. It wasn't strong but took a long time. Glad that he's ok. #tokyo #earthquake
Hurricanes indirect eyewitness reports
Texas has me going for a spin....my hometown was evacuated for the hurricane then an earthquake in Dallas where my entire family is
My city is getting a rain storm from the hurricane and hella winds but that's nothing compared to what's going on god i'm so worried

balancing technique (i.e. SMOTE), the last three rows represent class distribution where the number in parentheses shows how many artificially duplicated instances were added.

Compared to text features (i.e. baseline), the flood results (Table 4.13) show slightly better performance (e.g. F-scores) when using domain-expert features. Similarly, when combining both text and domain-expert features, even better results are obtained. However, compared to the don't know class that achieved an F-score of 0.745, the minority classes of eyewitnesses and non-eyewitnesses still suffer.

In the case of hurricanes (Table 4.14), the domain-expert features tend to give the eyewitness class a good advantage over the plain text features. For the other two classes, however, the difference is not substantial. In addition, better performance was observed when both text and domain characteristics were combined. The minority groups seem to suffer again.

Table 4.15 shows the results of the earthquakes. Surprisingly, domain-expert features do not seem to support much in this situation. In fact, in the don't know class, a substantial drop is observed. The output tends to improve a bit when mixing domain-expert and text features. As there is a major difference in class distributions, these experiments were challenging.

The results of the wildfires are shown in Table 4.16. When using domain-expert features, there was an increase in efficiency compared to using only text features.

Table 4.11: Vulnerable direct eyewitness reports from manual analysis

Floods vulnerable direct eyewitness reports
`Why am I always napping when a flash flood warning comes on to my phone? #scared`
`Those flash flood alerts will kill me one day, they scare the f**k out of me`
Hurricanes vulnerable direct eyewitness reports
`Hurricane Harvey is approaching..Dun dun dun.. first hurricane I will experience in Texas in my new home omg I hope my area doesn't flood`
`I'm so scared I hope this hurricane don't flood my apartment or my car`

Table 4.12: Crowdsourcing results for second data sample

Event type	**Direct eyewitness**	**Indirect eyewitness**	**Vulnerable direct eyewitness**	**Non-eyewitness**	**Don't know**
Floods	320	85	222	551	822
Earthquakes	1557	43	-	200	200
Hurricanes	321	67	77	1199	336
Forest fires	122	44	23	1379	432

However, the don't know class tend to perform better than the other two classes when domain-expert and text features are combined.

Overall, the domain-expert features seem to help achieve better performance compared to only text features in most instances. Also, the efficiency of classifiers seems to be substantially gained by a combination of both text and domain features. In particular, the top features were tweet-length, magnitude token (a bi-gram consisting of earthquake and magnitude), felt, etc. in the case of earthquakes. Whereas in the case of floods and forest fires, the most useful features were personal possessive, disaster effect reporting (e.g., raining, burning), terms suggesting severity (e.g. strong, intense), position comparison, flash flood, etc. In addition, in the case of hurricanes, characteristics such as disaster severity, caution and advice to others (e.g., watch out, warning, be careful), time indicating words, perceptual senses, etc., were defined as helpful ones.

Given that all the datasets were highly imbalanced, the output of classifiers was evaluated after applying more labelled information to minority groups. All of the model training variations seem to outperform using this strategy. The AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curves were drawn from the best models to better understand the efficiency of the classifiers (Figure 4.1), which in most cases are models that depend on a combination of

Table 4.13: Flood results for all four variations of our trained models

Category	Precision	Recall	F-score	Class Dist.
Text-based features (baseline)				
Eyewitness	0.584	0.488	0.532	627
Non-eyewitness	0.706	0.575	0.634	551
Don't know	0.656	0.820	0.729	822
Domain-expert features				
Eyewitness	0.638	0.478	0.547	627
Non-eyewitness	0.642	0.664	0.653	551
Don't know	0.635	0.742	0.685	822
Domain-expert + text features				
Eyewitness	0.717	0.469	0.567	627
Non-eyewitness	0.748	0.653	0.698	551
Don't know	0.648	0.875	**0.745**	822
Domain-expert + text features with class balancing				
Eyewitness	0.760	0.648	**0.699**	815 (+30%)
Non-eyewitness	0.774	0.763	**0.768**	716 (+30%)
Don't know	0.688	0.798	0.739	822

domain-expert and textual features combined with class balancing. Generally, the closer the ROC curve is to the upper left corner, the greater the model's overall precision .

The results of eyewitness and non-eyewitness classifiers have important insights for emergency responders when extracting information from social media during emergencies. First of all, the types of eyewitness reports give deeper appreciation of the availability of different types of credible information. Particularly, indirect and vulnerable direct eyewitness reports are a valuable information resource for the people in warning regions and to get more information about missing persons in badly impacted areas. Secondly, eyewitness and non-eyewitness classifiers are solely based on linguistic features. This fact makes these models robust to use any kind of text-based social media platform to reproduce the results. Finally, the use of SMOTE (class balancing technique) in

Table 4.14: Hurricane results for all four variations of our trained models

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.646	0.419	0.508	465
Non-eyewitness	0.773	0.852	0.810	1199
Don't know	0.605	0.679	0.640	336
Domain-expert features				
Eyewitness	0.655	0.546	0.596	465
Non-eyewitness	0.776	0.881	0.825	1199
Don't know	0.645	0.482	0.552	336
Domain-expert + text features				
Eyewitness	0.734	0.503	0.597	465
Non-eyewitness	0.788	0.910	**0.844**	1199
Don't know	0.686	0.604	0.642	336
Domain-expert + text features with class balancing				
Eyewitness	0.816	0.796	**0.806**	930 (+100%)
Non-eyewitness	0.838	0.843	0.841	1199
Don't know	0.801	0.820	**0.810**	672 (+100%)

these models show that the problem of class imbalancing can be addressed by such solutions and good quality results can be obtained.

4.3 CONTEXT AND ADDITIONAL INFORMATION EXTRACTION MODEL

This section describes the results of crowdsourcing experiment performed on Twitter threads to assess the potential of a thread to provide additional information to the initiating tweet. Moreover, the results of thematic lexicons and text summaries generated by information extraction models are also elaborated in detail.

Table 4.15: Earthquake results for all four variations of our trained models

Text-based features (baseline)

Category	Precision	Recall	F-score	Class Dist.
Eyewitness	0.878	0.977	0.925	1600
Non-eyewitness	0.893	0.585	0.707	200
Don't know	0.629	0.280	0.388	200

Domain-expert features

Eyewitness	0.871	0.969	0.917	1600
Non-eyewitness	0.787	0.645	0.709	200
Don't know	0.333	0.095	0.148	200

Domain-expert + text features

Eyewitness	0.865	0.987	0.922	1600
Non-eyewitness	0.912	0.620	0.738	200
Don't know	0.641	0.125	0.209	200

Domain-expert + text features with class balancing

Eyewitness	0.892	0.966	**0.927**	1600
Non-eyewitness	0.932	0.793	**0.857**	400 (+100%)
Don't know	0.801	0.653	**0.719**	400 (+100%)

4.3.1 *Crowdsourcing results*

The results of crowdsourcing experiment (described in section 3.4.2) where the crowd was asked to identify if the rest of the thread adds more information to the initiating post show that 70.5 percent of threads add more information to what is found in the initial tweet, while 29.5 percent do not. In addition, based on crowdworkers responses, the number of themes present in the initial tweet was compared with the number of themes present in the entire thread (Figure 4.2). The results show that the initial tweet has more instances than the remainder of the thread in the case of event reporting, location, and time theme. In comparison, in the case of intensity theme, the thread has a few more instances reported than the first tweet. Moreover, very few instances are present in the data for the themes of casualty and injury reports and help calls. In the threads, there is only one casualty and injury report and two help calls present

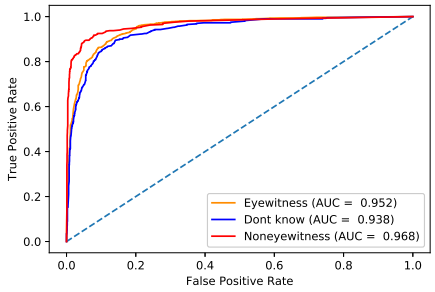
Table 4.16: Forest fire results for all four variations of our trained models

Category	Precision	Recall	F-score	Class Dist.
Text-based features (baseline)				
Eyewitness	0.649	0.265	0.376	189
Non-eyewitness	0.857	0.941	0.897	1379
Don't know	0.748	0.708	0.728	432
Domain-expert features				
Eyewitness	0.703	0.339	0.457	189
Non-eyewitness	0.863	0.943	0.901	1379
Don't know	0.737	0.688	0.711	432
Domain-expert + text features				
Eyewitness	0.794	0.265	0.397	189
Non-eyewitness	0.867	0.946	**0.905**	1379
Don't know	0.730	0.731	0.731	432
Domain-expert + text features with class balancing				
Eyewitness	0.897	0.714	**0.795**	378 (+100%)
Non-eyewitness	0.753	0.727	0.740	432
Don't know	0.876	0.935	**0.905**	1379

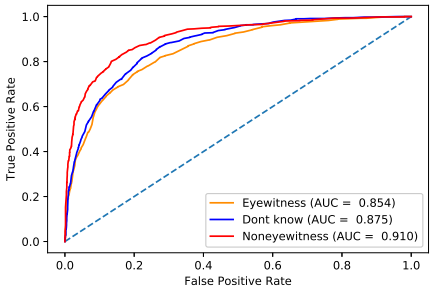
in the first tweet as well as in the rest of the thread. One potential reason for this limited number of records of casualty and damage reports and help calls is that none of the earthquake incidents in the data has caused mass disruption or casualties . Therefore, comparatively large volumes of event reporting, location, time, and intensity themes can be found.

4.3.2 *Thematic lexicons*

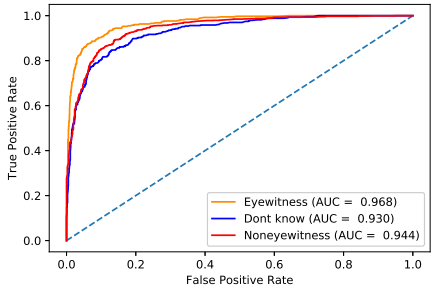
The characteristics of thematic lexicons are summarized in Table 4.17. For the event reporting theme four seed words generally describing the event of an earthquake were selected. The seed words retrieved 61 initial synset definitions. After the relevance check, 21 synset definitions were selected that retrieved 61 synsets. This results in a total number of 65 terms in the event reporting thematic lexicon. For time (relative) theme, 11 seed words were chosen and retrieved 20



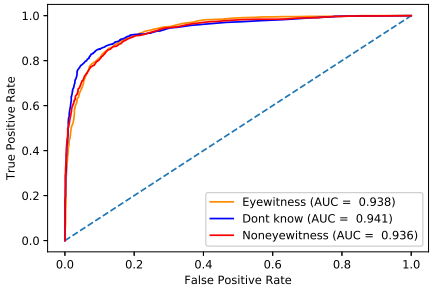
(I) Earthquake



(II) Flood



(III) Forest fire



(IV) Hurricane

Figure 4.1: ROC curves of all three classes of the best model

initial synset definitions. After the relevance check, only seven synset definitions were used to extract synsets. A total of 18 synsets were retrieved and therefore, the time (relative) theme lexicon consists of 29 words.

Table 4.17: Characteristics of thematic lexicons

Theme	**Seed words**	**Initial synset definitions**	**Selected synset definitions**	**Retrieved synsets**	**Total number of terms**
Event reporting	4	61	21	61	65
Time (relative)	11	20	7	18	29
Intensity	56	371	105	393	449
Casualty and damage reports	9	227	80	225	234
Help calls	4	81	21	101	105

The intensity theme has the highest number of seed words i.e. 56 to describe various intensity scales of an earthquake event. These seed words retrieved 371 initial synset definitions. After the relevance check, 105 synset definitions were selected and 393 synsets were retrieved. The intensity theme lexicon contains

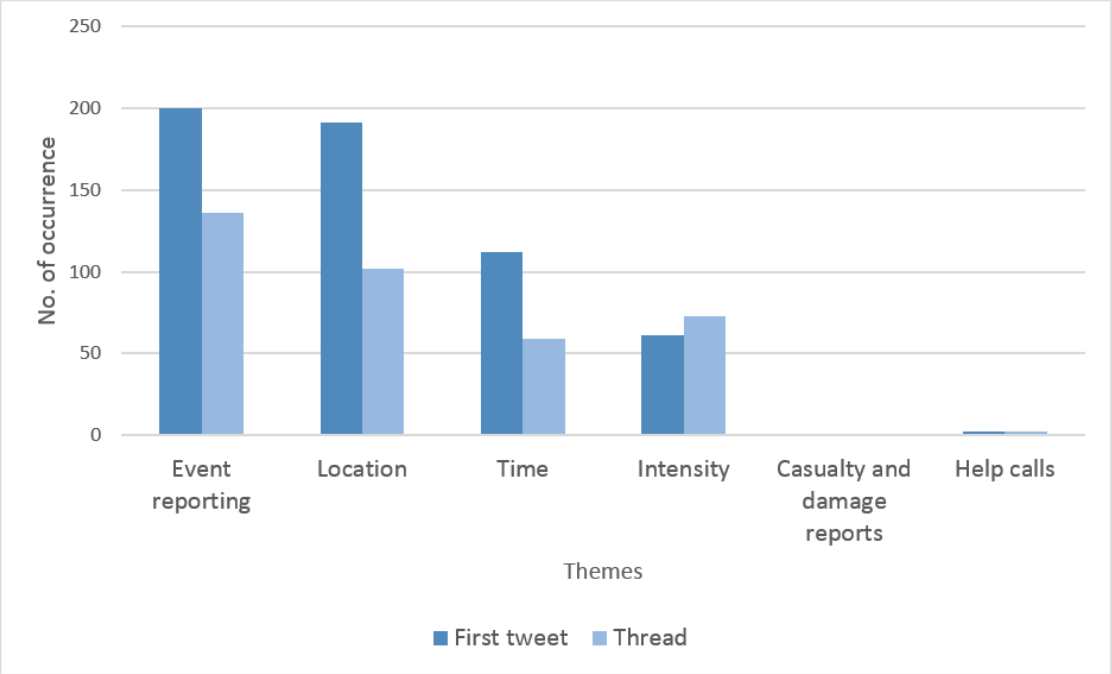


Figure 4.2: Comparison of the number of themes found in the first tweet and the thread

449 words in total. For the casualty and damage reports theme, there were nine seed words to retrieve 227 initial synset definitions, out of which 80 relevant synset definitions were selected after performing the relevance check. A total of 225 synsets were retrieved that result in 234 words in the casualty and damage reports lexicon. Finally, for the help call theme, four seed words were selected that retrieved 81 initial synset definitions. After the relevance check, only 21 relevant synset definitions were selected and 101 synsets were retrieved. This resulted in a total number of 105 words in the help calls thematic lexicon.

The word cloud of all thematic lexicons with the words found in all Twitter threads is shown in Figure 4.3. *Felt* is the most frequently used word for the event reporting theme (Figure 4.3a) with 427 occurrences, followed by *earthquake* with 377 occurrences. Whereas with 264 occurrences, *just* is the most frequently used term for time (relative) theme (Figure 4.3b). This outcome coincides with previous work [*Zahra et al.*, 2020], where the results suggested that the presence of the term *just* is a strong indication of a personal earthquake event observation. In addition, for earthquake events, social media users tend to report their observations immediately in comparison with other natural disasters, so the use of such immediate temporal markers is very frequent. Whereas, *Good* is the most frequently used term for the intensity theme (Figure 4.3c) with 38 occurrences followed by different instances of some obvious terms such as *big*, *strong*, *small*, etc. Exploring individual tweets reveals that to report an earthquake event, users often use phrases such as *felt a good jolt*.

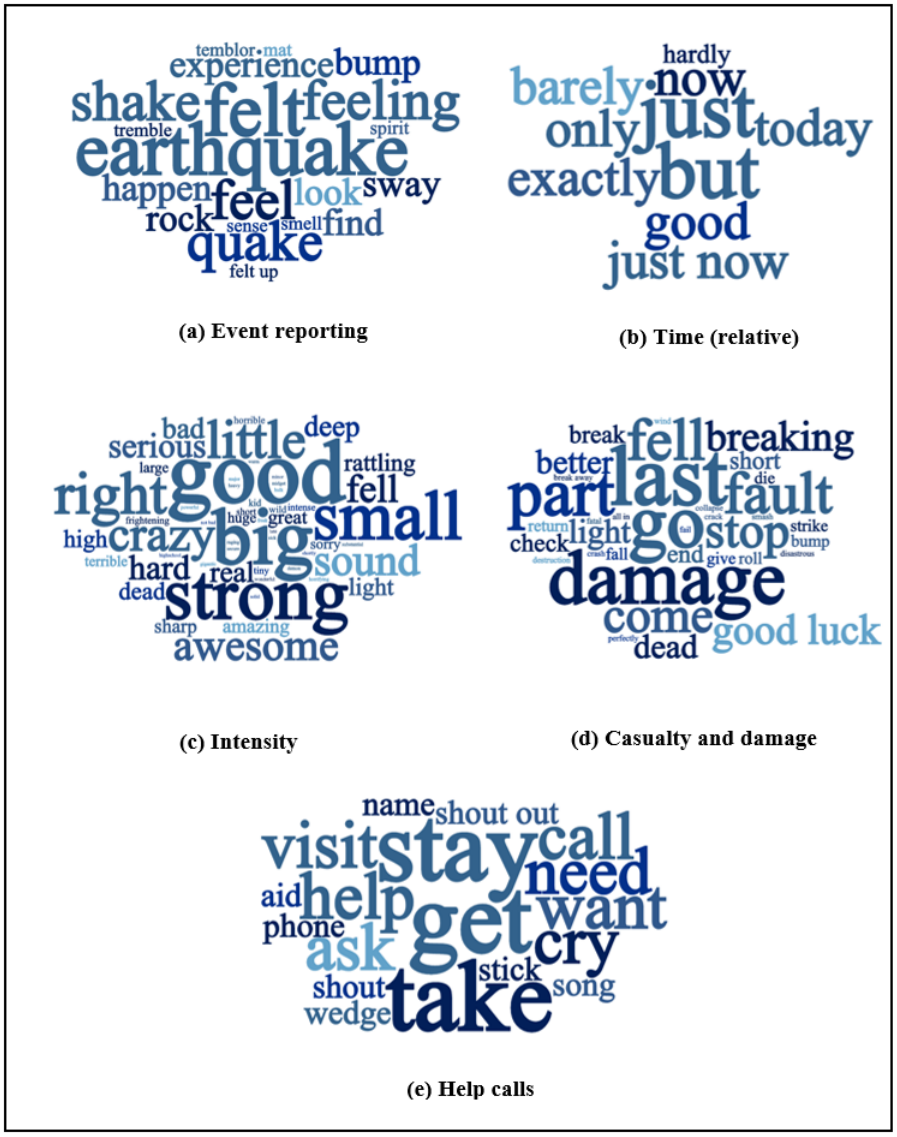


Figure 4.3: Lexicon words that occur in Twitter threads – the size of the word corresponds to the number of times it occurred in the data

Irrelevant terms such as *last* (17 times) and *go* (15 times) were the most frequent (due to their high frequency in language) for the relatively rare casualty and damage theme (Figure 4.3d). However, in the dataset, terms such as *harm* (11 times), *death* (5 times), and *fall* (3 times) were also found. Terms like *stay* (39 times), *get* (26 times), and *take* (23 times) occurred most frequently for the help calls theme (Figure 4.3e), followed by some obvious terms such as *need* (5 times), and *help* (4 times). Although actual help calls occurred only once (Figure 4.2) the word *help* occurred a few more times in another context, such as *Stay safe and keep in touch. Let us know if you need help with anything*. The complete lexicons developed in this research are available on GitHub[1].

1 https://github.com/rddspatial/text-summarization

4.3.3 *Summary evaluation*

A random selection of 50 summaries created by each information extraction model was done to evaluate which model performs best in preserving maximum information in the thread summary. The summaries were evaluated based on the presence of words from disaster-related themes. The analysis was based on context and semantics, i.e. simple presence of a term does not earn a score. On the contrary, a score is obtained by the meaningful presence of a word belonging to one of the six themes. To calculate inter-rater agreement, two people performed the evaluation. The best summary was later ranked on a scale of one to four, where one means the highest rank and four means the lowest rank. The findings show that on average the highest rank was 1.6 for TBoW model, followed by Word2Vec 1.7. The GloVe model has an average rank of 2.22, and the lowest average rank is 2.58 for the BOW model.

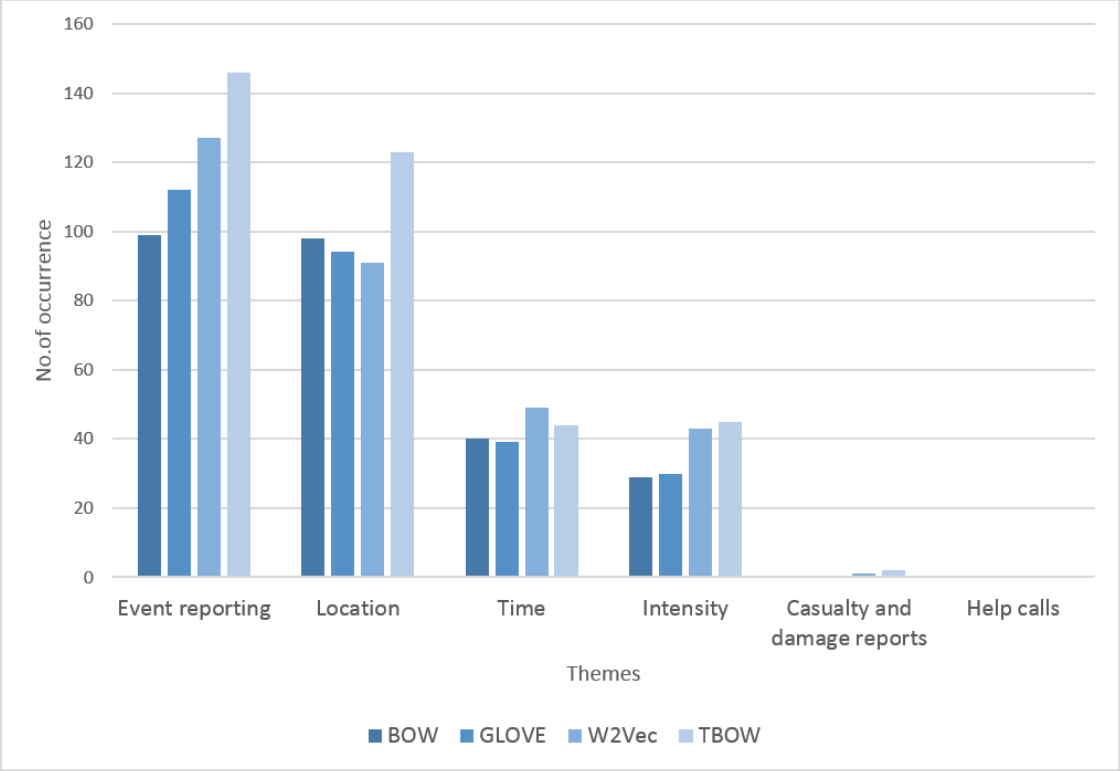


Figure 4.4: Individual themes present in four summaries

The number of themes present in all the summaries is shown in Figure 4.4. The comparison of all models shows that TBoW maintains the highest number of themes for event reporting, location, intensity, and casualty and damage reports. The word2vec model for the time theme outperforms the others.

Table 4.18 shows an example of text summaries created for one example thread by four information extraction models. The BoW summary has the smallest ratings, i.e. five, meaning five instances of different disaster-related themes are

Table 4.18: Example of an extractive text summary

Model	Summary	Total score
BoW	Just felt an Earthquake here in antaMonica Didn't feel a thing in Simi - not that far away. That sounded far less corny in my head. It's just the earth celebrating the new year a few days late.	5
GloVe	Didn't feel a thing in Santa Monica. Didn't feel a thing in Simi - not that far away. But, i don't have wine. You're in California ... you have lots of earthquakes. I felt it in hollywood. Going to find flashlights and shoes just in case... Hope all is well.	6
Word2Vec	Didn't feel a thing in Simi - not that far away. That sounded far less corny in my head. It's just the earth celebrating the new year a few days late. Hope all is well. Please keep us updated. i got nothin. didn't feel a thing in Santa Monica.	4
TBoW	Just felt an Earthquake here in SantaMonica.Didn't feel a thing in Simi - not that far away..no kidding! i got nothin..I felt it in hollywood..It's just the earth celebrating the new year a few days late. That sounded far less corny in my head..You're in California ... you have lots of earthquakes..didn't feel a thing in Santa Monica.	9

included in the summary. With a total score of six, the second summary produced by the model trained on GloVe preserves more relatively more information. With a total score of four, the third summary created by the model trained on Word2Vec word embeddings retains (in this particular case) relatively less disaster-related themes. Finally, with a total score of nine, the summary created by the TBoW model using disaster-related thematic lexicons retains the highest number of thematic instances. The full summaries generated by all four information extraction models are available on GitHub[2].

The results presented in this section have important insights for social media research community as well as for the future of information extraction models from social media data. As described in section 2.4.4, most of the current research relies on single posts for information extraction that is usually missing the essential context. The crowdsourcing results show that a thread can be

2 https://github.com/rddspatial/text-summarization

a valuable source of information with various disaster-related themes to add essential context and additional information to the initiating post. These results have also very important implications for emergency responders who can avoid information overload by extracting only valuable information from these threads for various disaster-related themes in the form of a summary. Although, at the moment, Twitter does not support downloading of a thread via its API, however, in future ever evolving social media platforms can adapt to the needs of research community.

5

DISCUSSION

The information extraction from short, unstructured, and informal social media data is a challenging task. This research aimed to demonstrate that social media posts contain useful information for emergency responders despite various challenges associated with the processing and filtering of this data. I developed various information extraction models focusing information needs of emergency responders during or after a disaster using social media data.

In particular, I addressed the following three challenges associated with the extraction of information from social media posts in the context of natural disasters:

- Timely extraction of relevant information from noise.
- Extraction of credible eyewitness reports from non-eyewitness reports.
- Extraction of information from social media threads for various disaster-related themes to add context and additional information to the initiating post.

To overcome these challenges, I used supervised machine-learning algorithms for text analysis and developed information extraction models to filter and classify relevant information automatically. To demonstrate the practicality of developed methods, I used Twitter data collected for more than 36 months from Twitter streaming API based on disaster-related keywords query. The tweet corpus contains more than 90 Million tweets that covers many natural disasters from all over the world. In the following, sections 5.1 to 5.4 address the research questions stated in chapter 1.

5.1 ROLE OF GEOGRAPHIC FEATURES IN INFORMATION EXTRACTION (RQ1)

Social media data is produced in bulk quantity at a high velocity [*Rajadesingan and Liu*, 2014] particularly during an emergency event. Considering the huge volume of the data, machine-learning algorithms (such as random forest, Naive Bayes) have been used to automatically classify relevant and not relevant information [*Batool et al.*, 2013] ; [*Purwarianti et al.*, 2016]. However, to generate high quality results, well-prepared training datasets are required to train these algorithms. Verma *et al.* 2011 state the classification of new events using the classifiers trained on previous events is a challenging task; however, the results are better for similar disasters. Based on these claims, I hypothesized that

geographic features play an important role in determining relevant information using classifiers trained on previous events of the same kind. The results of the experiments performed on the tweets posted during Italy and Myanmar earthquake (See section 4.1) support this hypothesis.

To test this hypothesis, I designed an experiment with three different cases and six different scenarios. The results of case I (scenario I) show (table 4.1 that Naive Bayes has well-learnt the classification task with training and test data from the same event. When these results are compared with the results of case I (scenario II) in table 4.2, it can be observed that the precision of *information* class has declined by 10 percent with almost a stable value of recall. However, it is important to note that in this scenario a new training data was not prepared for the new event (i.e. Myanmar earthquake), whereas, an already prepared training data (from Italy earthquake) was used to train the classifier by only swapping geographic features. The decline of 10 percent precision in detecting *information* class has saved the valuable time and resources spent on preparing new training data.

The case II with both scenarios was designed to observe the consistency of the performance of the classifier. The results (table 4.3 of Case II (scenario I) are consistent with previous results with a high precision and recall with *information* class. Furthermore, this time for scenario II results (table 4.4), precision declines by only 5 percent whereas the recall declines by almost 14 percent which means that this time less total relevant results were correctly classified. However, with the actual recall value of above 79 percent (table 4.4), the performance of the classifier is considered very good.

Both scenarios of case III showed consistent and very good results when test data was used from both events with a ratio of 50 percent.

The results of this experiment have important insights for the research community to further analyse the role of geographic features and identify other linguistic features when using machine learning algorithms for text analysis in extracting the relevant information.

5.2 A TAXONOMY OF CREDIBLE INFORMATION (RQ2)

Eyewitness reports are critical to analysing the credibility of the posted content on social media [*Truelove et al.*, 2015]. Social media users often share personal observations and first-hand reports during various incidents. The manual analysis of eyewitness reports (see section 4.2.1) revealed that a generalized eyewitness reports taxonomy for natural and human-induced disasters developed by Truelove *et al.* 2015 is not sufficient to cover the content shared during natural disasters because of distinct properties. The man-made disasters (such as protests, blasts and other terrorist activities) are not predictable. Therefore, no warnings are issued for such types of disasters.

Therefore, I designed an eyewitness reports taxonomy exclusive to natural disasters. The taxonomy includes direct eyewitness, vulnerable direct eyewitness, and indirect eyewitness reports besides non-eyewitness and don't know classes. The direct eyewitnesses are first-hand observations of the event and/or its impact, whereas the vulnerable direct eyewitness are reports shared by people who are present in a region for which a disaster warning has been issued. An indirect eyewitness report is from the friends and relatives of direct eyewitnesses who have shared first-hand reports with them. In the case of a natural disaster, in addition to a direct eyewitness report, indirect and vulnerable direct eyewitness reports can potentially be a very important piece of information for emergency responders. The vulnerable eyewitness reports can help them with an estimation of people present in the region when the disaster hits the location. Similarly, indirect eyewitness reports can help them to get more information about missing people from their friends and family.

The results of manual analysis (table 4.7) for eyewitness reports reveal that eyewitness, non-eyewitness, and don't know classes are greatly imbalanced in the dataset. For example, in the case of floods, almost 87 percent of data is comprised of don't know class. Similarly for earthquakes and hurricanes the percentage of don't know class is 65 and 80 respectively. For crowdsourcing experiment, the results (table 4.12) show a comparatively better situation for don't know class with 41, 10, 16, and 21 percent for floods, earthquakes, hurricanes, and forest fires respectively. This phenomenon can be interpreted as inherent errors of the random sampling technique. Another possible interpretation can be the time span of collected data which is almost one month for manual analysis data and almost two years for crowdsourced experiment data. The longer the time span of the data might mean higher probability of extracting relevant information.

For the sub classes of eyewitness reports, e.g. direct eyewitness, indirect eyewitness, and vulnerable direct eyewitness, the results (table 4.8) show a very small fraction of indirect eyewitness instances i.e. 1, 3, and 5 percent for floods, earthquakes, and hurricanes respectively with respect to total eyewitness instances. This percentage is very low when compared to direct and vulnerable eyewitness reports. Therefore, to prepare more labelled data, a crowdsourcing experiment 3.3.2 was performed on the second data sample. However, the results in table 4.12 reveal that the number of instances in all eyewitness classes are still not enough for a machine learning algorithm to learn and predict these classes with a decent accuracy. Moreover, all three classes overlap certain characteristics (tables 3.4, 3.5, 3.6) with a very few distinct characteristics that makes it even more difficult for the classifier to predict the correct class.

5.3 ROLE OF LINGUISTIC FEATURES IN EXTRACTING CREDIBLE INFORMATION (RQ3)

To be a witness of an event, the reporter should be present in the region to be able to observe or experience the event and its impact personally. A naive approach to infer an eyewitness report is to use geographic information of the users at the time of posting a tweet. There are three types geographic information that can found in and/or with a tweet. The first geographic information is found in metadata of a tweet in the form of a geotag. A geotag is the *live* location of user which can be attached to a tweet in the form of a pair of coordinates. However, a very small fraction of tweets contain a geotag with different spatial granularity [*Middleton et al.*, 2014]. Moreover, as per the announcement[1] from Twitter, the conventional precise geotagging function has been removed since June 2019. That is another motivation for the researchers that their research designs should not fundamentally rely on such dynamic platforms [*Hu and Wang*, 2020].

The second form of geographic information is the *location* field which users fill at the time of creating their account. This is a free-text format field where users can (virtually) add anything. Hecht *et al.* 2011 analysis reveal that 34 percent of social media users do not provide real location information, rather users tend to frequently add fake locations (such as Mars, Moon) or sarcastic comments (such as Justin Bieber's heart). Moreover, this is a static field that does not automatically change when users move to different locations.

The third form of geographic information is the geographic features shared in a tweet content. Tweeters either explicitly share geographic locations in their tweets (such as *Earthquake in Los Angeles California 4.6 at 11:38pm today 18-11-2020.*) or in the form of acronyms, abbreviations or hashtags (such as *That was first time I really felt an earthquake in LA*). In case of the first example, geocoding geographic location mentioned in the tweet content does not guarantee that the user who shared the tweet is also present in the region. And in case of the second example, advanced methods are required to geoparse such geographic locations.

Another important prerequisite to follow before using the location information of the users is to adhere to the ethical concerns to protect their privacy. The general data protection regulation[2] provides a detailed guideline on using location information and personal data of social media users. Therefore, in this research, I used a text-based approach over a location-based approach to identify eyewitness reports. I hypothesised that eyewitness reports contain various linguistic features that can play an important role in identifying them from non-eyewitness reports without using geographic information. For this purpose, I

1 https://twitter.com/TwitterSupport/status/1141039841993355264
2 https://www.into.ie/app/uploads/2019/10/GDPR_FAQ.pdf

identified various eyewitness characteristics (called domain-expert features) during manual analysis (see tables 3.4, 3.5, and 3.6). Some of the characteristics such as characteristic 11 in table 3.4 were only observed in a particular type of disaster i.e. in this case earthquake. The unpredictable and sudden nature of the event creates a state of panic where users write a very short text *earthquake!!* to express a wide range of information such as, personal experience, time i.e. the tweet was either posted in the middle of the earthquake or right after the event. Similarly the characteristic 12 in table 3.4 *caution and advice for others* was found only in floods, hurricanes, and forest fire because of the relatively predictable nature of the disaster. While developing the eyewitness classifiers all theses characteristics were combined to develop a robust model that is capable of classifying all of the disasters.

The eyewitness classification models were developed using a set of text-based features and domain expert features with a class balancing technique (SMOTE). The results in tables 4.13, 4.14, 4.15, and 4.16 reveal that on average a combination of domain expert features with text-based features coupled with SMOTE improved the performance of eyewitness classifiers when compared with the results of models that were only trained on text-based features.

5.4 INFORMATION EXTRACTION FROM SOCIAL MEDIA THREADS (RQ4, RQ5)

The limit on the number of characters on microblogs restrict users to write short and to the point text that often leads to missing the context of shared information. This causes a huge amount of data to be discarded that can potentially be useful information. To overcome this challenge, I hypothesised that social media threads can add context and additional information to the initiating post considering the common properties of a conventional conversation [*Ten Have*, 1990].

To perform the experiment, 200 Twitter threads posted during natural disasters by eyewitnesses were analysed using a crowdsourcing platform to determine the information content in the threads with respect to the initiating post. To achieve this goal, a job was designed to ask the crowd if a thread adds more information to the initiating tweet. The term "more information" is subjective, therefore, I defined six disaster-related themes to categorize the information efficiently. Each thread received at least three trusted judgements to calculate the inter-rated agreement. The results (section 4.3.1 of the experiment revealed that more than 70 percent threads add more information to the initiating tweet for various disaster-related themes.

The most commonly reported theme in the threads was event reporting followed by location. The event reports found in the thread are an important theme as assurance of an event from another person adds to the credibility of shared information. Similarly, a superficial analysis of the location theme present in

the threads revealed that it contains more detailed information (more locations where the event happened with finer granularity such as city or area level) that can potentially help to disambiguate the location shared in initiating posts.

When analysing a text corpus, it is important to understand the context and semantics of each word to extract the required information. For this purpose, word-embedding algorithms such as GloVe [*Pennington et al.*, 2014] and Word2Vec [*Mikolov et al.*, 2013] have been used widely. These algorithms are trained on voluminous datasets to determine the vector space of each word in a given sentence. Chowdhury *et al.* 2019 state that vocabulary used reporting various aspects of natural disasters is limited. Therefore, I used a lexicon-based approach to extract relevant information for various disaster-related themes from Twitter threads. The results of lexicon words found in Twitter threads are shown in figure 4.3. For themes like *event reporting*, *time (relative)*, and *intensity*, most of the lexicon terms are very relevant. On the other hand, for *casualty and damage* and *help calls* theme more irrelevant terms can be observed. One possible interpretation can be the lack of such reports in the threads as shown in figure 4.2.

To demonstrate the usability of a lexicon-based approach to extract relevant information, four information extraction models were developed (section 3.4). Considering the time pressure on emergency responders during an ongoing operation to analyse the information, I used extractive text summarization to develop a summary of the relevant information from the threads while excluding rest of the content. The results (section 4.3.3) revealed that the text summary generated by using the first information extraction model BoW preserved the amount of information related to different disaster-related themes with a 2.58 average rank. The second information extraction model generated word embeddings using GloVe that was trained on a general text corpus from Wikipedia 2014 articles and Gigaword 5 dataset to determine the context of each word. The text summaries generated by this model scored an average rank of 2.22. The third information extraction model word2vec was trained on 52 million tweets posted during various disasters [*Imran et al.*, 2016]. It showed even better results by scoring an average rank of 1.7. Lastly, the information extraction model developed using a lexicon-based approach (TBoW) on average scored a rank of 1.6 by preserving the highest amount of disaster-related information. These results coincide with the expectations that models trained on general text (i.e. GloVe) are less efficient in determining the context of the words as compared to the model (i.e. word2vec) trained on a similar kind of text (i.e. tweets posted during disasters). Moreover, the highest score of the model using a lexicon-based approach shows that reasonably high quality results can be produced without developing word embeddings trained on huge datasets.

5.5 OVERALL LIMITATIONS

This research has several limitations related to data and methodological approaches. As a data source for all the experiments, I used only the tweets posted in English language during different disasters. Language plays an important role on Twitter with more than half of the tweets posted in other than English language [*Hong et al.*, 2011]. This can possibly mean that I missed a huge amount of data particularly for disasters that occurred in regions where English is not the native language.

A very important limitation of using tweets for research is that the Twitter streaming API allows the downloading of an *unknown* percentage of the actual tweets posted at a given time. Therefore, researchers often call it "working within a black box" [*Driscoll and Walker*, 2014]. A comparison between the data downloaded by Twitter streaming API and Gnip firehose (a paid commercial service that ensures that every tweet posted meeting the query keywords is downloaded) reveals that the difference of downloaded data depends on the actual Twitter activity [*Driscoll and Walker*, 2014]. The high cost associated with paid commercial services like Gnip is however a big impediment for researchers to use these services.

Another limitation of using social media data for the purpose of information gathering is the role of social bots and so-called social media influencers who use this platform for deliberately spreading misinformation and rumours. Davis *et al.* 2016 developed an online service that analyses a Twitter account using over 1,000 features to determine if that particular account is real or a social bot. However, they have not evaluated the accuracy of their system. Moreover, deliberate efforts of spreading misinformation by social media influencers and other users are also a big drawback.

6

CONCLUSION AND OUTLOOK

In this work, I explored optimal solutions of various challenges associated with the processing of unstructured, informal, and short social media posts as a source of useful and credible information for practitioners and emergency responders during natural disasters. I hypothesised that information extraction models trained on machine learning algorithms can automatically extract useful and credible information from social media posts timely. By using a text corpus of millions of tweets posted during many natural disasters, I demonstrated that such short and informal posts could carry useful and critical information about the emergency events.

In an initial study, conducted on tweets posted during two earthquakes in two different geographic locations, the role of toponyms was analysed in determining the information class. The idea was to use a machine-learning algorithm (trained on an old disaster) on a new disaster that occurred in a different geographic location. This experiment aimed to present a solution that can potentially save the time that is required to prepare a new training dataset for every disaster. The experiment showed promising and consistent results in predicting *information* class by only swapping the geographic locations of the new disaster with the old one while keeping the rest of the content unchanged.

In the next study, I analysed the content of *information* tweets for credibility. This analysis was motivated by one of the measures to determine the credibility of social media posts i.e. eyewitness reports. Therefore, a tweet corpus posted during various natural disaster events was analysed to identify different types of eyewitness reports. These eyewitness reports were manually examined for eyewitness characteristics that were later used to train a supervised machine-learning algorithm to automatically categorize eyewitness and non-eyewitness reports. When eyewitness characteristics (identified during manual analysis) were coupled with text-based features on relatively balanced dataset, the performance of classification models was substantially increased.

Finally, the analysis of *credible information tweets* revealed that tweets are often too short and therefore the context and semantics of shared content is often missing. This challenge motivated the analysis of Twitter threads as an additional source of information to the initiating post. The results reveal that threads are a useful source of extracting information about various disaster-related information themes. To synthesize the threads, extractive text summarization technique was used. The information extraction model using disaster-related them-

atic lexicons proved to preserve highest content of information when compared to other models that were trained on word embeddings such as GloVe and word2vec.

Despite all the challenges associated with the processing of social media posts, this data has an established reputation as a unique source to collect up to date and useful information about an event. I suggest that future work analyses the ways to integrate other regional and local official data sources such as United States Geological survey that collects earthquake felt reports from all around the world and GeoNet platforms that collect reports about seismic activity and other disasters in New Zealand. Disaster reports collected from these sources can be used to assess the credibility of an event reported on social media in first place. Moreover, to assess the credibility of individual help calls exploring the role of multimodal social media data can provide important insights such as image data shared in a report can be used to assess the credibility of the text that contains a help call. Furthermore, to address the scarcity of location data associated with a post that contains actionable information (i.e. an evacuation request) the possibility of extracting additional location information from social media threads can be analysed.

REFERENCES

Abbasi, M. A., and H. Liu, Measuring user credibility in social media, in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction. Springer*, vol. 7812 LNCS, pp. 441–448, Berlin, doi: 10.1007/978-3-642-37210-0_48, 2013.

Abel, F., I. Celik, G.-J. Houben, and P. Siehndel, Leveraging the semantics of tweets for adaptive faceted search on twitter, in *International Semantic Web Conference*, pp. 1–17, Springer, 2011.

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

Aggarwal, A., A. Rajadesingan, and P. Kumaraguru, Phishari: Automatic real-time phishing detection on twitter, in *2012 eCrime Researchers Summit*, pp. 1–12, IEEE, 2012.

Alabi, O. F., A survey of facebook addiction level among selected nigerian university undergraduates, *New Media and Mass Communication*, *10*(2012), 70–80, 2013.

Alam, F., F. Ofli, M. Imran, and M. Aupetit, A twitter tale of three hurricanes: Harvey, irma, and maria, in *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2018.

Alexander, D., *Natural disasters*, Routledge, 1993.

Alrubaian, M., M. Al-Qurishi, M. M. Hassan, and A. Alamri, A Credibility Analysis System for Assessing Information on Twitter, *IEEE Transactions on Dependable and Secure Computing*, *15*(4), 661–674, doi: 10.1109/TDSC.2016.2602338, 2018.

Ashktorab, Z., C. Brown, M. Nandi, and A. Culotta, Tweedr: Mining twitter to inform disaster response., in *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pp. 269–272, 2014.

Asur, S., B. A. Huberman, G. Szabo, and C. Wang, Trends in social media: Persistence and decay, in *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

Aubrecht, C., S. Fuchs, and C. Neuhold, Spatio-temporal aspects and dimensions in integrated disaster risk management, *Natural Hazards*, *68*(3), 1205–1216, 2013.

Batool, R., A. M. Khattak, J. Maqbool, and S. Lee, Precise tweet classification and sentiment analysis, in *Proceedings of the 12th International Conference on Computer and Information Science (ICIS) IEEE/ACIS*, pp. 461–466, IEEE, 2013.

Bifet, A., and E. Frank, Sentiment knowledge discovery in twitter streaming data, in *International conference on discovery science*, pp. 1–15, Springer, 2010.

Botzen, W., and J. Van Den Bergh, Managing natural disaster risks in a changing climate, *Environmental Hazards*, *8*(3), 209–225, 2009.

Boyd, D., S. Golder, and G. Lotan, Tweet, tweet, retweet: Conversational aspects of retweeting on twitter, in *2010 43rd Hawaii International Conference on System Sciences*, pp. 1–10, IEEE, 2010.

Bruns, A., and J. Burgess, Researching news discussion on twitter: New methodologies, *Journalism Studies*, *13*(5-6), 801–814, 2012.

Bryant, E. A., L. M. Head, and J. Morrison, Planning for natural hazards—how can we mitigate the impacts?, in *Proceedings of a symposium with the same title*, 2005.

Bullock, J. A., G. D. Haddow, and D. P. Coppola, *Introduction to emergency management*, Butterworth-Heinemann, 2017.

Buntain, C., and J. Golbeck, Automatically Identifying Fake News in Popular Twitter Threads, in *017 IEEE International Conference on Smart Cloud (SmartCloud)*, pp. 208–215, 2017.

Burton, I., *The environment as hazard*, Guilford press, 1993.

Castillo, C., M. Mendoza, and B. Poblete, Information credibility on twitter, in *Proceedings of the 20th international conference on World wide web*, pp. 675–684, 2011.

Castillo, C., M. Mendoza, and B. Poblete, Predicting information credibility in time-sensitive social media, *Internet Research*, 2013.

Cerrai, D., Q. Yang, X. Shen, M. Koukoula, and E. N. Anagnostou, Brief communication: Hurricane dorian: automated near-real-time mapping of the "unprecedented" flooding in the bahamas using synthetic aperture radar, *Natural Hazards and Earth System Sciences*, *20*(5), 1463–1468, 2020.

Chang, H.-C., A new perspective on twitter hashtag use: Diffusion of innovation theory, *Proceedings of the American Society for Information Science and Technology*, *47*(1), 1–4, 2010.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, *Journal of Artificial Intelligence Research 16*, *16*, 321–357, doi: 10.1002/eap.2043, 2002.

Chen, Y., and S. Skiena, Building sentiment lexicons for all major languages, in *52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference*, vol. 2, pp. 383–389, Association for Computational Linguistics, doi: 10.3115/v1/p14-2063, 2014.

Chowdhury, J. R., C. Caragea, and D. Caragea, Keyphrase extraction from disaster-related tweets, in *Proceedings of the 28th International Conference on World Wide Web (WWW)*, pp. 1555–1566, 2019.

Cobo, A., D. Parra, and J. Navón, Identifying relevant messages in a twitter-based citizen channel for natural disaster situations, in *Proceedings of the 24th International Conference on World Wide Web*, pp. 1189–1194, 2015.

Cooper, C., and R. Block, *Disaster: Hurricane Katrina and the failure of homeland security*, Macmillan, 2007.

Crooks, A., A. Croitoru, A. Stefanidis, and J. Radzikowski, # earthquake: Twitter as a distributed sensor system, *Transactions in GIS*, *17*(1), 124–147, 2013.

Das, T. K., and P. M. Kumar, Big data analytics: A framework for unstructured data analysis, *International Journal of Engineering Science & Technology*, *5*(1), 153, 2013.

Davis, C. A., O. Varol, E. Ferrara, A. Flammini, and F. Menczer, Botornot: A system to evaluate social bots, in *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 273–274, 2016.

De Zeeuw, H., R. Van Veenhuizen, and M. Dubbeling, The role of urban agriculture in building resilient cities in developing countries, *The Journal of Agricultural Science*, *149*(S1), 153, 2011.

Del Vicario, M., A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, The spreading of misinformation online, *Proceedings of the National Academy of Sciences*, *113*(3), 554–559, 2016.

Doggett, E., and A. Cantarero, Identifying eyewitness news-worthy events on twitter, in *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pp. 7–13, 2016.

Driscoll, K., and S. Walker, Big data, big questions | working within a black box: Transparency in the collection and production of big twitter data, *International Journal of Communication*, *8*, 20, 2014.

Dugdale, J., B. Van de Walle, and C. Koeppinghoff, Social media and sms in the haiti earthquake, in *Proceedings of the 21st International Conference on World Wide Web (WWW)*, pp. 713–714, 2012.

Duni, L., and N. Theodoulidis, Short note on the november 26, 2019, durres (albania) m6. 4 earthquake: Strong ground motion with emphasis in durres city, *EMSC on Line Report. Available online: https://www. google. com. hk/url*, 2019.

Escobar, H., Amazon fires clearly linked to deforestation, scientists say, 2019.

Fang, R., A. Nourbakhsh, X. Liu, S. Shah, and Q. Li, Witness identification in twitter, in *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pp. 65–73, 2016.

Fellbaum, C., Wordnet, in *Theory and applications of ontology: computer applications*, pp. 231–243, Springer, 2010.

Ferreira, R., L. De Souza Cabral, R. D. Lins, G. Pereira E Silva, F. Freitas, G. D. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, Assessing sentence scoring techniques for extractive text summarization, *Expert Systems with Applications*, *40*(14), 5755–5764, doi: 10.1016/j.eswa.2013.04.023, 2013.

Floridi, L., *Information: A very short introduction*, OUP Oxford, 2010.

Ghosh, A., G. Li, T. Veale, P. Rosso, E. Shutova, J. Barnden, and A. Reyes, Sentiment analysis of figurative language in twitter, in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, pp. 470–478, 2015.

Goodchild, M. F., Citizens as sensors: the world of volunteered geography, *GeoJournal*, *69*(4), 211–221, 2007.

Goodchild, M. F., and J. A. Glennon, Crowdsourcing geographic information for disaster response: a research frontier, *International Journal of Digital Earth*, *3*(3), 231–241, doi: 10.1080/17538941003759255, 2010.

Gruebner, O., M. Sykora, S. R. Lowe, K. Shankardass, S. Galea, and S. Subramanian, Big data opportunities for social behavioral and mental health research, *Social Science & Medicine*, *189*, 167–169, 2017.

Gupta, A., and P. Kumaraguru, Credibility ranking of tweets during high impact events, *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, pp. 2–8, 2012.

Gupta, A., H. Lamba, P. Kumaraguru, and A. Joshi, Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy, in *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pp. 729–736, 2013.

Gupta, A., P. Kumaraguru, C. Castillo, and P. Meier, Tweetcred: Real-time credibility assessment of content on twitter, in *International Conference on Social Informatics.*, vol. 8851, pp. 228–243, Springer, Cham, doi: 10.1007/978-3-319-13734-6_16, 2014.

Gupta, K., Challenges in developing urban flood resilience in india, *Philosophical Transactions of the Royal Society A*, *378*(2168), 20190,211, 2020.

Hagras, M., G. Hassan, and N. Farag, Towards natural disasters detection from twitter using topic modelling, in *European Conference on Electrical Engineering and Computer Science (EECS)*, pp. 272–279, IEEE, 2017.

Hecht, B., L. Hong, B. Suh, and E. H. Chi, Tweets from justin bieber's heart: the dynamics of the location field in user profiles, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 237–246, 2011.

Hilligoss, B., and S. Y. Rieh, Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context, *Information Processing and Management*, *44*(4), 1467–1484, doi: 10.1016/j.ipm.2007.10.001, 2008.

Hong, L., G. Convertino, and E. H. Chi, Language matters in twitter: A large scale study, in *Fifth International AAAI Conference on Weblogs and Social Media*, Citeseer, 2011.

Howden, M., How humanitarian logistics information systems can improve humanitarian supply chains: a view from the field, in *Proceedings of the 6th international ISCRAM conference, Gothenburg, Sweden*, 2009.

Hu, Y., and R.-Q. Wang, Understanding the removal of precise geotagging in tweets, *Nature Human Behaviour*, pp. 1–3, 2020.

Imran, M., S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, Extracting information nuggets from disaster-related messages in social media., in *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2013.

Imran, M., C. Castillo, J. Lucas, P. Meier, and S. Vieweg, Aidr: Artificial intelligence for disaster response, in *Proceedings of the 23rd International Conference on World Wide Web (WWW)*, pp. 159–162, 2014.

Imran, M., P. Mitra, and C. Castillo, Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages, *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pp. 1638–1643, 2016.

Jain, A., D. Bhatia, and M. K. Thakur, Extractive Text Summarization Using Word Vector Embedding, *Proceedings of International Conference on Machine Learning and Data Science (MLDS)*, pp. 51–55, doi: 10.1109/MLDS.2017.12, 2017.

Jiang, J., Information extraction from text, in *Mining text data*, pp. 11–41, Springer, 2012.

Kang, B., J. O'Donovan, and T. Höllerer, Modeling topic specific credibility on twitter, in *Proceedings of the ACM international Conference on Intelligent User Interfaces - IUI '12*, pp. 179–188, Lisbon, Portugal, doi: 10.1145/2166966.2166998, 2012.

Karami, A., V. Shah, R. Vaezi, and A. Bansal, Twitter speaks: A case of national disaster situational awareness, *Journal of Information Science*, *46*(3), 313–324, 2020.

Khan, F. H., S. Bashir, and U. Qamar, Tom: Twitter opinion mining framework using hybrid classification scheme, *Decision Support Systems*, *57*, 245–257, 2014.

Khan, H., L. G. Vasilescu, A. Khan, et al., Disaster management cycle-a theoretical approach, *Journal of Management and Marketing*, *6*(1), 43–50, 2008.

Kumar, S., G. Barbier, M. A. Ali Abbasi, and H. Liu, TweetTracker: An Analysis Tool for Humanitarian and Disaster Relief, in *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 661–662, 2011.

Laituri, M., and K. Kodrich, On line disaster response community: People as sensors of high magnitude disasters using internet gis, *Sensors*, *8*(5), 3037–3055, 2008.

Ledeneva, Y., A. Gelbukh, and R. A. García-Hernández, Terms derived from frequent sequences for extractive text summarization, in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 593–604, Springer, 2008.

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

Li, J., and H. R. Rao, Twitter as a rapid response news service: An exploration in the context of the 2008 china earthquake, *The Electronic Journal of Information Systems in Developing Countries*, *42*(1), 1–22, 2010.

Liu, I. L., C. M. Cheung, and M. K. Lee, User satisfaction with microblogging: Information dissemination versus social networking, *Journal of the Association for Information Science and Technology*, *67*(1), 56–70, 2016.

Liu, L., Y. Lu, M. Yang, Q. Qu, J. Zhu, and H. Li, Generative adversarial network for abstractive text summarization, in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 8109–8110, 2018.

Loper, E., and S. Bird, NLTK: The Natural Language Toolkit, in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, doi: 10.3115/1118108.1118117, 2002.

Mendoza, M., B. Poblete, and C. Castillo, Twitter under crisis: Can we trust what we rt?, in *Proceedings of the First Workshop on Social Media Analytics*, pp. 71–79, 2010.

Middleton, S. E., L. Middleton, and S. Modafferi, Real-time crisis mapping of natural disasters using social media, *IEEE Intelligent Systems*, *29*(2), 9–17, doi: 10.1109/MIS.2013.126, 2014.

Mihalcea, R., and P. Rarau, TextRank: Bringing Order into Texts, in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, doi: 10.1016/0305-0491(73)90144-2, 2004.

Mikolov, T., K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, in *In Proceedings of ICLR Workshops Track*, 2013.

Moawad, I. F., and M. Aref, Semantic graph reduction approach for abstractive Text Summarization, in *Proceedings of International Conference on Computer Engineering and Systems*, pp. 132–138, IEEE, doi: 10.1109/ICCES.2012.6408498, 2012.

Musaev, A., and Q. Hou, Gathering high quality information on landslides from Twitter by relevance ranking of users and tweets, in *Proceeding of 2nd International Conference on Collaboration and Internet Computing*, pp. 276–284, doi: 10.1109/CIC.2016.43, 2016.

Naseem, U., and K. Musial, Dice: Deep intelligent contextual embedding for twitter sentiment analysis, in *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 953–958, IEEE, 2019.

Nichols, J., J. Mahmud, and C. Drews, Summarizing sporting events using Twitter, in *International Conference on Intelligent User Interfaces, Proceedings IUI*, pp. 189–198, doi: 10.1145/2166966.2166999, 2012.

Oh, O., K. H. Kwon, and H. R. Rao, An exploration of social media in extreme events: Rumor theory and twitter during the haiti earthquake 2010., in *Icis*, vol. 231, pp. 7332–7336, 2010.

Osborne, M., S. Petrovic, R. McCreadie, C. Macdonald, and I. Ounis, Bieber no more: First story detection using twitter and wikipedia, in *Sigir 2012 Workshop on Time-Aware Information Access*, pp. 16–76, Citeseer, 2012.

Pelling, M., Urbanization and disaster risk, *Cyberseminar on Population and Natural Hazards*, 2007.

Pennington, J., R. Socher, and C. D. Manning, GloVe: Global Vectors for Word Representation, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Petrović, S., M. Osborne, and V. Lavrenko, Streaming first story detection with application to twitter, in *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference (NAACL HLT)*, June, pp. 181–189, 2010.

Plisson, J., N. Lavrac, and D. D. Mladenić, A rule based approach to word lemmatization, in *Proceedings of the 7th International Multiconference Information Society (IS'04)*, pp. 83–86, 2004.

Poser, K., and D. Dransch, Volunteered geographic information for disaster management with application to rapid flood damage estimation, *Geomatica*, *64*(1), 89–98, 2010.

Purohit, H., A. Hampton, V. L. Shalin, A. P. Sheth, J. Flach, and S. Bhatt, What kind of #conversation is Twitter? Mining #psycholinguistic cues for emergency coordination, *Computers in Human Behavior*, *29*(6), 2438–2447, doi: 10.1016/j.chb.2013.05.007, 2013.

Purwarianti, A., A. Andhika, A. F. Wicaksono, I. Afif, and F. Ferdian, Inanlp: Indonesia natural language processing toolkit, case study: Complaint tweet classification, in *International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA)*, pp. 1–5, IEEE, 2016.

Rajadesingan, A., and H. Liu, Identifying users with opposing opinions in twitter debates, in *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pp. 153–160, Springer, 2014.

Reuter, C., M. A. Kaufhold, T. Spielhofer, and A. S. Hahne, Social media in emergencies: A representative study on citizens' perception in Germany, in *Proceedings of the ACM on Human-Computer Interaction*, p. 90, doi: 10.1145/3134725, 2017.

Romero, S., and K. Becker, A framework for event classification in tweets based on hybrid semantic enrichment, *Expert Systems with Applications*, *118*, 522–538, doi: 10.1016/j.eswa.2018.10.028, 2019.

Rowe, M., M. Stankovic, and A. Dadzie, Making Sense of Microposts (# MSM2012), *Ceur-Ws.Org*, 2012.

Sakaki, T., M. Okazaki, and Y. Matsuo, Earthquake shakes twitter users: real-time event detection by social sensors, in *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 851–860, 2010.

Sakaki, T., M. Okazaki, and Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 919–931, 2013.

Senaratne, H., A. Mobasheri, A. L. Ali, C. Capineri, and M. M. Haklay, A review of volunteered geographic information quality assessment methods, *International Journal of Geographical Information Science*, *8816*(June), 1–29, doi: 10.1080/13658816.2016.1189556, 2016.

Shaw, R., Post disaster recovery: Issues and challenges, in *Disaster Recovery*, pp. 1–13, Springer, 2014.

Spence, P. R., K. A. Lachlan, X. Lin, and M. del Greco, Variability in twitter content across the stages of a natural disaster: Implications for crisis communication, *Communication Quarterly*, *63*(2), 171–186, 2015.

Spinsanti, L., and F. Ostermann, Automated geographic context analysis for volunteered information, *Applied Geography*, *43*, 36–44, 2013.

Starbird, K., J. Maddock, M. Orand, P. Achterman, and R. M. Mason, Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing, in *IConference Proceedings*, iSchools, 2014.

Stowe, K., J. Anderson, M. Palmer, L. Palen, and K. M. Anderson, 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, 2018.

Takahashi, B., E. C. Tandoc, and C. Carmichael, Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines, *Computers in Human Behavior*, *50*, 392–398, doi: 10.1016/j.chb.2015.04.020, 2015.

Tao, K., F. Abel, C. Hauff, and G.-J. Houben, Twinder: A Search Engine for Twitter Streams, in *International Conference on Web Engineering. Springer, Berlin, Heidelberg*, 2012.

Tapia, A. H., K. Bajpai, B. J. Jansen, and J. Yen, Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations, in *8th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2011.

Téllez-Valero, A., M. Montes-y Gómez, and L. Villaseñor-Pineda, A machine learning approach to information extraction, in *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 539–547, Springer, 2005.

Ten Have, P., Methodological issues in conversation analysis1, *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *27*(1), 23–51, 1990.

Terpstra, T., R. Stronkman, A. de Vries, and G. L. Paradies, Towards a realtime twitter analysis during crises for operational crisis management, in *9th International Conference on InformationSystems for Crisis Response and Management (ISCRAM)*, 2012.

Thompson, S., N. Altay, W. G. Green III, and J. Lapetina, Improving disaster response efforts with decision support systems, *International Journal of Emergency Management*, *3*(4), 250–263, 2006.

Toriumi, F., T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda, Information sharing on twitter during the 2011 catastrophic earthquake, in *Proceedings of the 22nd International Conference on World Wide Web (WWW)*, pp. 1025–1028, 2013.

Truelove, M., M. Vasardani, and S. Winter, Testing a model of witness accounts in social media, in *Proceedings of the 8th Workshop on Geographic Information Retrieval - GIR '14*, pp. 1–8, doi: 10.1145/2675354.2675699, 2014.

Truelove, M., M. Vasardani, and S. Winter, Towards credibility of micro-blogs: characterising witness accounts, *GeoJournal*, *80*(3), 339–359, 2015.

Truong, B., C. Caragea, A. Squicciarini, and A. H. Tapia, Identifying valuable information from twitter during natural disasters, *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–4, 2014.

Tuten, T. L., *Advertising 2.0: social media marketing in a web 2.0 world*, Westport, CT: Praeger, 2008.

Verma, S., S. Vieweg, W. J. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson, Natural language processing to the rescue? extracting" situational awareness" tweets during mass emergency., in *ICWSM*, pp. 385–392, Citeseer, 2011.

Vicario, M. D., A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, E. Stanely, and W. Quattrociocchi, The spreading of misinformation online, *Proceedings of the National Academy of Sciences*, *113*(3), 554–559, doi: 10.1073/pnas.1517441113, 2016.

Vieweg, S., A. L. Hughes, K. Starbird, and L. Palen, Microblogging during two natural hazards events: what twitter may contribute to situational awareness, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1079–1088, 2010.

Wan, Y., and Q. Gao, An ensemble sentiment classification system of twitter data for airline services analysis, in *IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1318–1325, 2015.

Ward, M., et al., Impact of 2019–2020 mega-fires on australian fauna habitat, *Nature Ecology & Evolution*, pp. 1–6, 2020.

Wasike, B. S., Framing news in 140 characters: How social media editors frame the news and interact with audiences via twitter., *Global Media Journal: Canadian Edition*, *6*(1), 2013.

Wood, H. O., and F. Neumann, Modified Mercalli intensity scale of 1931, *Bulletin of the Seismological Society of America*, *21*(4), 277–283, 1931.

Wu, W., B. Zhang, and M. Ostendorf, Automatic generation of personalized annotation tags for twitter users, in *Human language technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 689–692, 2010.

Xu, B., X. Guo, Y. Ye, and J. Cheng, An improved random forest classifier for text categorization., *Journal of Computers*, *7*(12), 2913–2920, 2012.

Yin, D., Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, Detection of harassment on web 2.0, *Proceedings of the Content Analysis in the WEB*, *2*, 1–7, 2009.

Zahra, K., F. O. Ostermann, and R. S. Purves, Geographic variability of Twitter usage characteristics during disaster events, *Geo-Spatial Information Science*, *20*(3), 231–240, doi: 10.1080/10095020.2017.1371903, 2017.

Zahra, K., M. Imran, and F. O. Ostermann, Automatic identification of eyewitness messages on twitter during disasters, *Information Processing & Management*, *57*(1), 102,107, 2020.

Zeng, D., H. Chen, R. Lusch, and S.-H. Li, Social media analytics and intelligence, *IEEE Intelligent Systems*, *25*(6), 13–16, 2010.

Zhao, D., and M. B. Rosson, How and why people twitter: the role that microblogging plays in informal communication at work, in *Proceedings of the ACM International Conference on Supporting Group Work*, pp. 243–252, 2009.

Part II

PUBLICATIONS

A

PUBLICATION I: GEOGRAPHIC VARIABILITY OF TWITTER USAGE CHARACTERISTICS DURING DISASTER EVENTS

Zahra, K., Ostermann, F. O., & Purves, R. S. (2017). Geographic variability of Twitter usage characteristics during disaster events. *Geo-spatial information science*, 20(3), 231-240.

GEO-SPATIAL INFORMATION SCIENCE, 2017
VOL. 20, NO. 3, 231–240
https://doi.org/10.1080/10095020.2017.1371903

Taylor & Francis
Taylor & Francis Group

Geographic variability of Twitter usage characteristics during disaster events

Kiran Zahra[a], Frank O. Ostermann[b] and Ross S. Purves[a]

[a]Department of Geography, University of Zurich, Zurich, Switzerland; [b]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands

ABSTRACT

Twitter is a well-known microblogging platform for rapid diffusion of views, ideas, and information. During disasters, it has widely been used to communicate evacuation plans, distribute calls for help, and assist in damage assessment. The reliability of such information is very important for decision-making in a crisis situation, but also difficult to assess. There is little research so far on the transferability of quality assessment methods from one geographic region to another. The main contribution of this research is to study Twitter usage characteristics of users based in different geographic locations during disasters. We examine tweeting activity during two earthquakes in Italy and Myanmar. We compare the granularity of geographic references used, user profile characteristics that are related to credibility, and the performance of Naïve Bayes models for classifying Tweets when used on data from a different region than the one used to train the model. Our results show similar geographic granularity for Myanmar and Italy earthquake events, but the Myanmar earthquake event has less information from locations nearby when compared to Italy. Additionally, there are significant and complex differences in user and usage characteristics, but a high performance for the Naïve Bayes classifier even when applied to data from a different geographic region. This research provides a basis for further research in credibility assessment of users reporting about disasters

ARTICLE HISTORY
Received 23 June 2017
Accepted 12 August 2017

KEYWORDS
Geographic feature granularity; Volunteered Geographic Information (VGI); Naïve Bayes; Twitter; credibility; Geonames

1. Introduction

The growth of social media over the last decade, and its possible use as a source of information about a wide variety of topics including events, news, personal opinions, and many more (Hossmann et al. 2011); (Terpstra et al. 2012) is unquestionable. One widely studied investigated potential use is real-time monitoring of events (Middleton, Middleton, and Modafferi 2014). In particular, where events take the form of natural disasters additional information with respect to casualties, damage, situational updates, and evacuation plans has the potential to be extremely valuable (Verma et al. 2011).

However, not everything shared on social media can be considered as useful and actionable information with respect to natural disasters, since people also share spam, personal opinions, and material to explicitly harass other users (Senaratne et al. 2017). Even if we collect Tweets based on particular keywords related to a specific theme, the retrieved content may still not be relevant since many words and phrases are polysemous and may also be used as synonyms or metaphors (Sakaki, Okazaki, and Matsuo 2013). Thus, one may "tremble" in fear, "like an avalanche," and we may be "flooded" with information, and "fire" is used in many metaphors about emotions. This makes the adoption of methods which can analyze the semantics behind particular terms very important if we wish to categorize information harvested from social media as relevant or irrelevant pieces of information with respect to a particular class of events.

Twitter currently offers access to real-time data in the form of Tweets through its streaming Application Programming Interface (API). This API requires certain parameters to capture Tweets such as particular keywords, Tweets sent from particular users, or Tweets originating from a particular region. For our project, we wrote a script in R to capture Tweets based on disaster-related keywords such as earthquake, flood, hurricane, etc. During the data collection phase of our project, we observed a sudden rise in the number of Tweets contemporaneously with events such as earthquakes or storms. This observation forms the basis of many event detection applications which claim to detect events in near real time (e.g. Sakaki, Okazaki, and Matsuo 2010).

The normal daily count of Tweets containing our keywords is around 50,000, but it rises tenfold to maxima of around 500,000 Tweets in case of disasters. It appears that users connect to Twitter even to verify a small earthquake experienced by themselves (example Tweet text: "Was that #earthquake in Cali, or someone was rocking my chair?"), or to know about damages and

CONTACT Kiran Zahra kiran.zahra@geo.uzh.ch

casualties caused by a major earthquake. This behavior is well known, and multiple studies have used Twitter to detect events such as earthquakes and attempt to determine their geographical extent or magnitude (Sakaki, Okazaki, and Matsuo 2010), among other things. However, little attention appears to have been paid to issues relating either to the semantics of Tweets or the specific quality of information, as opposed to many more general studies on the quality of Twitter and Volunteered Geographic Information more generally. Especially, the potential geographic variability in the usage of Twitter remains a concern to be addressed, as it impacts on the potential transferability of methods to assess and evaluate Tweets.

In the case-study reported in this paper, which extends a workshop paper on the same topic (Zahra and Purves 2017), we selected two natural disasters which occurred on the same date in two different geographic regions of the world to explore the geographic variability of Tweets and its impact on information content, credibility-related characteristics, and trained models to classify Tweets. The first disaster was an earthquake which occurred in Italy on 24 August 2016 at 03:36 local time, and the second one was an earthquake in Myanmar on the same date at 17:04 local time. The two earthquakes were both of strong magnitudes (Italy 6.2 and Myanmar 6.8 on the Richter scale).

Since Tweets contain free text, Twitter users can report on disasters in many different ways. One critical feature in terms of information content that relates to fitness-for-purpose of Tweets is the granularity of the reported geographic location. We defined granularity with respect to a Tweet as referring to the specificity or precision of the area described in a Tweet – thus a Tweet reporting on an event in Italy is of coarse granularity, and of limited information use, while one reporting on an event near the commune of Accumoli in the Province of Rieti in Italy has a fine granularity and higher information value.

There are four types of location information associated with a Tweet: GPS coordinates formated as GeoJSON in the "coordinates" metadata field, a place indicated by the user in the "place" metadata field using Twitter's database of places, a location mentioned in the user profile's "location" free-form metadata field, and a location mentioned in the Tweet's content. We focus on the latter, because only 1–2% of all Tweets have GPS coordinates, the "place" metadata is often too coarse at the country level, the user profile location is often incorrect and static (Hecht et al. 2011), and we are interested in the location being tweeted about. We consider any Tweet containing locational information about the earthquake to be a potential source of information.

Geonames is an open source gazetteer that offers a standardized administrative hierarchy for different countries of the world, thereby assisting in the analysis of the granularity of place names (toponyms) used between different regions. Our first research question took advantage of this feature:

RQ1: How does the spatial granularity with which an event is reported in terms of toponym hierarchy according to Geonames vary in two different continents?

One important aspect of data quality is the credibility of a Tweet, that is to say how likely is it that the content is for example, accurate, authoritative, objective, and current (Gupta and Kumaraguru 2012). Since Tweets are user-generated data, produced for many different reasons, they are also associated with varying quality with respect to particular contexts (Senaratne et al. 2017). We assume that the usage characteristics of contributors can help to assess the credibility of a particular Tweet. In our second research question, we therefore explore the different user-based features of Tweets.

RQ2: What is the difference in user attributes which can help assessing credibility of Tweets during natural disasters from Europe and Asia?

Another important requirement for using Twitter in the contest of disaster is to be able to distinguish between signal and noise. We needed a simple, repeatable, and reproducible method to classify Tweets as disaster related and containing useful information. We therefore used a common approach in text classification, the supervised machine learning algorithm Naïve Bayes. The performance of any supervised machine learning algorithm is dependent on the training data-set used. During a real disaster, time is of the essence, and building a new training data-set for every event could result in a significant delay in classification (Spinsanti and Ostermann 2013).

One possible solution is crowdsourcing the labeling for timely preparation of training data for a particular disaster, which can be volunteered with no or limited quality assurance or may also be generated as a paid task with associated costs (Imran et al. 2014). While some researchers claim that classifiers trained for one disaster work well for another disaster of the same nature (Verma et al. 2011), others have shown that classification of specifically geographic information is a challenging task, often requiring local knowledge (Ostermann, Tomko, and Purves 2013). In our case, we used data related to two disasters of the same nature in two different continents. To explore the need to prepare new training data for every disaster, we formulated the following research question.

RQ3: How well does Naïve Bayes perform with respect to text classification of informational content for another event of the same nature, when training data for the classifier is trained using an event of a similar nature in a different location?

The overall aim of this research is to analyze Twitter usage characteristics of users residing in two different continents of the world typically characterized as

developed (Italy) and developing (Myanmar) regions. The main contribution of this study is to analyze how users from a developed country and a developing country report similar kind of disasters and how different or similar are the user-based credibility assessment features of Twitter users who are reporting about the disaster. This paper also analyzes the granularity of toponyms in Tweets.

2. Related work

In the following, we briefly introduce related work with respect to each of our three research questions. The potential role of VGI in disaster management (Goodchild and Glennon 2010) has become more important as mobile technologies have become increasingly ubiquitous (Sarda and Chouhan 2017). Thus, emerging technologies and the increased use of social media have changed the speed and ways in which people use and share information during disasters (Hughes et al. 2008). Ostermann and Spinsanti (2011) highlight the main challenges in using such content including a lack of structure to generate information (particularly in case of Twitter), the huge volume of data and a lack of quality control. As governmental authorities and disaster response agencies as well as individuals continue to use such data for disaster management, these challenges need to be addressed (Haworth and Bruce 2015).

Our first research question therefore concerns the granularity of locations present in Tweet content. Despite a wealth of research attempting to georeference Tweets, including many approaches using language-based models where toponyms are treated as potential features (e.g. Kinsella, Murdock, and O'Hare 2011), there is a dearth of research exploring the specifics of locational information associated with Tweets. Many locational models are implicitly very coarse, for example measuring accuracy with respect to 0.1° grids (e.g. Wing and Baldridge 2011).

However, conversely, studies using georeferenced Tweets typically assume that the coordinates associated with a Tweet accurately reflect the location of the content (e.g. Li, Lei, and Khadiwala 2012) despite more recent work suggesting a weak relationship between the locations of points of interest (POI) and content associated with these POIs (Hahmann, Purves, and Burghardt 2014). In practice, disaster-related applications of Twitter often seem to assume that data delivered are of a granularity appropriate to the task at hand, without any clear analysis of the ways in which locations are described in Tweet content, the association between content and locations and any analysis of variation in granularity as a function of the region being studied. While such assumptions may be justifiable when Tweets are averaged to create, for example, density surfaces, the granularity of locational information with respect to individual Tweets and their information content is important if these are to be treated as actionable information.

Our second research question focuses on the extraction and analysis of attributes argued to be associated with credibility in Twitter. According to the Merriam Webster dictionary, credibility is defined as "the quality of being believed or accepted as true, real, or honest" [1]. Despite the sheer volume of data shared on Twitter, not every Tweet provides information and facts related to an event (Gupta and Kumaraguru 2012). Rather, trending topics on Twitter, including disasters, can provide an opportunity for spammers to share spams using keywords associated with trending topics and generate revenue (Benevenuto et al. 2010). Such intrusions from spammers and other sources can make the credibility of information mined from social media platforms questionable (Morris et al. 2012). Senaratne et al. (2017) discussed possible quality indicators for VGI and argue that when International Standard Organization (ISO) standard measures are not applicable to assess quality, researchers tend to use more abstract indicators including credibility, trustworthiness, text content quality, etc. O'Donovan et al. (2012) suggest using features in Tweets such as "hashtags, reTweets and mentions" to predict the credibility of Tweet content. Conversely, Canini, Suh, and Pirolli (2011) use the approach of ranking individual social media users on the relevance and their expertise on the content they share to assess credibility of the content. Ostermann and Spinsanti (2012) successfully use geographic context information to assist in filtering relevant information on forest fires. Castillo, Mendoza, and Poblete (2011) combine four sets of features based on propagation, message, topic, and users to determine credibility of trending topics. They demonstrated that machine learning algorithms trained on these features can automatically classify credible and not credible trends with good precision and recall.

However, Gupta and Kumaraguru (2012) aim to assess credibility at the level of individual Tweets and argue that assessing credibility at a trending topic level is insufficient as trending topic about an earthquake may be true but Tweets about misleading magnitude can question the credibility. They used message and source-based features to determine credibility of individual Tweets. Castillo, Mendoza, and Poblete (2012) discuss different set of features such as Tweet length, friends count, followers count, etc. to determine credibility of Tweets. Becker, Naaman, and Gravano (2011) studied the techniques of assessing the quality of Tweets based on relevance to a particular topic instead of studying the truthfulness and factual credibility of Tweet content which is an important perspective in case of disasters. It is thus clear that credibility is a complex and important topic, where it is unclear which features and approaches are most appropriate in assessing credibility. Furthermore, it is also unclear how features thought to be associated with credibility vary in space as a result of, for example, different patterns in the use of Twitter in different locations.

Our final research question concerned the extraction of Tweets containing information using machine learning approaches. Extracting useful Tweets using crowdsourcing is a key task when dealing with large volumes of data which have been argued to be rapid and effective ways of collecting data for time-sensitive events such as natural disasters (e.g. Wald et al. 2011; Imran et al. 2014; Haubrock et al. 2017).

3. Methods

In this section following we first explain how our data-sets were collected, before describing our methods for exploring geographic granularity, study of credibility related features, and classification of information.

3.1. Data collection

We collected the Twitter data based on disaster-related keywords from the Twitter Streaming API. This API allows retrieval of Tweets in real time. The streaming API provides access to some 1–40% of Tweets.[2] We chose keywords to query the Twitter streaming API on general words used in English to refer to a hazard which can cause disaster. Query keywords used in the API are space sensitive but not case sensitive. The full set of keywords we used is illustrated in Table 1 and the data-set detailed in Table 2.

We aimed to collect only Tweets written in English, with no spatial restrictions, for the following reasons:

- English is one of the most frequently learned and spoken second languages worldwide.
- Many researchers have used English Tweets in their research.
- We were not familiar with all regional languages spoken in earthquake hit areas, making analysis in local languages difficult.

3.2. Geographic granularity of Tweets

We analyzed Tweet text to assess how users in different regions of the world (Asia and Europe) report an earthquake with its location. We selected 500 (Verma et al. 2011) Tweets through stratified sampling for each earthquake and manually analyzed the content and identified every geographic location (place name) reported in the Tweet text and the number of times it appeared in the sample data-set. These geographic locations were then identified in Geonames gazetteer, and we added feature classes to every location as per gazetteer on the list. While searching Geonames for geographic locations we came across ambiguous cases typically during geocoding:

(1) Presence of the same geographic location in different countries.
(2) Same geographic location categorized in different administrative hierarchies, e.g. Deoghar in India, is categorized as second-order administrative division as well as a populated place (City or Town).

To resolve the first ambiguity, we went through the full content of Tweet text to try to resolve the appropriate country. For the second case, we assumed that users are talking about finer granularity locations in the Geonames hierarchy (thus are more likely to be naming towns or villages than a containing administrative region of the same name).

We retrieved all the Tweets reporting on Myanmar and Italy earthquake from our database using these toponyms with "earthquake" as keywords which resulted in 47,557 Tweets for Myanmar and 234,620 Tweets for Italy. We counted the number of times a toponym occurred in whole data and compared them with number of occurrences of toponyms in sample data. We made this comparison to know the difference between the facts drawn from sample data vs. actual data. In sample data, there were some toponyms such as Tyrrhenian Sea, 66.6 miles from Vatican City, Himalayas, and South Indian Ocean which were not considered for toponyms count.

We performed a second comparison between hierarchies of toponyms according to Geonames gazetteer to analyze how users report about earthquake location in different times of during and post-disaster phases. We divided our data into 2-h intervals to make the difference more visible and counted the occurrence of every location in the group of geographic hierarchy occurring in the data. We have post disaster data for Italy, since the earthquake occurred at 01:36 UTC and first Tweet in our data is at 08:57 UTC, but for Myanmar, the data are during and post-disaster as earthquake occurred at 10:34 UTC and first Tweet in our data is at 10:36 UTC.

Table 1. Set of Keywords used to query Twitter Streaming API.

List of keywords used for the querying		
Tsunami	flood	Earthquake
Landslide	earth quake	fore shock
fore-shock	after shock	after-shock
landslide	land slide	Avalanche
rockfall	rock fall	mud slide
mudslide	earth slip	earth-slip
cloudburst	cloud burst	heavy rainfall
extensive rainfall	heavy rain	extensive rain
rain storm	forest fire	inundation
overflow	flash-flood	–

3.3. User-based attribute assessment

We adopted user-based features (Castillo, Mendoza, and Poblete 2011) for this case study to assess user-based attributes which are important for credibility assessment of Tweets (Table 3).

We selected the user provided "location" field to filter Tweets from our data-set for Italy and Myanmar. This field is entered by users at the time of creating their account, or may be added later, and is a free-text format

Table 2. Data-set details.

Size	Tweets	Start time	End time
2.54 GB	488,175	Wednesday 24 August 2016 08:57	Thursday 25 August 2016 08:57

Table 3. User-based attributes.

User-based features	Description
Registration age	The time passed since the author registered their account
Statuses count	The number of Tweets sent by the user
Followers count	Number of people following this user
Friends count	Number of people user is following
Verified	If the account has been verified
Has description	A non-empty bio
Has URL	A non-empty homepage URL

field. We wrote a new query for our research question two, because we wanted to collect only the Tweets for which the users claimed to be in earthquake hit location. For the Italian earthquake, we filtered our data-set based on a query which selected all the records which contain Italy in location field. For Myanmar earthquake, we used four countries India, Bangladesh, Myanmar, and Thailand, because Myanmar earthquake was felt in these four countries. This query returned 4773 records for Italy earthquake and 16,797 records for Myanmar earthquake. We selected 500 records by random sampling for each event to analyze credibility related user-based attributes of Tweets originating from these two regions. We assume that credibility is a function of user-based features, as follows:

$$C = f(FrC,\ SC,\ FoC,\ AG,\ U,\ D,\ V) \quad (1)$$

where C is credibility, FrC, SC, FoC, and AG are friends count, statuses count, followers count, and account age (in years), respectively. Other features such as U represent whether users are associated with a Uniform Resource Locator (URL), D whether users have added a description or bio, and V if a user has a verified account. These three features are represented by Boolean values. We compared the properties of each feature for our two areas, to test the hypothesis that credibility-related attributes varied according to locations.

3.4. Classification rules

We defined two categories to classify our data into two classes: Information and Not information. These classes are defined as follow:

- Information: Tweet text about disaster event and its location.
- Not Information: everything else falls in this category.

We used a supervised machine learning algorithm, Naïve Bayes, to classify Tweets according to frequency of earthquake-related terms in the sample corpus. We used a ratio of 7:3 for training and test data and tested Naïve Bayes on three different cases. We annotated 350 Tweets as "not information" class to train the classifier on not information class and also prepared 300 Tweets (150 information, 150 not information) as Italy earthquake test data and 300 Tweets (150 information, 150 not information) as Myanmar earthquake test data. This data remained the same in all three cases to train Naïve Bayes on "not information" class and test the classifier on Italy and Myanmar earthquake event.

For the first case, we annotated 350 Tweets from the Italian earthquake as information class, coupled with 350 Tweets as not information class to train the classifier, and we tested it on 300 Tweets prepared as Italy test data. Then we replaced independent entity geographic feature "Italy" in Italian training data-set for information class with "Myanmar" keeping rest of the content same and used this new data to train the classifier to run on Myanmar test data to explore the ability of the classifier to identify Tweets containing information when trained on annotated Tweets from a different region only with same geographic location in text.

For the second case, we annotated 350 Tweets from Myanmar earthquake event as information class to train the classifier on information class, and we tested it on Myanmar test data. Then we replaced independent entity geographic feature in Myanmar training data-set with "Italy" keeping rest of the content same and used this new data to train the classifier to run on Myanmar test data.

For the third case, the information class contained 175 Tweets from Italy and 175 Tweets from Myanmar earthquake event and tested it on Italy and Myanmar test data.

4. Results and interpretation

4.1. Geographic granularity according to Geonames

Figures 1 and 2 show the places named and their frequencies in Myanmar and Italy sample data-set. We attempted to use the hierarchy of administrative regions as used by Geonames to explore the granularity of the spatial information available.

However, though Myanmar appears to have information of finer granularities as Italy, it is clear that the toponyms used in Italy cover a much more tightly defined region, while for Myanmar, many Tweets appear to be from the surrounding countries.

Since our initial results are based on stratified sampling of Tweets, we also retrieved all Tweets containing these toponyms and explored their relative distribution in our corpus as a whole. The approach taken means that we only retrieve Tweets identified by our manual annotation, but gives insight into the usage of these place names in a much larger sample.

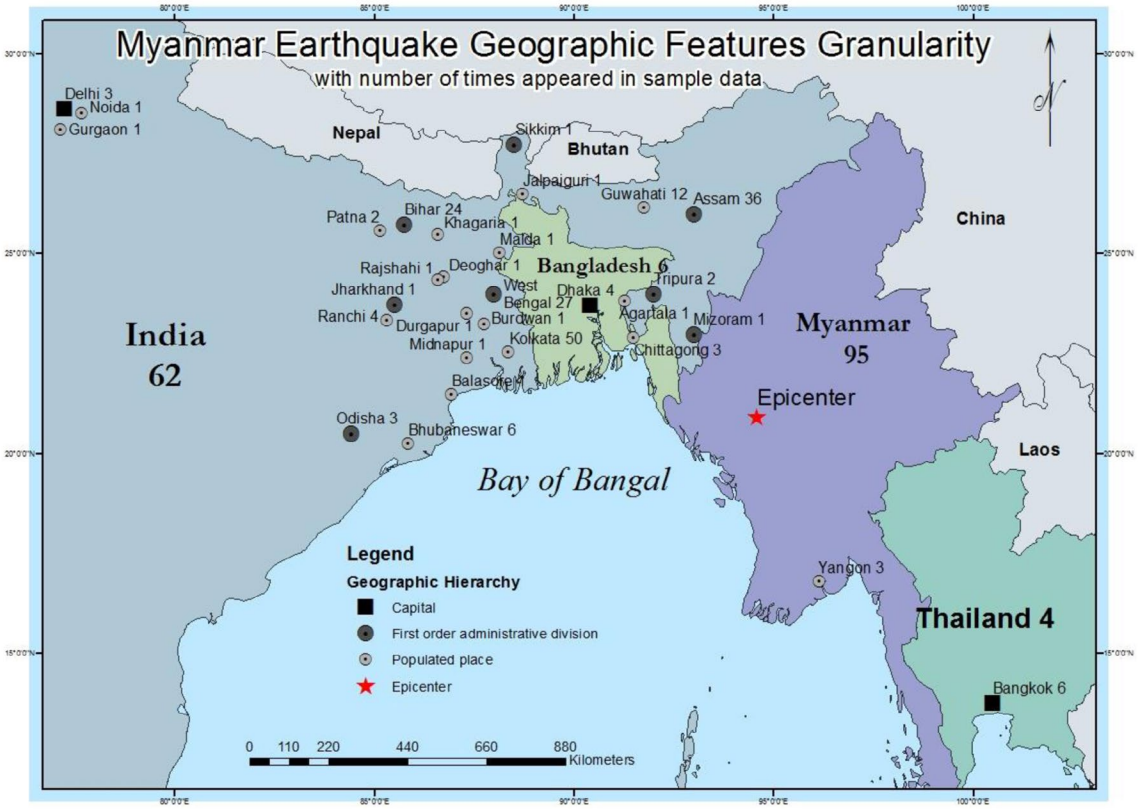


Figure 1. Myanmar earthquake geographic feature granularity according to Geonames.

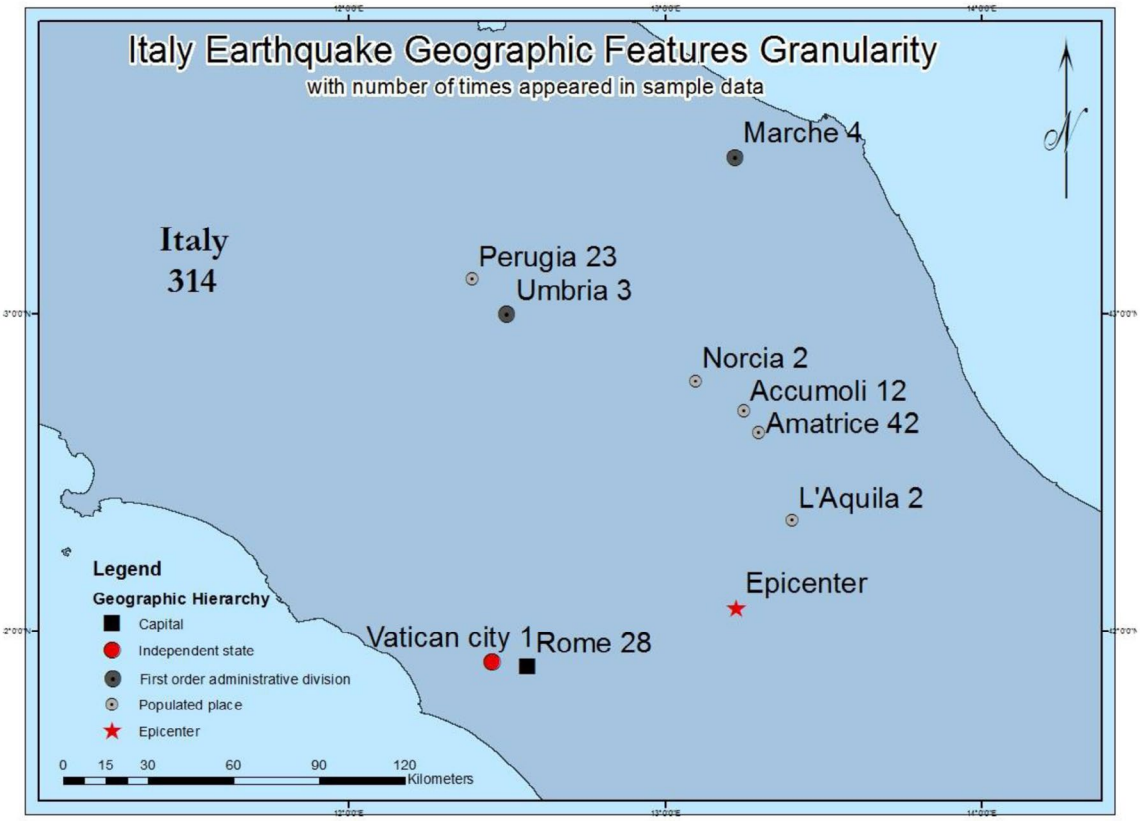


Figure 2. Italy earthquake geographic feature granularity according to Geonames.

The relatively small number of toponyms in the sample for Italy clearly shows that our sample data represent the overall distribution of toponyms well, and indeed this is confirmed by a Kendall's Tau Rank Correlations of 0.7 for Italy ($p < 0.05$). Notable are the prominence of very coarse grained toponyms (e.g.

Italy) and Rome (as the capital city) which provide very limited spatial information with respect to the earthquake.

In the case of Myanmar the picture is more complex, since very little data actually come from the country itself, other than in the form of the country name, and a large number of Tweets are from India. These Tweets appear to be primarily from regions where the earthquake was physically felt, but demonstrate a clear bias away from the areas most seriously affected by this event toward those where, we speculate, engagement with social media in general, and Twitter in particular is higher. Nonetheless, our sample appears to reflect overall behavior well with Kendall's Tau Rank Correlations of 0.67 for Myanmar ($p < 0.01$).

Figures 3 and 4 show the usage of toponyms over time for the two incidents. Notable is the relatively constant ratio of usage, with in both cases the country name being by far the most common, followed by populated places. Both data-sets also show a slow decline in Tweets using these place names immediately after the event.

4.2. User-based attributes assessment

We assessed the difference between a number of variables commonly associated with credibility for two events with the same number of Tweets and occurring at similar times. The count of friends, statuses, followers, and account ages are illustrated in Table 4. We tested significance of differences using a Mann–Whitney U test, and found that the count of friends, followers, and average account age were all significantly different ($p < 0.05$). However, these differences were asymmetric with accounts in Italy being associated with more friends and a greater account age, while those in Myanmar had more statuses (though not significantly) and more followers. Finally, we found that users in Italy were more likely to have URLs associated with their accounts, while there was little difference in the number of users with descriptions between the two locations.

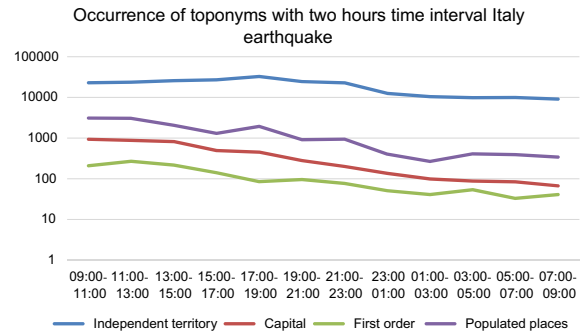


Figure 3. Occurrence of toponyms over time in Italy.

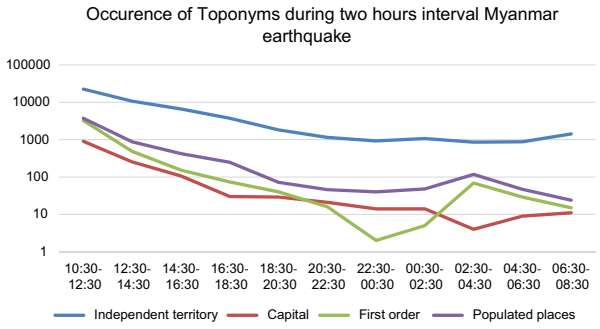


Figure 4. Occurrence of toponyms over time in Myanmar.

Table 4. Differences between credibility-related attributes.

Attribute	Italy	Myanmar
Friends count	1320 ± 3839	1073 ± 2369
Statuses count	31,498 ± 79,831	52,067 ± 101,762
Followers count	2082 ± 5479	3966 ± 22,622
Account age	5.32 ± 2.25	3.45 ± 2.53

These results point to the difficulty of assessing credibility using simple measures which are not normalized for local differences, since it appears that for events of the same class in different locations we find users with very different average behaviors, implying that a globally applied credibility metric is likely to capture differences in the local properties of Twitter users rather than differences in the credibility of content at these locations.

4.3. Classification results

For the first case, we used our test data to evaluate our classifier's performance on data from Italy (Table 5). The precision of our classifier was very high 98% for Tweets classified as containing information, suggesting that almost all Tweets classified using this approach contain information, while a recall of 93% means that a small number of Tweets were falsely discarded. When running the classifier on a different geographical region by replacing only geographical features in text, the performance decreased somewhat but remained relatively high (Table 6). This result has important implications, as it suggests that training data from other regions may help us extract information from Tweets.

For the second case, we applied the same approach as for case one but swap Italy with Myanmar. The results

Table 5. Confusion matrix for Italy (Case 1).

		Actual class		
Predicted class	Class	Information	Not information	Precision
	Information	147	3	98%
	Not information	11	139	92.667%
	Recall	93.038%	97.887%	–

Table 6. Confusion matrix for Myanmar (Case 1).

		Actual class		
Predicted class	Class	Information	Not information	Precision
	Information	133	17	88.667%
	Not information	12	138	92%
	Recall	91.724%	89.032%	–

Table 7. Confusion matrix for Myanmar (Case 2).

		Actual class		
Predicted class	Class	Information	Not information	Precision
	Information	149	1	99.333%
	Not information	11	139	92.667%
	Recall	93.125%	99.286%	–

Table 8. Confusion matrix for Italy (Case 2).

		Actual class		
Predicted class	Class	Information	Not information	Precision
	Information	141	9	94%
	Not information	37	113	75.333%
	Recall	79.213%	92.623%	–

Table 9. Confusion matrix for Italy (Case 3).

		Actual class		
Predicted class	Class	Information	Not information	Precision
	Information	146	4	97.333%
	Not information	36	114	76%
	Recall	80.22%	96.61%	–

Table 10. Confusion matrix for Myanmar (Case 3).

		Actual class		
Predicted class	Class	Information	Not information	Precision
	Information	146	4	97.333%
	Not information	37	113	75.333%
	Recall	79.781%	96.581%	–

(Table 7) for Myanmar show very high precision 99% with classifier trained on Myanmar data. For Italy (Table 8), the classifier was trained on Myanmar data but performed very well with 94% precision for Tweets reporting on Italian earthquake.

For the third case, results show (Tables 9 and 10) again very high precision 97% for both Italy and Myanmar as the classifier was trained on 50% data from Italy and 50% data from Myanmar.

5. Concluding discussion

In this paper, we set out to compare data related to natural hazard events that occurred more or less contemporaneously in two very different locations, Myanmar and Italy. When exploring the granularity of locations reported in Tweets, an initial analysis based only on hierarchies derived from Geonames suggested that the toponyms used in Myanmar of finer granularities were more common than in Italy. However, mapping the data clearly show the more or less total absence of detailed data in Myanmar, as compared to the finer data in Italy. These results reinforce the importance of considering data divides (e.g. Murthy 2011; Graham et al. 2014) when analyzing such data, and also reflect the difficulties of using VGI itself (here in the form of Geonames and Twitter) to do so. Bagan, the city in Myanmar which was the most affected city in terms of damages and casualties, was not reported in our sample data even once. It could be argued that this is simply a function of the sample of data which we annotated.

However, by extending our analysis using toponyms found as search terms in Tweets referring to earthquakes, we found that Kendall Tau rank correlations were high and statistically significant. This result has important implications, as it suggests, firstly, that a well-stratified data-set is adequate for the exploration of toponym usage. However, in terms of actionable information the slow decline of toponym usage immediately after such significant events suggests the challenges present in identifying truly actionable and local information from such data streams.

Our results exploring attributes commonly used in the assessment of credibility suggest an equally complex picture. We expected a clear difference between Tweets related to Italy and Myanmar, but in fact observe that, at least for user attributes these perhaps better reflect different user characteristics (users reporting on events in Asia appear to Tweet more often and have more followers, while those reporting on Europe seem to have older accounts and more friends). Since our results point to differences in the use of Twitter in different locations, and these differences are reflected in attributes previously associated with credibility, we suggest that efforts on understanding credibility would be better focused on content rather than proxy information. One promising approach to such problems is that proposed by Truelove, Vasardani, and Winter (2015) who aim to identify first-person witness accounts in Twitter.

Underpinning the importance of looking at content when trying to understand the nature of informational content was the performance of simple, off-the-shelf machine learning methods in classifying Tweets. Here we found that, independent of the location of Tweets being classified or the training data used we could identify Tweets containing information with a precision of the order of between 88 and 99%. Recall, while the lower value (with a minimum of 79%) was also satisfying, and again we argue that this result points to the importance of analyzing content when assessing the quality of information provided by Twitter with respect to natural hazards.

This paper contributes to our understanding of how social media sources are being used in different geographic regions, and the important question which analytical approaches may be suitable to transfer and reproduce methods from one geographic region to another. While our results once again highlight the often underestimated digital divide, the successful use of relatively straightforward analytical methods to both data-sets promises that global body of knowledge and methodological toolkit is possible. However, it is important to also be clear in stating that

our approaches currently are far from being capable of identifying actionable and additional information suitable for use in applications of such data. We suggest that in the rush to exploit social media to produce academic output, there is also an urgent need for more thoughtful and critical work such as the analysis of Mission 4636 after the Haiti earthquake by Munro (2013) and we repeat verbatim his important conclusion "It is recommended that future humanitarian deployments of crowdsourcing focus on information processing within the populations they serve, engaging those with crucial local knowledge wherever they happen to be in the world."

Notes

1. https://www.merriam-webster.com/dictionary/credibility
2. https://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/

Acknowledgments

The first author of this paper would like to thank her colleagues at the University of Zurich Ms. Flurina Wartmann and Mr. Oliver Burkhard for helping at various stages of analyzing the data and results.

Notes on contributors

Kiran Zahra is a PhD student in department of Geography at the University of Zurich. Her research interests are Twitter, natural disaster management, geographic information retrieval, and data quality.

Frank O. Ostermann joined the ITC in 2014 as an assistant professor for Cloud and Crowd geo-information processing. His main research interests are the opportunities and challenges of distributed collection, storage, processing and sharing of crowd-sourced and volunteered geographic information. Since 2009, he holds a PhD (Dr. Sc. Nat.) in Geographic Information Science from the University of Zurich. Previous work includes three years as a post-doctoral researcher at the Joint Research Center of the European Commission, as well as several years as a research assistant at the University of Zurich and Hamburg on several EU-funded projects on user-generated geographic content, and spatio-temporal data analysis and visualization in urban contexts.

Ross S. Purves is concerned with methods to analyze and understand geography through a range of different approaches, including the analysis of user generated content and text more generally. He is based at the Department of Geography of the University of Zurich, where he leads the Geocomputation Group in the GIScience Centre.

References

Becker, H., M. Naaman, and L. Gravano. 2011. "Selecting Quality Twitter Content for Events." Paper presented at The International Conference on Weblogs and Social Media, Barcelona, Spain, July 17–21.

Benevenuto, F., G. Magno, T. Rodrigues, and V. Almeida. 2010. "Detecting Spammers on Twitter." Paper presented at The Seventh annual Collaboration, Electronic messaging, AntiAbuse and Spam Conference, Redmond, WA, USA, July 13–14.

Canini, K. R., B. Suh, and P. L. Pirolli. 2011. "Finding Credible Information Sources in Social Networks Based on Content and Social Structure." Paper presented at The IEEE Third International Conference on Privacy, Security, Risk and Trust, Boston, MA, USA, October 9–11.

Castillo, C., M. Mendoza, and B. Poblete. 2011. "Information Credibility on Twitter." Paper presented at The 20th International Conference on World Wide Web, Hyderabad, India, March 28–April 1, 675–684.

Castillo, C., M. Mendoza, and B. Poblete. 2012. "Predicting Information Credibility in Time-sensitive Social Media." *Internet Research* 23 (5): 560–588. doi:10.1108/IntR-05-2012-0095.

Goodchild, M. F., and J. A. Glennon. 2010. "Crowdsourcing Geographic Information for Disaster Response: A Research Frontier." *International Journal of Digital Earth* 3 (3): 231–241.

Graham, M., B. Hogan, R. K. Straumann, and A. Medhat. 2014. "Uneven Geographies of User-generated Information: Patterns of Increasing Informational Poverty." *Annals of the Association of American Geographers* 104 (4): 746–764.

Gupta, A., and P. Kumaraguru. 2012. "Credibility Ranking of Tweets during High Impact Events." Paper presented at The 1st Workshop on Privacy and Security in Online Social Media, Lyon, France, April 17. doi:10.1145/2185354.2185356.

Hahmann, S., R. S. Purves, and D. Burghardt. 2014. "Twitter Location (Sometimes) Matters: Exploring the Relationship between Georeferenced Tweet Content and Nearby Feature Classes." *Journal of Spatial Information Science* 2014 (9): 801–802.

Haubrock, S., C. Little, S. Mcbride, and N. Balfour. 2017. "Keeping up with the Citizens-collecting Earthquake Observations in New Zealand." Paper presented at AGILE Conference 2017, Wageningen, The Netherlands, May 9–12. https://www.cs.nuim.ie/~pmooney/VGI.

Haworth, B., and E. Bruce. 2015. "A Review of Volunteered Geographic Information for Disaster Management." *Geography Compass* 9 (5): 237–250.

Hecht, B., L. Hong, B. Suh, and E. H. Chi. 2011. "Tweets from Justin Bieber's Heart: The Dynamics of the Location Field in User Profiles." Paper presented at The ACM Conference on Human Factors in Computing Systems (CHI 2011), Vancouver, BC, Canada, May 7–12, 237–246.

Hossmann, T., P. Carta, D. Schatzmann, F. Legendre, P. Gunningberg, and C. Rohner. 2011. "Twitter in Disaster Mode:Security Architecture." Paper presented at SWID '11 – The Special Workshop on Internet and Disasters, Tokyo, Japan, December 6–9. doi:10.1145/2079360.2079367.

Hughes, A. L., L. Palen, J. Sutton, S. B. Liu, and S. Vieweg. 2008. "Site-seeing in Disaster: An Examination of On-line Social Convergence." Paper presented at The 5th International ISCRAM Conference, Washington, DC, May 5–7.

Imran, M., C. Castillo, J. Lucas, P. Meier, and S. Vieweg. 2014. "AIDR: Artificial Intelligence for Disaster Response." Paper presented at The 23rd International Conference on World Wide Web, Seoul, April 7–11, 159–162. doi:10.1145/2567948.2577034.

Kinsella, S., V. Murdock, and N. O'Hare. 2011. "'I'm Eating a Sandwich in Glasgow': Modeling Locations with Tweets." Paper presented at The 3rd International Workshop on Search and Mining User-generated Contents, Glasgow, UK, October 24–28, 61–68.

Li, R., K. H. Lei, R. Khadiwala, and C. C. Chang. 2012. "TEDAS: A Twitter-based Event Detection and Analysis System." Paper presented at The 28th IEEE International Conference on Data Engineering, Washington, DC, USA, April 1–5, 1273–1276. doi:10.1109/ICDE.2012.125.

Middleton, S. E., L. Middleton, and S. Modafferi. 2014. "Real-time Crisis Mapping of Natural Disasters Using Social Media." *IEEE Intelligent Systems* 29 (2): 9–17. doi:10.1109/MIS.2013.126.

Morris, M. R., S. Counts, A. Roseway, A. Hoff, and J. Schwarz. 2012. "Tweeting is Believing? Understanding Microblog Credibility Perceptions." Paper presented at The ACM 2012 Conference on Computer Supported Cooperative Work, Seattle, WA, USA, February 11–15. 441–450.

Munro, R. 2013. "Crowdsourcing and the Crisis-affected Community." *Information Retrieval* 16 (2): 210–266.

Murthy, D. 2011. "Twitter: Microphone for the Masses?" *Media Culture and Society* 33 (5): 779–789.

O'Donovan, J., B. Kang, G. Meyer, T. Höllerer, and S. Adalii. 2012. "Credibility in Context: An Analysis of Feature Distributions in Twitter." Paper presented at 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust, Washington, DC, USA, September 3–5, 293–301.

Ostermann, F., and L. Spinsanti. 2012. "Context Analysis of Volunteered Geographic Information from Social Media Networks to Support Disaster Management: A Case Study on Forest Fires." *International Journal of Information Systems for Crisis Response and Management* 4 (4): 16–37.

Ostermann, F. O., and L. Spinsanti. 2011. "A Conceptual Workflow for Automatically Assessing the Quality of Volunteered Geographic Information for Crisis Management." Paper presented at AGILE Conference 2011. Utrecht, The Netherlands, April 18–21.

Ostermann, F. O., M. Tomko, and R. Purves. 2013. "User Evaluation of Automatically Generated Keywords and Toponyms for Geo-referenced Images." *Journal of the American Society for Information Science and Technology* 64 (3): 480–499.

Sakaki, T., M. Okazaki, and Y. Matsuo. 2010. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors." Paper presented at The 19th international conference on World Wide Web, Raleigh, NC, USA, April 26–30, 851–860. doi:10.1145/1772690.1772777.

Sakaki, T., M. Okazaki, and Y. Matsuo. 2013. "Tweet Analysis for Real-time Event Detection and Earthquake Reporting System Development." *IEEE Transactions on Knowledge and Data Engineering* 25 (4): 919–931. doi:10.1109/TKDE.2012.29.

Sarda, P., and R. L. Chouhan. 2017. "Extracting Non-situational Information from Twitter During Disaster Events." *Journal of Cases on Information Technology* 19 (1): 15–23.

Senaratne, H., A. Mobasheri, A. L. Ali, C. Capineri, and M. Haklay. 2017. "A Review of Volunteered Geographic Information Quality Assessment Methods." *International Journal of Geographical Information Science* 31 (1): 139–167.

Spinsanti, L., and F. Ostermann. 2013. "Automated Geographic Context Analysis for Volunteered Information." *Applied Geography* 43: 36–44.

Terpstra, T., A. D. Vries, R. Stronkman, and G. L. Paradies. 2012. "Towards a Realtime Twitter Analysis during Crises for Operational Crisis Management." Paper presented at The 9th International ISCRAM Conference, Vancouver, Canada, April 22–25.

Truelove, M., M. Vasardani, and S. Winter. 2015. "Towards Credibility of Micro-blogs: Characterising Witness Accounts." *GeoJournal* 80 (3): 339–359.

Verma, S., S. Vieweg, W. Corvey, L. Palen, J. H. Martin, M. Palmer, A. Schram, and K. M. Anderson. 2011. "NLP to the Rescue?: Extracting 'Situational Awareness' Tweets during Mass Emergency." Paper presented at The Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, July 17–21.

Wald, D. J., V. Quitoriano, C. B. Worden, and M. Hopper. 2011. "USGS 'Did You Feel It?' Internet-based Macroseismic Intensity Maps." *Annals of Geophysics* 54 (6): 688–707.

Wing, B. P., and J. Baldridge. 2011. "Simple Supervised Document Geolocation with Geodesic Grids." Paper presented at The Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, June, 19–24.

Zahra, K., and R. Purves. 2017. "Analysing Tweets Describing during Natural Disasters in Europe and Asia." Paper presented at AGILE Conference 2017. Wageningen, The Netherlands, May 9–12. https://www.cs.nuim.ie/~pmooney/VGI.

B

PUBLICATION II: AUTOMATIC IDENTIFICATION OF EYEWITNESS MESSAGES ON TWITTER DURING DISASTERS

Zahra, K., Imran, M., & Ostermann, F. O. (2020). Automatic identification of eyewitness messages on twitter during disasters. *Information processing and management*, 57(1), 102107.

Information Processing and Management 57 (2020) 102107

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Automatic identification of eyewitness messages on twitter during disasters

Kiran Zahra[a,*], Muhammad Imran[b], Frank O. Ostermann[c]

[a] University of Zurich, Switzerland
[b] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[c] University of Twente, The Netherlands

ARTICLE INFO

Keywords:
Social media
Eyewitness identification
Machine learning
Disaster response

ABSTRACT

Social media platforms such as Twitter provide convenient ways to share and consume important information during disasters and emergencies. Information from bystanders and eyewitnesses can be useful for law enforcement agencies and humanitarian organizations to get firsthand and credible information about an ongoing situation to gain situational awareness among other potential uses. However, the identification of eyewitness reports on Twitter is a challenging task. This work investigates different types of sources on tweets related to eyewitnesses and classifies them into three types (i) direct eyewitnesses, (ii) indirect eyewitnesses, and (iii) vulnerable eyewitnesses. Moreover, we investigate various characteristics associated with each kind of eyewitness type. We observe that words related to perceptual senses (feeling, seeing, hearing) tend to be present in direct eyewitness messages, whereas emotions, thoughts, and prayers are more common in indirect witnesses. We use these characteristics and labeled data to train several machine learning classifiers. Our results performed on several real-world Twitter datasets reveal that textual features (bag-of-words) when combined with domain-expert features achieve better classification performance. Our approach contributes a successful example for combining crowdsourced and machine learning analysis, and increases our understanding and capability of identifying valuable eyewitness reports during disasters.

1. Introduction

At times of disasters caused by natural and anthropogenic hazards, people use social media platforms such as Twitter and Facebook to share information (Imran, Castillo, Diaz, & Vieweg, 2015; Vieweg, Hughes, Starbird, & Palen, 2010) that can potentially be useful for disaster response. This information includes reports of injured and dead people, urgent needs of affected people, reports of missing and found people, and reports of unrest and looting, among others (Imran et al., 2015). Social media not only contains useful information, it also breaks stories and events faster than many other traditional information or news sources such as TV. For instance, the first report of the Westgate Mall attack[1] in Nairobi, Kenya in 2013 was published on Twitter, almost 33 minutes before a local TV channel reported the event. Similarly, the news about the Boston bombing incident[2] appeared on Twitter before any other

* Corresponding author.
E-mail addresses: kiran.zahra@geo.uzh.ch (K. Zahra), mimran@hbku.edu.qa (M. Imran), f.o.ostermann@utwente.nl (F.O. Ostermann).
[1] https://en.wikipedia.org/wiki/Westgate_shopping_mall_attack .
[2] https://en.wikipedia.org/wiki/Boston_Marathon_bombing .

https://doi.org/10.1016/j.ipm.2019.102107
Received 8 April 2019; Received in revised form 17 July 2019; Accepted 26 August 2019
Available online 27 September 2019

news channel reported the event. Likewise, in the case of the California earthquake[3] it was observed that the first half dozen tweets were recorded by Twitter about a minute earlier than the recorded time of the event according to the USGS. These first-hand reports come from eyewitnesses and bystanders, i.e. people who directly observe the occurrence of an event (Diakopoulos, De Choudhury, & Naaman, 2012; Zahra, Imran, & Ostermann, 2018).

At the onset of a disaster event, people share massive amounts of data such as damage reports, casualties, but much of that data has redundant information, e.g. through sharing the same news article or video. For instance, millions of messages were posted on Twitter during Hurricane Harvey in 2017, many containing similar information.[4] Nevertheless, studies have revealed that the information sources also include many local citizens, bystanders, and eyewitnesses. From the perspective of an information seeker (affected citizen or institutional response agency), information from eyewitness reports is preferred over other types of information sources (e.g. people outside the disaster area). Law enforcement agencies and first responders always look for first-hand and credible information for decision-making. Humanitarian organizations look for timely and trustworthy information that is directly observed from the disaster-hit areas to better estimate the severity and scale of damage, and the amount of aid required to help save lives and fulfill the urgent needs of affected people.

Gaining rapid access to the information shared by eyewitness reports, especially during an ongoing disaster event, is thus useful but challenging to obtain (Imran, Mitra, & Srivastava, 2016). The most straightforward approach to identify local residents in disaster-hit areas is through geotagged information, e.g. Twitter messages (called tweets) with attached coordinates from a global navigation satellite system (e.g. the Global Positioning System, or GPS). However, given that only 1–3% of tweets are geotagged, relying solely on those to identify local residents may not provide enough data required for decision-making. Moreover, not all tweets from the disaster-struck area automatically come from eyewitnesses. Many social media platforms allow the user to enter manually a home location in the user profile, but research has shown that at least in the case of Twitter it is very noisy and inaccurate, and it does not indicate location of the source at the time when a tweet is made (Lee, Ganti, Srivatsa, & Liu, 2014).

Given the above issues, it remains a challenge to process potentially millions of tweets and identify eyewitnesses reliably. In Doggett and Cantarero (2016), the authors identify a set of eyewitness and non-eyewitness linguistic features to categorize eyewitness news-worthy events on human-induced disasters such as protests, shooting, and police activities. Likewise, in Fang, Nourbakhsh, Liu, Shah, and Li (2016), the authors highlight a similar set of linguistic (e.g. personal or impersonal expressions, time awareness) and meta-features (e.g. client application) to identify witness reports on various natural and human-induced disasters. They also used the topic (e.g. accidents, crimes and disasters) of tweets as a feature to automatically classify tweets as witness reports. The work presented in Tanev, Zavarella, and Steinberger (2017) identified a set of eyewitness features from several dimensions and categorized stylistic, lexical Twitter metadata and semantic features, and Truelove, Vasardani, and Winter (2014) developed a generalized conceptual model of different types of eyewitness reports for several events such as concerts, shark sightings, cyclones, and protests. However, none of the works (i) develop their classifiers through a combination of expert-driven and data-driven feature engineering, and (ii) differentiate between particular types of eyewitnesses found during natural disasters, and (iii) operationalize different characteristics associated with those types, and finally (iv) validate the results for different disaster types using crowdsourced annotations.

This paper aims to address this research gap by designing an eyewitness reports taxonomy focusing on the needs of disaster response agencies during natural disasters. Moreover, we explore different types of features associated with tweets to train machine learning models for the automatic classification of tweets. For this purpose, we first manually learn different characteristics (mainly language based) from tweets posted by eyewitnesses. We use these characteristics as independent (domain-experts) features together with content-based features from tweets to train several machine learning models. We establish that when domain-expert features are combined with the text-based features of tweets, these models outperform those which are trained on independent features. Since creating a balanced labeled dataset from social media labeled data is a challenging task, we employee a state-of-the-art class balancing technique and demonstrate that a model trained on a balanced data can achieve even higher results. Lastly, we contribute a successful example of combined crowdsourced training of machine learning classification (Imran, Lykourentzou, Naudet, & Castillo, 2013; Ostermann, Garcia-Chapeton, Kraak, & Zurita-Milla, 2018). The contributions of this work are summarized as follows:

- Designing a taxonomy consisting of different sub-types of eyewitnesses i.e., direct eyewitness, indirect eyewitness, vulnerable direct eyewitness.
- A generalized methodology, which can be applied on textual data from other domains to extract eyewitness reports, that uses textual content-based features and domain-expert features without relying on platform-specific features such as Twitter metadata.
- Combining textual and domain-expert features to train and evaluate automatic machine learning classifiers on real-world disaster-related Twitter datasets.
- Last but not least, we offer all the labeled data obtained through crowdsourcing and manual analysis to the research community to further develop and extend this line of research. The dataset will be shared at the CrisisNLP repository: https://crisisnlp.qcri.org/.

The following section gives a brief overview of related work, before we describe in more detail the methodology and obtained results in their respective sections. Before concluding, we then discuss lessons learnt and evaluate our work.

[3] http://latimesblogs.latimes.com/technology/2008/07/twitter-earthqu.html .
[4] https://crisiscomputing.qcri.org/2017/09/27/hurricane_harvey_and_the_role_ai/ .

2. Literature review

Our study uses a Twitter dataset. Twitter is a well established source to harvest opportunistically information during crisis events. Twitter messages are called tweets, and are micro blog posts, i.e. small packets of information of originally 140 characters (now doubled to 280). By default, all posts are public and can be found by everyone using the proper search parameters. Tweets can use hashtags to facilitate this search. Additionally, users can retweet any tweet and follow each other to see posts within their network on their timeline. Therefore, sharing information has zero marginal cost for the user: Posting a tweet requires only a smartphone, a Twitter app or log-in through the web interface, and a comparatively narrow network bandwidth (more if videos are to be shared). According to Twitter usage statistics,[5] around 500 million tweets are posted per day. A challenge for utilizing this massive volume of information is the often informal, unstructured, and noisy nature of Twitter posts and communication.

Twitter also has a well-establish history in breaking news in real-time. These news are often generated/started by a person who has witnessed the event. A very recent example is the eyewitness report of an emergency landing of Delta Aircraft due to its engine failure in Alaska. This eyewitness report has enough credibility to become the news source for a traditional news agency.[6] Another classic example of an eyewitness report on Twitter is the New York airplane crash in Hudson bay in 2009. This tweet also became the headline of The Daily Telegraph.[7] In Kwak, Lee, Park, and Moon (2010), the authors argue that Twitter serves also as news source and not only as social media platform. Their research reveals that over 85% of trending topics are news headlines.

Because many Twitter users post their personal experiences during a natural disaster, people are motivated to search for breaking news and real-time content on Twitter (Teevan, Ramage, & Morris, 2011) as in case of disasters (Allen, 2014; Amaratunga, 2014; Kryvasheyeu et al., 2016; Schnebele et al., 2013). In Oh, Agrawal, and Rao (2013), the authors explore the use of Twitter during social crisis situations. Academic research into Twitter and disaster management has mainly focused on user contributed data in disaster response (Haworth & Bruce, 2015) and relief phase (Landwehr & Carley, 2014) such as the Haiti earthquake in 2010 (Meier, 2012) or during forest fires (Ostermann & Spinsanti, 2012).

In the case of emergency events, extraction of relevant and reliable information from noise and redundant information is critical. Originally, researchers and relief organizations alike attempted to crowdsource the curation of information processing during disasters. Early systems such as CrisisCamp[8] and Ushahidi[9] are well-established platforms to gather and network volunteers from all over the globe to help solving different problems using a collective wisdom. Imran et al. (2014) used Standby Task Force[10] volunteers to label if the tweet belongs to information category. The work presented in Purohit et al. (2014) used Crowdflower[11] (now figure-eight) to classify tweets in categories such as requests for help or offers to provide help, among others. They also used crowdsourcing to label the resources available during the crisis. However, their results focused on the evaluation of their machine learning model. They did not evaluate the performance of their crowdsourcing task itself. In Snow et al. (2008), the authors used Amazon Mechanical Turk[12] (AMT) to evaluate the effectiveness of tasks performed by various “expert” and “non-expert volunteers” and suggested a technique to remove bias. In Callison-Burch (2009), the authors used AMT to evaluate the translation of different texts a much faster and cheaper than the conventional ways.

To facilitate and support these tasks, humanitarian and disaster relief organizations also developed real-time tweet crawler applications such as TweetTracker (Kumar, Barbier, Abbasi, & Liu, 2011), Artificial Intelligence for Disaster Response (AIDR) (Imran, Castillo, Lucas, Meier, & Vieweg, 2014), Twitcident (Abel, Hauff, Houben, Stronkman, & Tao, 2012), ScatterBlogs for situational awareness (Thom et al., 2015), cross-language aspects on Twitter (Imran et al., 2016), or during a particular disaster such as Typhoon Haiyan in the Philippines (Takahashi, Tandoc, & Carmichael, 2015). In Zahra, Ostermann, and Purves (2017), the authors discussed a time-saving and efficient technique to filter the noise from informative tweets. However, crowdsourced social media data generated by often anonymous users suffers from an absence of quality assurance (Goodchild & Li, 2012) on the truthfulness, objectivity, and credibility of the information.

Disaster response organizations search for eyewitness reports as those are considered more credible (Truelove, Vasardani, & Winter, 2015). Researchers have studied possibilities to identify eyewitness reports out of millions of tweets for journalism (Diakopoulos et al., 2012), criminal justice, and natural disasters (Olteanu, Vieweg, & Castillo, 2015). The work in Morstatter, Lubold, Pon-Barry, Pfeffer, and Liu (2014) relates the identification of eyewitness tweets to the use of language and linguistic patterns within the region during different crisis events. They also identified a set of features to automatically classify eyewitness reports. In Kumar, Morstatter, Zafarani, and Liu (2013), the authors use location information of the users to assess local users and remote users on crisis reports.

However, both research strands have worked mostly in isolation until now, with the potential of location information not fully exploited for establishing whether a source is an eyewitness, who might also be vulnerable and at risk. This paper aims to develop a holistic categorization of characteristics of eyewitness reports for frequent types of natural disasters.

[5] http://www.internetlivestats.com/twitter-statistics/ .
[6] https://channelnewsasia.com/news/world/delta-flight-middle-of-the-ocean-seattle-beijing-emergency-land-11062706 .
[7] https://www.telegraph.co.uk/technology/twitter/4269765/New-York-plane-crash-Twitter-breaks-the-news-again.html .
[8] https://crisiscommons.org/crisiscamp/ .
[9] https://www.ushahidi.com/ .
[10] http://www.standbytaskforce.org/about-us/our-history/ .
[11] https://www.figure-eight.com/ .
[12] https://www.mturk.com/ .

3. Methodology & experimental framework

This work identifies eyewitnesses messages from Twitter data using the following four steps:

1. **Disaster-related data collection:** First, we collect Twitter data related to four different types of natural disasters. Specifically, several months of data about earthquakes, floods, wildfires, hurricanes are collected and used in this work. We then retrieved two data samples from our corpus. Our first sample is used for an initial manual analysis (see next step) and includes data from three natural disasters types: earthquake, hurricane, and floods. Our second sample is used for crowdsourced annotation and includes data on four natural disaster types: earthquake, hurricane, floods, and wildfire. Data collection details are described in the next section.
2. **Manual Analysis to determine eyewitnesses types and characteristics:**To determine different types of eyewitness reports, we first analyse our first data sample taken from the collected data for three disaster types. Next, tweets which are identified as posted by eyewitnesses are further analyzed to understand different linguistic characteristics using their content (i.e., message text). The following annotation guidelines were developed and followed for the manual analysis:
   - **Identify tweet source:** This task aims to determine the source of a given tweet i.e., whether it is posted by an eyewitness or not, using only the message content. For this purpose, we consider the following three categories for the analysis:
     (i) **Eyewitness:** if the message is posted by an eyewitness
     (ii) **Non-eyewitness:** if the message is posted by anyone else other than an eyewitness
     (iii) **Don't know:** if it is not possible to determine which of the above two categories
   - **Identify eyewitness type:** If the previous task identifies a tweet as posted by an eyewitness, then this task aims to further determine whether the author of the tweet is a direct eyewitness, or passes on information he/she learned from a familiar source or is otherwise familiar with. We term that second type of eyewitness an *indirect eyewitness*. We also categorize a group of eyewitness tweets as *vulnerable eyewitness* where people were anticipating a disaster and were present in the region for which disaster warnings were issued.
   - **Identify eyewitness message characteristics:** This task aims to identify various linguistic characteristics and clues from the contents of eyewitness tweets. In the remainder of the paper, we refer to these as domain-expert features.
3. **Crowdsourcing to obtain labeled data:**The types of eyewitnesses and content characteristics identified in the previous step are then used to obtain labeled data using a paid crowdsourcing platform on our second data sample which includes tweets from four natural disasters i.e. earthquake, hurricane, flood, and wildfire.
4. **Training supervised machine learning models:**We consider the task of determining whether a tweet is posted by an eyewitness or not as a classification task. We use supervised machine learning techniques to train models on the crowdsourced labeled data using the following steps:
   - **Automatic feature extraction**: we first extract textual features from the textual content of the labeled tweets. For this purpose, we use two types of textual features (i) uni-grams and (ii) bi-grams and compute their TF-IDF scores Hong, Dan, and Davison (2011).
   - **Feature selection**: Feature selection techniques help identify features which help classifiers discriminate well between different classes and also help in generalization. We use the information gain feature selection technique to choose the top performing features for each class.
   - **Supervised models learning**: Among different learning schemes, Random Forest is considered best for the classification of textual data (Xu, Guo, Ye, & Cheng, 2012). We use Random Forest to train our classification models. Specifically, for each event type (e.g., earthquake), we train four different models as follows:
     (i) **Training using textual features (baseline):** In this case, we only use Bag-of-Words (BOW) based features extracted from the content of the labeled tweets.
     (ii) **Training using domain-expert features:** We extract features from tweets content using the characteristics identified by the domain-experts and use them to train new models.
     (iii) **Training using text and domain-expert features:** In this case, both text-based and domain-experts features are combined to train models.
     (iv) **Training using text and domain-expert features with class balancing:** Models trained on imbalanced classes always suffer performance issues. To tackle this problem, we used SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), which is a well-known class balancing technique. We retrained our models on reasonably balanced classes using the combined features from text and domain-experts.

The models performance is evaluated using the cross-validation (10-fold) technique and presented using standard performance evaluation metrics such as precision, recall, and F-measure.

4. Data, manual analysis and crowdsourcing

In this section, we provide details of the data collection, manual analysis, and crowdsourced annotation.

Table 1
Frequency of eyewitness, non-eyewitness, and unknown reports.

Event type	Sampled Tweets	Eyewitness	Non eyewitness	Don't know
Floods	2000	148	113	1739
Earthquakes	2000	367	321	1312
Hurricanes	2000	296	100	1604

4.1. Data collection

We used the Twitter Streaming API to collect data from July 2016 to May 2018 using a methodology described in this paper (Zahra et al., 2017). Specifically, we used *earthquake, foreshock, aftershock, flood, inundation, extensive rain, heavy rain, hurricane, cloud-burst, forest fire*, and *wildfire* keywords to collect in total 25 million tweets related to earthquakes, hurricanes, floods, and forestfires. For the manual analysis, we used two samples from this tweet corpus. Our first sample was retrieved from 1 to 28 August 2017 from three disaster types i.e., earthquake, flood, and hurricane. We chose this time period because of the occurrence of several such disaster events during that period. Our second sample was retrieved from July 2016 to May 2018 - but excluding the first sample's time period for earthquake - for the same three disaster types, and wildfire. Each sample consisted of 2000 randomly selected tweets from each disaster type. Our first data sample was used for the manual analysis and our second data sample was used for crowdsourcing.

4.2. Manual analysis results

Following the *Identify tweet source* annotation task guidelines described in the previous section, two authors of this paper manually analysed every tweet in the sample. Table 1 shows the results of this manual analysis. The number of tweets posted by eyewitnesses is very limited. A total of 148, 367, and 296 messages were found as posted by an eyewitness for floods, earthquakes, and hurricanes respectively. For many tweets it was not possible to determine with sufficient reliability whether they were posted by eyewitnesses or not, i.e., the *don't know* cases in the last column of Table 1. This difficulty already hints at the challenge automated classification may face.

Next, by following the *Identify eyewitness type* annotation task guidelines, tweets posted by eyewitnesses are analyzed further to determine if there are different types of eyewitness reports. Table 2 shows the results of this analysis. Mainly, three types of eyewitnesses are identified namely (i) direct eyewitness, (ii) indirect eyewitness, and (iii) vulnerable direct eyewitness. We provide details for each of the type in the following subsections.

4.2.1. Direct eyewitness

A direct eyewitness report represents first-hand knowledge and experience of an event. There are different ways in which direct eyewitness reports can provide information on events. Table 3 shows some examples of direct eyewitness reports taken from the manually analyzed data from all three disaster types.

The first message on floods reports the personal experience of the author about a flood situation. The second message is even more interesting since the author not only reports the event, he/she also complains about a lack of notifications or flood warnings in his/her area. Similarly, in the third message the author reports about high flood waters and that he/ she has got stuck due to it. The fourth message is also about a personal experience of a flash flood situation. All the earthquake-related messages in Table 3 express personal experiences of the authors about some earthquake events. We observe that in most of the earthquake cases, people express or relate their messages to the sense of feeling such as "just felt" or "feeling shaking".

Regarding hurricane-related messages, the first and second example report about winds and heavy rain, which are obvious signs of a direct personal experience. Both authors experienced the situation and reported it, while the author of the third message not only reported a flood situation but also gave an indication that the situation could get worse. The last example is again a personal experience of an event where the author is also reporting a power outage.

One common observation from the analysis of direct eyewitness reports is that eyewitnesses often mention the severity of situation they are in. Moreover, people associate their messages to different senses like "seeing", "feeling", "hearing" or "smelling". For example, in case of an earthquake they relate it to the sense of "feeling" such as, *Just felt an earthquake...* Likewise, in case of a storm or floods, tweets are related to the sense of seeing or hearing such as; *I've never seen or heard such a violent thunder/hail/rain storm as the one we've just experienced.*

Table 2
Frequency of different types of eyewitness reports.

Event type	Direct eyewitness	Indirect eyewitness	Vulnerable direct eyewitness
Floods	62	2	84
Earthquakes	354	13	0
Hurricanes	95	16	185

Table 3
Direct eyewitness reports from manual analysis .

No.	Floods direct eyewitness reports
(1)	I almost died driving home from work because it started to downpour and flood on the freeway and lightning and its 99 f**king degrees out
(2)	No one even notified me that this flood in our area has reached almost 3 feet. but atleast i was able to reach home safely.
(3)	Stuck in New Brunswick. High flood waters near Rutgers. Rt 1 south #Avoid
(4)	I just experienced a flash flood. they're intense
	Earthquakes direct eyewitnesses reports
(1)	Most intense earthquake i've experienced in japan so far... that is
(2)	Big midnight earthquake and aftershocks now
(3)	Just felt the house shaking in Tokyo. Been awhile since I felt an earthquake. I hope it wasn't a bad one anywhere on the island.
	Hurricanes direct eyewitness reports
(1)	Please pray for us right now, the winds and rain is heavy and the hurricane hasn't even hit us yet. #hurricaneharvey2017
(2)	This hurricane ain't no joke, the rain and winds are heavy right now. #hurricaneharvey2017
(3)	It's starting to flood in our area (hurricane Harvey) so if I don't respond back within a 7++ days expect for the worse hope we'll be safe
(4)	first time is street is starting to flood and the power went out, hurricane harvey finally hit us

4.2.2. Indirect eyewitness

During our manual analysis, we found several tweets where the author was sharing information from direct witnesses. Most often, they were sharing valuable information received from friends, relatives, and their social circle. Although we found only a small number of tweets from this category in our dataset, they are an interesting and potentially useful category, as they also allow information to cross platforms or communication channels (e.g. someone tweets about a story heard over the phone). Table 4 shows some examples of indirect eyewitness reports taken from the manually analyzed data from all three disaster types.

There were only two messages found in the flood dataset where indirect eyewitnesses were reporting about disasters by referring to their family members, while for earthquakes, one indirect eyewitness was reporting about an ongoing earthquake with emotions of worry for his/her family who were direct witnesses of this earthquake. The second example shows a unique case where an indirect eyewitness is reporting about an earthquake he/she was experiencing live but distantly during a video call with one of their family members. The last example in this section is reporting about the safety of direct witnesses from a relative.

Finally, indirect eyewitness reports on hurricanes were very interesting. The first and second examples are about an indirect eyewitness' hometown conditions mixed with emotions of worry. The third example shows concern of the indirect eyewitnesses about their relatives' property due to the prospective hazard. In the last example, the indirect eyewitness is sharing the direct eyewitness report of his friend.

4.2.3. Vulnerable direct eyewitness

During our manual analysis of sampled tweets we noticed tweets where users were anticipating a disaster and were reporting warnings and alerts they received from local authorities on their cell phones. This type of messages was only found in the floods and hurricanes datasets, probably due to the more predictable nature of those events. These tweets constitute an interesting subgroup of direct eyewitness information, because identifying people at risk is important information for crisis managers to allocate resources effectively. Table 5 shows some examples of vulnerable direct eyewitness reports.

The first message in the floods section reports the personal experience of the author about a flood warning where he relates the

Table 4
Indirect eyewitness reports from manual analysis .

No.	Floods indirect eyewitness reports
(1)	Some days in Thailand has been insane, there has been massive flood on the road to the city (only have image on my dad's phone)
(2)	The hsm school and my uncles house are right behind eachother and they were ruined in the flash flood)):
	Earthquakes indirect eyewitnesses reports
(1)	F*cking hell... my wife and kids are in Tokyo and they're in the middle of an earthquake Jesus Murphy just how crap can one day get?
(2)	Was Facetiming my brother in Tokyo when an earthquake. It wasn't strong but took a long time. Glad that he's ok. #tokyo #earthquake
(3)	Finally able to hear from my uncle and know that he and his daughters are safe, the earthquake did not affect them to much #bless #mexico
	Hurricanes indirect eyewitness reports
(1)	Texas has me going for a spin....my hometown was evacuated for the hurricane then an earthquake in Dallas where my entire family is
(2)	My city is getting a rain storm from the hurricane and hella winds but that's nothing compared to what's going on god i'm so worried
(3)	So this hurricane is heading for my brother and sister-in-law's brand new winery. Hope it doesn't get flooded before https://t.co/VBfAchRplM
(4)	Heard from friends in Houston, Austin and San Antonio. High winds and heavy rain last night. Everyone is safe. #hurricaneharvey

Table 5
Vulnerable direct eyewitness reports from manual analysis.

No.	**Floods vulnerable direct eyewitness reports**
(1)	Flash flood warning yet it's not even raining
(2)	Why am I always napping when a flash flood warning comes on to my phone? #scared
(3)	Those flash flood alerts will kill me one day, they scare the f**k out of me
(4)	Ima throw my phone if I get another flood warning
	Hurricanes vulnerable direct eyewitness reports
(1)	Hurricane Harvey is approaching.Dun dun dun. first hurricane I will experience in Texas in my new home omg I hope my area doesn't flood
(2)	Staying home for the hurricane,hopefully it doesn't flood
(3)	I'm so scared I hope this hurricane don't flood my apartment or my car
(4)	Big hurricane is supposed to hit the area tonight and i live in one of the flood zones...

alert with the current weather situation. The second and third messages depict emotions of fear created because of a hazard warning. On the other hand, in the fourth example the eyewitness is angry because of so many flood warnings. On the same note, in the hurricane section the first, second, and third examples are showing mixed emotions of hope and fear while reporting about an approaching hurricane. In last example, the eyewitness is relating their vulnerability to the intensity of approaching hazard.

In this particular category of messages, we also noticed a mix of different types of emotions written in words (not emojis) such as hate, disgust, fear, anger, and humor.

4.2.4. Non-eyewitness reports

In our dataset, tweets which did not possess any explicit eyewitness characteristics, but possessed non eyewitness characteristics (Doggett & Cantarero, 2016). However, these tweets were nevertheless reporting about disasters and were categorized as non eyewitness reports. These reports were sharing disaster related information primarily from news media sources.

4.2.5. "Don't know" cases and noise

There were a number of tweets where disaster related keywords were used as metaphors, such as *Troll army will then flood social media with press cuttings, naughty headlines, whatsapp distortions to offset growing positive opinion*. Such messages were categorized as noise along with any messages containing disaster-related keywords in URL's instead of text body. Furthermore, there were several tweets which were possibly eyewitness reports but were too ambiguous to classify them as such. We put these tweets also in this category.

4.3. Characteristics of different types of eyewitness reports

This section describes the manual analysis task to extract common characteristics of reports posted by the three types of eyewitnesses identified in the previous tasks. We performed this step on our two data samples. However, we found the same characteristics in both datasets.

Information posting on social media platforms are usually restricted by several constraints, e.g. a length limit of originally 140 (now 280) characters on Twitter. Such constraints force social media users to apply creative ways to shorten their messages while conveying their actual intent. As a consequence, Twitter communications differ from usual daily life communications such as emails, blogs etc. We believe that the identification of characteristics associated with each type of eyewitness report will help (i) differentiate among eyewitness types and (ii) also, more importantly, to build automatic computational methods and systems to automatically identify and categorize eyewitness messages.

Table 6
Direct eyewitness characteristics .

No.	Characteristic	Examples
(1)	Reporting small details of surroundings	window shaking, water in basement
(2)	Words indicating perceptual senses	seeing, hearing, feeling
(3)	Reporting impact of disaster	raining, school canceled, flight delayed
(4)	Words indicating intensity of disaster	intense, strong, dangerous, big
(5)	First person pronouns and adjectives	I, we, me
(6)	Personalized location markers	my office, our area
(7)	Exclamation and question marks	!, ?
(8)	Expletives	wtf, omg, s**t
(9)	Mention of a routine activity	sleeping, watching a movie
(10)	Time indicating words	now, at the moment, just
(11)	Short tweet length	one or two words
(12)	Caution and advice for others	watch out, be careful
(13)	Mention of disaster locations	area and street name, directions

Table 7
Indirect eyewitness characteristics .

No.	Characteristic	Examples
(1)	Mention of locations or people the author knows	mom, dad, hometown
(2)	First person adjective	my, our
(3)	Expressing emotions	thoughts, worry, relief
(4)	Reporting safety, damage, missing	missing, safe

4.3.1. Direct eyewitness characteristics

Table 6 lists all the characteristics we have observed in the direct eyewitness messages in earthquake, hurricane, flood, and wildfire types along with examples. As social media communications are short and to the point, users usually skip writing first person pronouns and adjectives. However, if a message has first person pronouns and adjectives, we observe that it is a strong indication of a direct eyewitness report (Fang et al., 2016).

Moreover, we observed that words related to perceptual senses such as *seeing, hearing, feeling* are also strong indications that a message originates from a direct eyewitness. Likewise, words indicating the intensity of a disaster situation such as intense, heavy, strong are extensively found in eyewitness messages posted during all four types of disasters. We suggest that the presence of intensity words is also a strong signal that the message is from an eyewitness, as a person far from the disaster area cannot describe the intensity of the situation. Eyewitnesses tend to mention more about their personalized locations such as my office, our area than non-eyewitnesses. Among other characteristics that are shared across all disaster types include use of exclamation and question marks and special/swear words like “wtf”, “omg”, and “s**t”. However, on the contrary, the characteristic# 11 i.e., *short tweet length* was only found in the earthquake dataset. Many examples were found where users shared tweets consisting of only one or two words to report an earthquake such as *earthquake*!. One probable cause could be the sudden and unpredictable nature of earthquake events. Furthermore, we noticed *caution and advice* and *mention of precise disaster locations* characteristics specifically in flood, hurricane, and wildfire disasters. One possible reason for this observation can be the relatively predictable and long-term nature of these disaster types.

4.3.2. Indirect eyewitness characteristics

Table 7 shows characteristics learnt from indirect eyewitness messages for all disasters. We observed that indirect eyewitness reports either mention a person or a place the contributing user already knows. The social circle of a user tends to be credible and so the indirect eyewitness is also considered credible. If an indirect eyewitness report is about the hometown of a user, then it is assumed that they know the geography of disaster hit region very well and can provide useful information if required. It was also observed that indirect eyewitness reports were either about emotions of worry or sense of relief. Indirect eyewitness reports were also about damage, safety or missing people/property.

4.3.3. Vulnerable direct eyewitness characteristics

Table 8 shows distinct characteristics of vulnerable direct eyewitness reports. This category was only found in flood, hurricane, and wildfire datasets due to their predictable nature. Characteristics 5 to 9 in Table 6 were common in both categories. Users were mostly found reporting about hazard warnings and associating it with current weather situations. As hazard alerts were often sudden in nature, they provoked different types of emotions due to sudden disruptions in user’s routine activities.

4.4. Crowdsourcing

We performed crowdsourced labeling on our second data sample for primarily two reasons: First, to acquire more training data so we can use machine learning algorithms to automatically classify the reports. Second, to validate our own eyewitness reports taxonomy developed from the manual analysis with our first data sample.

We used the Figure-eight platform to crowdsource the categorization of tweets into the eyewitness reports types identified during the manual analysis. Figure-eight is a paid and well-known crowdsourcing platform. We shared the messages with the crowd workers and asked them to categorize them according to the developed taxonomy. The crowdsourcing platform provides various quality control measures to evaluate the process and its results. The first measure is that the contributors themselves are categorized into three levels. Level one is comprised of all qualified contributors, and it got the fastest throughput. Level two is comprised of a smaller group of more experienced contributors delivering a higher accuracy of results. Level three is the smallest group of most experienced

Table 8
Vulnerable direct eyewitness characteristics.

No.	Characteristic	Examples
(1)	Warnings and alerts about expected disasters	flash flood warnings
(2)	Associating warnings with current weather situation	flash flood alert with rain
(3)	Expressing emotions	hate, disgust, anger, scare

Table 9
Crowdsourcing results for second data sample.

Event type	Direct	Indirect	Vulnerable	Non	Don't know
Floods	320	85	222	551	822
Earthquakes	1557	43	–	200	200
Hurricanes	321	67	77	1199	336
Wildfire	122	44	23	1379	432

contributors. We selected contributors from level two because of budget constraints, as the cost increases according with the level of the contributors. Moreover, the crowdsourcing platform provides additional quality control options such as posing initial test questions to annotators before authorizing them to contribute. If the annotator does not pass a threshold percentage of the test questions, he/she may not complete the job. A limit of minimum time spent on the task (in seconds) can also be set to make it sure that the annotator spent enough time in reading and understanding the task. We set a minimum limit of 80 percent accuracy of test questions which means users with 80 percent accuracy or higher of test questions will qualify for the job. Initially we created eight test questions to initiate a quiz and later we added another 13 test questions suggested by the platform based on trusted judgments. We also set a minimum time of 50 seconds for user to spend on reading and understanding the messages before moving on to the next set of messages. We asked for three judgments for every message to be able to assess the inter-rated agreement. We repeated the same step of identifying eyewitness reports features as in manual analysis from three classes (direct eyewitness, indirect eyewitness, and vulnerable direct eyewitness) from crowdsourced labelled dataset.

Crowdsourcing results are shown in Table 9. To evaluate the accuracy of crowdsourced annotation, we conducted an audit on the results. We select 50 sample messages from each dataset and two authors of the paper annotated them again. This annotation was later compared with the crowd annotation and showed 90% accuracy for floods, 92% for earthquake, 82% for hurricane, and 94% percent for wildfire datasets.

5. Experiments and results

The highly varying volume of Twitter streams makes it almost impossible for human analysts to identify eyewitness reports during peaks, e.g. during an evolving disaster. Therefore, we propose to use a supervised machine learning algorithm to automatically identify eyewitness reports.

5.1. Feature engineering

The characteristics of eyewitness messages identified by humans (described in Section 4.3) are operationalized into features to learn automatic classifiers. We refer to these features as domain-expert features.

For characteristics 2, 4, 8, and 12 from Table 6 and characteristic 1 from Table 8, we developed lists of words with the help of dictionaries and thesauri. For example, *words indication perceptual sense* uses words related to hearing and seeing, *indicating intensity of an event* has terms such as intense, small, dangerous, and big as basis and was then expanded with synonyms, while the expletives were build on Wiktionary[13] enriched by various slangs[14], since people often use various slangs on social media. *Caution and advice* words and their synonyms were searched in the dictionary and thesaurus. For characteristic 9 in Table 6 *mention of daily routines*, a list of daily routine activities[15] was used in the present continuous tense. Other characteristics (3, 10) in Table 6 were operationalized directly from message content: *reporting impact of disaster* and *time indicating words* were generated from direct eyewitness messages.

The remaining characteristics (5, 7, and 11) in Table 6, i.e. *first person pronouns and adjectives, exclamation and question marks*, and *short message length* were straightforward to implement by adding the corresponding terms (I, me, ...), characters (exclamation and question marks), and counting the words in the message.

Personalized location markers in Table 6 was overlapping with characteristic 5 and dropped. Likewise, characteristic 2 in Table 7 is overlapping with characteristic 5 in Table 6. Similarly, characteristic 2 in Table 8 is overlapping with characteristic 3 in Table 7. Characteristic 3 and 4 in Table 7 comprised a list of words extracted from indirect eyewitness reports. Stop words[16] were excluded. For each of the operationalized characteristics, we counted the number of occurrences of matching words (only uni-grams) or, in case of *short message length*, the binary absence or presence.

The first characteristic *reporting small details of surroundings* in Table 6 and the second characteristic *associating warnings with current weather situations* in Table 8 proved too abstract to operationalize and were not implemented. Likewise, the last characteristic *mentions of disaster locations* in Table 6 and first in Table 7 were not used, because tweets are often too short and informal at times and location identification as well as pronouns from social media data is another aspect of extensive research.

In addition to the domain-expert features, the second type of features we use are BoW-based. Specifically, uni-grams and bi-grams

[13] https://en.wiktionary.org/wiki/Category:English_swear_words .
[14] https://www.speakconfidentenglish.com/english-internet-slang/ .
[15] http://www.vocabulary.cl/Lists/Daily_Routines.htm .
[16] http://www.lextek.com/manuals/onix/stopwords1.html .

Table 10
Floods results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.584	0.488	0.532	627
Non-eyewitness	0.706	0.575	0.634	551
Don't know	0.656	0.820	0.729	822
Domain-expert features				
Eyewitness	0.638	0.478	0.547	627
Non-eyewitness	0.642	0.664	0.653	551
Don't know	0.635	0.742	0.685	822
Domain-expert + text features				
Eyewitness	0.717	0.469	0.567	627
Non-eyewitness	0.748	0.653	0.698	551
Don't know	0.648	0.875	**0.745**	822
Domain-expert + text features with class balancing				
Eyewitness	0.760	0.648	**0.699**	815 (+30%)
Non-eyewitness	0.774	0.763	**0.768**	716 (+30%)
Don't know	0.688	0.798	0.739	822

features are extracted from the textual content of tweets and their TF-IDF scores are used to train machine learning models.

5.2. Classification results

Using the methodological steps described in Section 3 and the features described in the previous subsection, we train several machine learning classifiers. We used the labeled data obtained from crowdsourcing. However, the indirect and vulnerable eyewitness classes were small, i.e. having fewer labeled messages. Further, the manual classification showed that even for a human these classes can be difficult to differentiate. For these reasons, we combined all three types of eyewitness classes, i.e. direct, indirect, and vulnerable into one class namely *"Eyewitness"*. Consequently, our classification task consists of three classes: (i) *Eyewitness*, (ii) *Non-eyewitness*, and (iii) *Don't know*.

Tables 10 –13 show the results of our analysis for floods, hurricanes, earthquakes, and wildfires respectively. The last columns of the tables show class distributions. The last three rows present class distribution after applying the class balancing technique (i.e. SMOTE) where the number in parentheses indicates how many artificially labeled instances we added to that class using the SMOTE technique.

The floods results (Table 10) show slightly better performance (e.g. F-scores) when using domain expert features compared to text features (baseline). However, even better results are obtained upon combining both text and domain features. However, the minority classes *eyewitness* and *non-eyewitness* still suffer compared to the *don't know* classes which achieved an F-score of 0.745. To balance the minority classes, we add 30% more instances, which clearly seem to help achieve better results for the both classes.

Table 11
Hurricanes results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.646	0.419	0.508	465
Non-eyewitness	0.773	0.852	0.810	1199
Don't know	0.605	0.679	0.640	336
Domain-expert features				
Eyewitness	0.655	0.546	0.596	465
Non-eyewitness	0.776	0.881	0.825	1199
Don't know	0.645	0.482	0.552	336
Domain-expert + text features				
Eyewitness	0.734	0.503	0.597	465
Non-eyewitness	0.788	0.910	**0.844**	1199
Don't know	0.686	0.604	0.642	336
Domain-expert + text features with class balancing				
Eyewitness	0.816	0.796	**0.806**	930 (+100%)
Non-eyewitness	0.838	0.843	0.841	1199
Don't know	0.801	0.820	**0.810**	672 (+100%)

Table 12
Earthquakes results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.878	0.977	0.925	1600
Non-eyewitness	0.893	0.585	0.707	200
Don't know	0.629	0.280	0.388	200
Domain-expert features				
Eyewitness	0.871	0.969	0.917	1600
Non-eyewitness	0.787	0.645	0.709	200
Don't know	0.333	0.095	0.148	200
Domain-expert + text features				
Eyewitness	0.865	0.987	0.922	1600
Non-eyewitness	0.912	0.620	0.738	200
Don't know	0.641	0.125	0.209	200
Domain-expert + text features with class balancing				
Eyewitness	0.892	0.966	**0.927**	1600
Non-eyewitness	0.932	0.793	**0.857**	400 (+100%)
Don't know	0.801	0.653	**0.719**	400 (+100%)

In the case of hurricanes (Table 11), the domain features seem to give good advantage over the plain text features for the eyewitness class. However, for the other two classes the difference is not significant. Furthermore, better performance was observed upon combining both text and domain features. However, the minority classes seem to suffer again. We added 100% more labeled instances to both minority classes, i.e. eyewitness and don't know, and obtained better performance, which outperforms all three models.

The earthquakes results are shown in Table 12. Surprisingly, in this case the domain features do not seem to help much. In fact, a significant drop is observed in the *don't know* class. On combining domain and text features, the performance seems to increase a bit. These experiments were challenging as there is a big difference in the class distributions. Just to highlight the significance of balanced classes, we added 100% more labeled instances to both minority classes, i.e. eyewitness and don't know, and obtained better results.

Table 13 shows the results of wildfires. We observe a good improvement in the performance when using domain features compared to using text features. However, when domain and text features are combined, the don't know class seems to perform better than the other two classes. To tackle with the class imbalance issue, we increased the labeled instance of the eyewitness class by 100%. Even after the dataset is not balanced, but it starts showing positive indication that more labeled data can achieve better results.

Overall, we observed that in most cases domain features seem to help achieve better performance compared to text-only features. Moreover, a combination of both text and domain features seem to significantly gain the classifiers performance. Specifically, in the case of earthquakes, *tweet-length, magnitude token*, a bi-gram consisting of *earthquake and magnitude, felt* etc. were among the top features. In the case of floods and forestfires, *personal possessive, reporting impact of disasters (e.g., raining, burning), words indicating*

Table 13
Wildfires results for all four variations of our trained models.

Text-based features (baseline)				
Category	**Precision**	**Recall**	**F-score**	**Class Dist.**
Eyewitness	0.649	0.265	0.376	189
Non-eyewitness	0.857	0.941	0.897	1379
Don't know	0.748	0.708	0.728	432
Domain-expert features				
Eyewitness	0.703	0.339	0.457	189
Non-eyewitness	0.863	0.943	0.901	1379
Don't know	0.737	0.688	0.711	432
Domain-expert + text features				
Eyewitness	0.794	0.265	0.397	189
Non-eyewitness	0.867	0.946	**0.905**	1379
Don't know	0.730	0.731	0.731	432
Domain-expert + text features with class balancing				
Eyewitness	0.897	0.714	**0.795**	378 (+100%)
Non-eyewitness	0.753	0.727	0.740	432
Don't know	0.876	0.935	**0.905**	1379

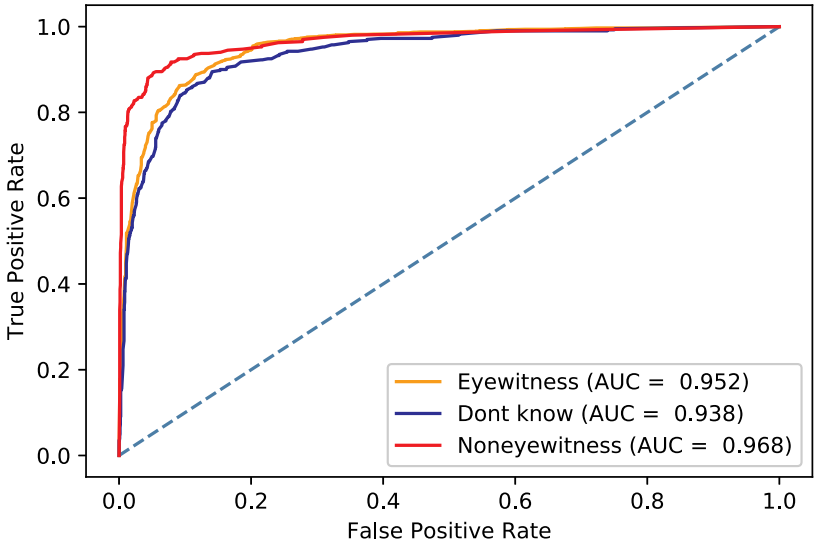


Fig. 1. Earthquake: ROC curves of all three classes of the best model.

intensity (e.g., heavy, intense), mention of locations, flash flood, etc. were among the most useful features. Furthermore, in the case of hurricanes, features including *intensity of disaster, caution and advice for others (e.g., watch out, warning, be careful), time indicating words, perceptual senses*, etc. were identified as useful ones.

Given all of our datasets are hugely imbalanced, we evaluated classifiers performance after adding more labeled data to minority classes. This approach seems to outperform all of our model training variations. To further understand the performance of our classifiers, we draw AUC (Area Under The Curve) ROC (Receiver Operating Characteristics) curves of the best models, which in most of the cases are the models that rely on domain-expert, textual features, and class balancing support. Generally, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the model. Fig. 1 shows AUC-ROC curves of the earthquake model. All classes obtained a reasonable AUC values i.e., & > 0.90&. Fig. 2 shows AUC-ROC curves of the flood model where the under performing class i.e., *eyewitness* can be easily noticed. This highlights the need of more labeled data as well as more distinguishable features among all three classes.

Fig. 3 shows the AUC-ROC curves of the forest fire model. In this particular case, the *eyewitness* class clearly outperforms and achieves an $AUC = 0.968$. The other two classes slight suffer, but still obtained a decent AUC. Fig. 4 shows the AUC-ROC curves of the hurricane model. The *eyewitness* class shows slightly lower performance compared to the other two classes. However, it achieves an $AUC = 0.938$, which is reasonable.

6. Discussion

Given the volatility of social media, our objective was to identify and engineer features for an automated classification that would be identifiable in similar social media platforms. Due to the ready availability of Twitter data, our study relies on tweets like most

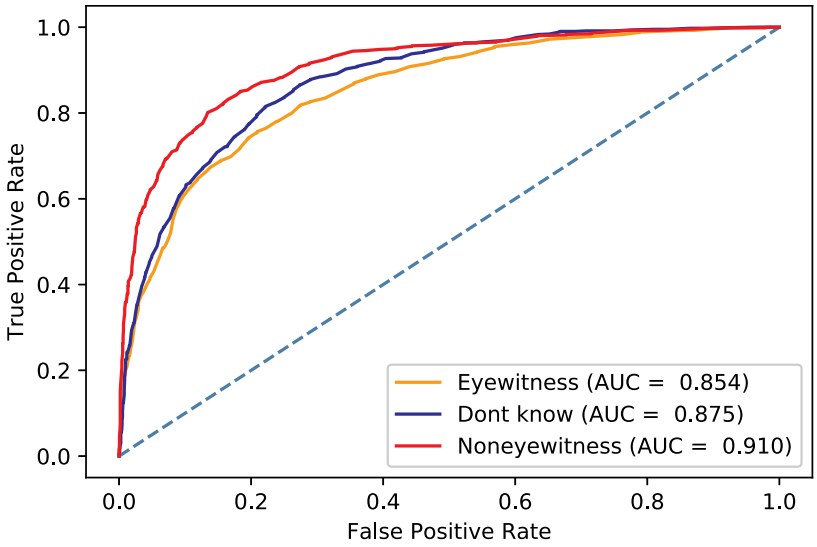


Fig. 2. Flood: ROC curves of all three classes of the best model.

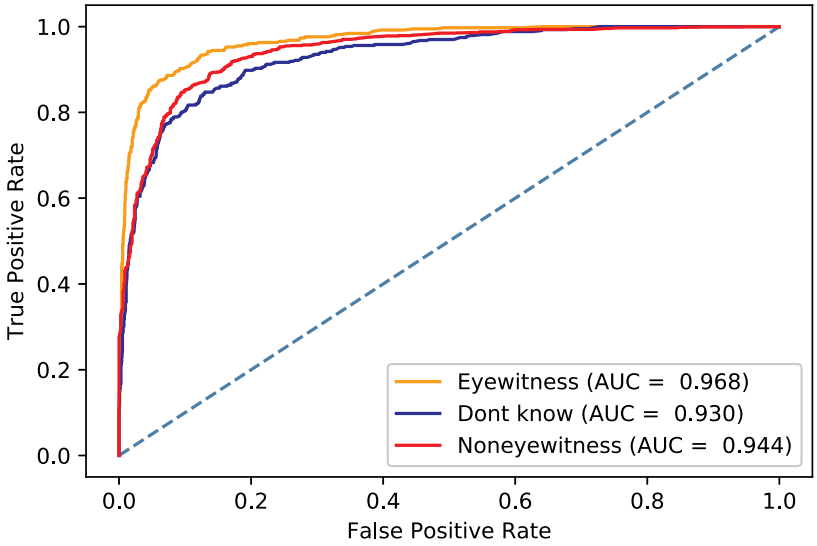


Fig. 3. Forestfire: ROC curves of all three classes of the best model.

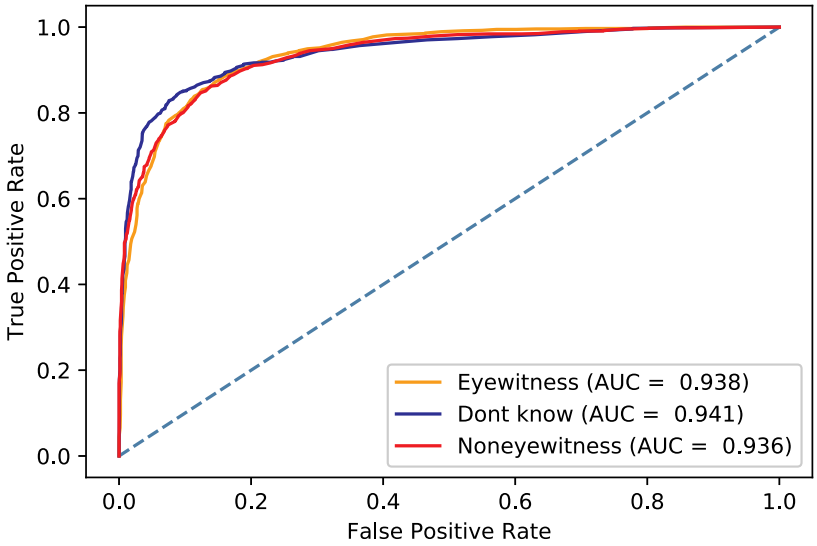


Fig. 4. Hurricane: ROC curves of all three classes of the best model.

other related studies do. However, we focused on extracting features from the text content (the messages), instead of metadata fields or the individual social network characteristics, which might differ structurally from other social media and social networks. Despite our previous observation that Twitter communication differs from other communication channels, we are therefore confident that validation studies using our approach with different social media data sources will be able to replicate our results.

Another important objective was to achieve good classification results for a variety of disaster types. We observed during manual analysis of our data for different disaster types that due to the different nature of disasters, some of the characteristics found in messages were different from each other. For example, because of unpredictable, sudden, and short nature of earthquakes, short tweets consisting of one or two words were only found in earthquake dataset. It was also observed that the frequency of words like *just, now, suddenly* was higher in the earthquake dataset as compared to other disasters which are relatively predictable and long-term in nature. Moreover, words related to *Caution and advice* were found more frequently in the flood, hurricane and wildfire dataset compared to others.

The feasibility to operationalize manual features varies, but most of them are very feasible to learn. However, some features such as *small details of vicinity* and *associating current weather conditions to the disaster* proved to be too abstract to operationalize well. Moreover, while many features can be kept or safely translated automatically after changing the language of the input data, others clearly are language dependent. For example, expletives features differ between languages, and some languages have very different grammar structures, so features depending on personal or possessive pronouns will require adjustment.

One can observe variations in classifiers performance. For instance, most earthquake-related classes scoring the highest F1-measures and floods the lowest, the performance is acceptable for all disaster types, possibly allowing our approach to be used on anthropogenic disasters as well.

A potential source for lower performance is the class imbalance that we had to deal with. We chose to use the SMOTE approach to

increase the availability of minority classes and improve balancing. Realistically, obtaining balanced labeled data from social media during an ongoing event is almost impossible. However, given our promising results obtained from nearly-balanced datasets motivates to put an effort during labeling to obtain balanced classes. During time-critical situations, one simple approach is to restrict adding new labels for a class which has majority while only allowing labels for minority classes. This is particularly suitable for human-in-the-loop systems.

On the positive side, our study provides another, reproducible example of the feasibility of crowdsourced annotation and labeling. Following best practices, we were able to increase our training dataset substantially with a modest investment of funds, while ensuring high quality of results and high inter-rater reliability. This is a further step towards a better integrated human-machine processing approach, with human validation of ”don't know” classifications a potential way to increase recall further.

Another innovative aspect of our research is that we focused on building an eyewitness reports taxonomy exclusive for disaster response agencies during natural disasters. Our taxonomy has three types of eyewitness reports: direct, indirect, and vulnerable eyewitness. Direct eyewitness reports are generated from the people who felt (hear, smell, saw) the disaster or its impacts by themselves. A direct eyewitness report restricts the geographic location of users e.g. reports originating from disaster-hit region. However, in this taxonomy we not only consider reports generated by the people who are present in disaster-hit region but also the reports generating from anywhere outside the disaster location about their family and friends who are present in disaster-hit region known as indirect eyewitness reports. Those people can be a useful resource to give more information about the whereabouts of missing people. Another interesting type of our taxonomy is vulnerable direct eyewitness. In this type we consider reports coming from the people who are anticipating a disaster and who are also present in the region for which disaster warning has been issued. The rationale of adding this type is that such reports can help disaster response agencies to launch precise rescue operations if situation gets worse in that region.

While our manual analysis identified several useful subclasses of eyewitnesses (direct, indirect, vulnerable), early experimentation with training the models showed that performance was low. The two main reasons were the semantic ambiguity of many instances and the even lower number of available instances. The semantic ambiguity made it difficult occasionally even for experienced human annotators to decide based on 140 characters of text whether the source of the message was directly observing or reporting someone else's observations, and whether danger was imminent (vulnerable). Coupled with the low number of example instances, we decided to combine all subclasses for the automated analysis.

One important limitation of our study is language. For practical reasons, we have limited our analysis to English language tweets. Depending on the language structure, some of our features, e.g. first person pronouns, might not work.

7. Conclusions

Finding firsthand and credible information during disasters and emergencies is an important task for relief organizations and law enforcement agencies. The extensive use of social media platforms during disasters provides numerous opportunities for humanitarian organizations to enhance their response. Among them, identification of bystanders and eyewitnesses can help to get important information. In this work, we presented an analysis of tweets collected related to four types of disasters to understand different types of eyewitness reports. Our manual analysis results show that we can categorize eyewitness reports into direct, indirect, and vulnerable direct eyewitnesses. Moreover, an important contribution of this work is to determine various characteristics associated with each type of eyewitness report. We observed that direct eyewitnesses use words related to perceptual senses such as seeing, hearing, feeling. Whereas, indirect eyewitness mainly express emotions such as thoughts, prayers, worry. And, the vulnerable category mostly share warnings and alerts about an expected disaster situation. We use these characteristics and manually labeled data obtained to perform extensive experimentation. Our results revealed that domain-expert features when combined with textual features outperform models which are only trained on text-based features. Moreover, we apply a class balancing technique to tackle the class-imbalance problem, which most of our datasets suffer with. The results obtained after applying class balancing reveal even better results.

Acknowledgement

This research has funded by Swiss government excellence scholarship (ESKAS), Einrichtungskredit, and Forschungskredit grant number K-75130-02 at the University of Zurich.

References

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., & Tao, K. (2012). *Twitcident: Fighting fire with information from social web streams. Proceedings of the 21st international conference on world wide web*. ACM305–308.

Allen, C. (2014). A resource for those preparing for and responding to natural disasters, humanitarian crises, and major healthcare emergencies. *Journal of Evidence-Based Medicine, 7*(4), 234–237.

Amaratunga, C. (2014). Building community disaster resilience through a virtual community of practice (VCOP). *International Journal of Disaster Resilience in the Built Environment, 5*(1), 66–78.

Callison-Burch, C. (2009). *Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. Proceedings of the 2009 conference on empirical methods in natural language processing: Volume 1*. Association for Computational Linguistics286–295.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Diakopoulos, N., De Choudhury, M., & Naaman, M. (2012). *Finding and assessing social media information sources in the context of journalism. Proceedings of the SIGCHI*

conference on human factors in computing systems. ACM2451–2460.
Doggett, E. V., & Cantarero, A. (2016). *Identifying eyewitness news-worthy events on twitter. Conference on empirical methods in natural language processing*7.
Fang, R., Nourbakhsh, A., Liu, X., Shah, S., & Li, Q. (2016). *Witness identification in twitter. Procceedings of the fourth international workshop on natural language processing for social media, Austin, TX, USA*65–73.
Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial Statistics, 1*, 110–120.
Haworth, B., & Bruce, E. (2015). A review of volunteered geographic information for disaster management. *Geography Compass, 9*(5), 237–250.
Hong, L., Dan, O., & Davison, B. D. (2011). *Predicting popular messages in twitter. Proceedings of the 20th international conference companion on world wide web*. ACM57–58.
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.
Imran, M., Castillo, C., Lucas, J., Meier, P., & Vieweg, S. (2014). *AIDR: Artificial intelligence for disaster response. Proceedings of the 23rd international conference on world wide web*. ACM159–162.
Imran, M., Lykourentzou, I., Naudet, Y., & Castillo, C. (2013). Engineering crowdsourced stream processing systems. arXiv preprint arXiv:1310.5463.
Imran, M., Mitra, P., & Srivastava, J. (2016). *Cross-language domain adaptation for classifying crisis-related short messages. Proceedings of the 13th international conference on information systems for crisis response and management (ISCRAM)*.
Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances, 2*(3), e1500779.
Kumar, S., Barbier, G., Abbasi, M. A., & Liu, H. (2011). *Tweettracker: An analysis tool for humanitarian and disaster relief. ICWSM*.
Kumar, S., Morstatter, F., Zafarani, R., & Liu, H. (2013). *Whom should I follow?: Identifying relevant users during crises. Proceedings of the 24th ACM conference on hypertext and social media*. ACM139–147.
Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is twitter, a social network or a news media? Proceedings of the 19th international conference on world wide web*. ACM591–600.
Landwehr, P. M., & Carley, K. M. (2014). *Social media in disaster relief. Data mining and knowledge discovery for big data*. Springer225–257.
Lee, K., Ganti, R. K., Srivatsa, M., & Liu, L. (2014). *When twitter meets foursquare: Tweet location prediction using foursquare. Proceedings of the 11th international conference on mobile and ubiquitous systems: Computing, networking and services*. ICST (Institute for Computer Sciences, Social-Informatics and âQ198–207.
Meier, P. (2012). Crisis mapping in action: How open source software and global volunteer networks are changing the world, one map at a time. *Journal of Map & Geography Libraries, 8*(2), 89–100.
Morstatter, F., Lubold, N., Pon-Barry, H., Pfeffer, J., & Liu, H. (2014). Finding eyewitness tweets during crises. arXiv preprint arXiv:1403.1773.
Oh, O., Agrawal, M., & Rao, H. R. (2013). Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises. *MIS Quarterly, 37*(2).
Olteanu, A., Vieweg, S., & Castillo, C. (2015). *What to expect when the unexpected happens: Social media communications across crises. Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. ACM994–1009.
Ostermann, F., Garcia-Chapeton, G., Kraak, M., & Zurita-Milla, R. (2018). *Towards a crowdsourced supervision of the analysis of user-generated geographic content: Engaging citizens in discovering urban places*.
Ostermann, F., & Spinsanti, L. (2012). Context analysis of volunteered geographic information from social media networks to support disaster management: A case study on forest fires. *International Journal of Information Systems for Crisis Response and Management (IJISCRAM), 4*(4), 16–37.
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).
Schnebele, E., et al. (2013). Improving remote sensing flood assessment using volunteered geographical data. *Natural Hazards and Earth System Sciences, 13*(3), 669.
Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). *Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics254–263.
Takahashi, B., Tandoc, E. C., & Carmichael, C. (2015). Communicating on twitter during a disaster: An analysis of tweets during typhoon Haiyan in the Philippines. *Computers in Human Behavior, 50*, 392–398.
Tanev, H., Zavarella, V., & Steinberger, J. (2017). *Monitoring disaster impact: Detecting micro-events and eyewitness reports in mainstream and social media*.
Teevan, J., Ramage, D., & Morris, M. R. (2011). *# twittersearch: A comparison of microblog search and web search. Proceedings of the fourth ACM international conference on web search and data mining*. ACM35–44.
Thom, D., Krüger, R., Ertl, T., Bechstedt, U., Platz, A., Zisgen, J., & Volland, B. (2015). *Can twitter really save your life? A case study of visual social media analytics for situation awareness. Visualization symposium (PACIFICVIS), 2015 IEEE pacific*. IEEE183–190.
Truelove, M., Vasardani, M., & Winter, S. (2014). *Testing a model of witness accounts in social media. Proceedings of the 8th workshop on geographic information retrieval*. ACM10.
Truelove, M., Vasardani, M., & Winter, S. (2015). Towards credibility of micro-blogs: Characterising witness accounts. *GeoJournal, 80*(3), 339–359.
Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). *Microblogging during two natural hazards events: What twitter may contribute to situational awareness. Proceedings of the SIGCHI conference on human factors in computing systems*. ACM1079–1088.
Xu, B., Guo, X., Ye, Y., & Cheng, J. (2012). An improved random forest classifier for text categorization. *JCP, 7*(12), 2913–2920.
Zahra, K., Imran, M., & Ostermann, F. O. (2018). *Understanding eyewitness reports on twitter during disasters. Proceedings of the 15th international conference on information systems for crisis response and management, Rochester, NY, USA, May 20–23*.
Zahra, K., Ostermann, F. O., & Purves, R. S. (2017). Geographic variability of twitter usage characteristics during disaster events. *Geo-Spatial Information Science, 20*(3), 231–240.

C

PUBLICATION III: TOWARDS AN AUTOMATED INFORMATION EXTRACTION MODEL FROM TWITTER THREADS DURING DISASTERS

Zahra, K., Das, R. D., Ostermann, F. O., and Purves, R. S. Towards an automated information extraction model from Twitter threads during disasters (submitted).

Towards an automated information extraction model from Twitter threads during disasters

Kiran Zahra[a], Rahul Deb Das[ab], Frank O. Ostermann[c], Ross S. Purves[a]

[a]Department of Geography, University of Zurich Switzerland

[b]IBM, Germany

[c]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente

Abstract

Social media plays a vital role as a communication source during large-scale disasters. The unstructured and informal nature of such short individual posts makes it difficult to extract useful information, often due to a lack of additional context. The potential of social media threads – sequences of posts – has not been explored as a source of adding context and more information to the initiating post. In this research, we explored Twitter threads as an information source and developed an information extraction model capable of extracting relevant information from threads posted during disasters. We used a crowdsourcing platform to determine whether a thread adds more information to the initial tweet and defined disaster-related information present in these threads into six themes – event reporting, location, time, intensity, casualty and damage reports, and help calls. For these themes, we created the respective thematic lexicons from WordNet. Moreover, we developed and compared four information extraction models trained on GloVe, word2vec, bag-of-words, and thematic bag-of-words to extract and summarize the most important information from the threads. Our results reveal that 70 % of all threads add more information to the initiating post for various disaster-related themes. Furthermore, the thematic bag-of-words information extraction model outperforms the other algorithms and models for preserving the highest number of disaster-related themes.

Keywords: social media threads, text summarization, disasters, lexicons, information extraction models, word embeddings

1. Introduction

Social media and particularly Twitter has become a very popular source of information for disaster management (Pourebrahim *et al.,* 2019; Hiltz *et al.,* 2020). Much current research extracting information about ongoing events from social media focuses on extracting relevant information at the level of individual tweets (Spence *et al.*, 2015; Zahra, Ostermann and Purves, 2017) – that is each tweet is treated as a small independent packet of information, and classified as relevant or not in isolation. However, the content of individual tweets often lacks useful context (Ritter *et al.*, 2011) – for example, consider the following tweet:

- *I felt it*.

This tweet is a reply to another tweet and, without additional information is impossible to interpret. Now consider the first tweet to which this is a reply:

- *Just felt an earthquake in San Jose...Bed started shaking and door kept rattling. Anyone else?*
- *I felt it*

The first tweet provides essential context to the second, such that it becomes a meaningful piece of information in itself, both telling us that a second person experienced the same earthquake as the first, and adding credibility to the information shared in the first tweet. In this paper, we rely on this observation, analysing tweets not individually, but as part of a dialogue. We assume that if we can identify initial tweets relevant to a disaster, initiating a dialogue between Twitter users (a so-called thread, e.g. Figure 1), then subsequent tweets can be analysed and classified as to whether they provide additional, useful information for analysts.

Figure 1: An example of a Twitter thread where an initial tweet explicitly describes an event and three responses, all confirm having experienced the same event, in one case also in a specific location of LA (the west side)

Disaster response organizations seek a range of information during different phases of the disaster management cycle. For example, practitioners working in a health department are keen to identify casualty and medically related help calls (Aung and Whittaker, 2013). By contrast, policymakers and resource allocators may be interested to know the extent of damage caused by the disaster to decide resource allocation (Kwok *et al.*, 2016). Therefore, it is important to know what types of information are shared in Twitter threads during disasters. Imran *et al.* (2013) developed a disaster-related message ontology based on tweets posted during a disaster. Their informative message classes include caution and advice, casualties and damage, donations, missing and found reports, and links (i.e. URLs) to further information sources. In our research, we focus on extracting different types of information from tweet content rather than exploring external links such as donation calls and URLs. To do this, we develop a new classification scheme that covers possible information themes shared in Twitter threads during a disaster. Chowdhury, Caragea and Caragea (2019) analyse disaster-related tweet content and noted that tweets posted during disasters have a limited and specific vocabulary when compared to tweets discussing general topics. Therefore, we create a set of thematic lexicons to assist in the categorization of disaster themes.

As a disaster unfolds, a particularly important task is the collection of relevant information in a timely way. Twitter produces high volumes of data at a high velocity (Sankaranarayanan *et al.*, 2009) such that it is impossible to manually read, analyse, and extract all the information

provided in Twitter threads in real-time. To overcome this challenge, we used extractive text summarization to extract and summarise important information from Twitter threads (Jain, Bhatia and Thakur, 2017). Our information extraction model used thematic lexicons to build text summaries based on all of the tweets found in a thread. We also compared a lexicon-based model approach with other approaches based on the GloVe and word2vec algorithms. As a baseline model, we used a simple BoW[1] model to quantify improvements in performance through the three approaches with which we experimented. In this research, we address the following three research questions:

RQ1. Do Twitter threads referring to disasters contain relevant information over and above that contained in the initiating tweet, and what themes are discussed in such threads?

RQ2. Can thematic lexicons containing disaster relevant keywords be created and used in information extraction from Twitter threads?

RQ3. How can disaster-related information in Twitter threads be summarized using extractive text summarization?

2. Background

Analysing different aspects of Twitter data, especially as a source of information during various disasters (Kankanamge *et al.*, 2020; Kaigo, 2012) continues to be an important focus of research. However, if classification is at the level of individual tweets, then a large number of potentially relevant tweets carrying useful information will be discarded due to lack of context. One approach to adding context to a tweet, taken by Nazer *et al.* (2016), is to use metadata (e.g. retweets, number of friends and followers) and content (e.g. topic, request specific keywords) as features. Their work focuses on a particular task – emergency dispatch, intending to identify calls for help.

Other authors have explored more generally the different types of informational content found in tweets during disasters. In early work, Ostermann and Spinsanti (2012) generated a list of relevant keywords based around discussions with domain experts on forest fires. Hodas *et al.* (2015) showed that tweet content posted during disasters focussed on announcing the emergency, giving and requesting advice, damage reports, anxiety, etc. They also developed a list of the most and least informative keywords for different types of disasters. Ashktorab *et al.* (2014) analysed the content and identified a range of categories including missing persons, electricity loss, hospital and health infrastructure, death/casualties, etc. They then implemented a classifier using machine-learning algorithms trained on manually annotated data. Similarly, Alam, Ofli and Imran (2018) classified tweets posted during various disasters into a range of categories including cautions, advice, warnings, injured, dead, rescue, volunteering, etc. Huang and Xiao (2015) also analysed the content of tweets posted during hurricanes but categorized them according to various disaster management phases. These included preparedness, plan, evacuation, tips, event tracking, food, casualty, damage, utilities, etc. They also identified a list of keywords associated with each category to aid classifying of individual tweets, however, their list of relevant keywords is based on a single event - Hurricane Sandy.

Olteanu *et al.* (2014) used a lexicon-based approach to automatically identify and filter relevant messages particular to a crisis event. Their methods dynamically update the terms in lexicons

[1] bag-of-words

based on specific crisis event. Chowdhury, Caragea, and Caragea (2020) also developed disaster lexicons from tweets posted during 37 disasters. Their lexicons contain informative terms posted during every disaster that includes event-specific information such as locations, disaster names, names of important persons, organizations, etc.

Twitter data is often too voluminous for manual summary and interpretation even after filtering and classifying. Text summarization is an effective way of reducing the size of a document and preserving key information at the same time. There are two main approaches to generating text summaries using different algorithms: abstractive text summarization (Moawad and Aref, 2012) and extractive text summarization (Ledeneva, Gelbukh and García-Hernández, 2008). The abstractive text summarization is "the task of generating a short and concise summary that captures the salient ideas of the source text" (Liu *et al.*, 2018). This means that the resultant summary may contain new phrases and sentences that are not part of the original text but are suggested by the algorithm to effectively communicate the information. In contrast, the extractive text summaries "produce a set of most significant sentences from a document, exactly as they appear" (Ferreira *et al.*, 2013). Thus, extractive summaries contain only sentences that appear in the source text. Nichols, Mahmud and Drews (2012) used extractive text summarization to extract important information posted in Twitter statuses during sports events. They also concatenated various text summaries to generate an event summary. Since our approach is based on identifying disaster-related information from various tweets, we employ extractive text summarization to generate summaries.

3. Methods

Our workflow contained four main elements, as illustrated in Figure 2:

1. Creating a corpus of 200 Twitter threads referring to earthquakes.
2. Thematic classification of the disaster-related content and crowdsourcing relevance judgements as to whether Twitter threads as whole contained additional information.
3. Development of thematic lexicons based on positively annotated thread content.
4. Comparison of four information extraction models to build extractive summaries of Twitter threads.

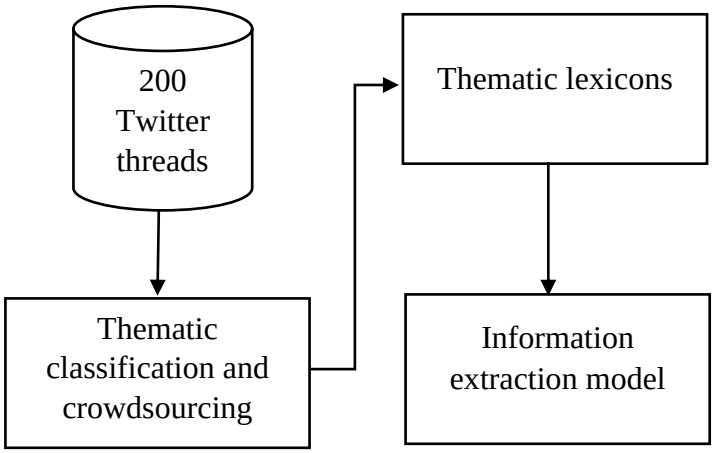


Figure 2: Overall workflow of the methods

3.1 Building a corpus of Twitter threads

The first stage of an experiment using text is to build a corpus. Twitter offers free access to its public tweets via streaming API in real-time. This API can collect individual tweets based on i.e. disaster-related keywords such as *earthquake*, or *flood* or location-based queries i.e. a bounding-box of coordinates. However, the API does not support downloading of threads that reply to an initial tweet. One possible solution to this problem could be scraping the Twitter web pages, however, this would violate Twitter's terms of service. Therefore, to collect threads, we manually searched Twitter[2] for tweets posted during an earthquake. As eyewitness reports and personal observations are considered a credible source of information (Truelove, Vasardani and Winter, 2014), our search strings were also comprised of eyewitness features and disaster-related keywords as described in (Zahra, Imran and Ostermann, 2020) such as "I felt an earthquake" or "I just felt an earthquake" to search for relevant threads. These strings result in a list of matching tweets. We manually analysed each tweet for the following criteria:

- Tweet text must be about an earthquake event.
- Tweet must have at least one reply to form a thread.

We collected the URLs of the first 200 tweets we found meeting the criteria, i.e. the initiating tweets of 200 earthquake-related Twitter threads. Later, two annotators extracted the following information from all tweets in the 200 threads by hand:

- User name
- Time of the tweet
- Tweet content

We assigned a unique conversation and a tweet identifier for every record. To keep the data collection task relatively uncomplicated, we did not analyse nested threads i.e. a thread inside threads. For longer threads, we apply a threshold value of 10 tweets, which means we collected a maximum of 10 tweets in chronological order in threads with more than ten tweets. Table 1 summarises the properties of the 200 threads annotated.

Table 1: Summary values for the 200 threads annotated

Per thread	**Average**	**Median**	**Total count**
Number of tweets	6.9	7	1380
Number of unique users	6.44	7	1288
Length of thread (characters)	409.185	402.5	81837
Length of thread (tokens)	70.62	67	14124

3.2 Thematic classification and crowdsourced tweet thread annotation

Extracting information from thread content related to disasters and relevant to emergency response requires that we define the nature of relevant information. We used six disaster-related themes after conducting general literature research (Imran *et al.*, 2014); (Tapia *et al.*, 2011); (Ashktorab *et al.*, 2014) and discussion with several experts. It is important to note that we aim to extract information only from tweet text – we ignore URLs and links to further information. The themes we identified are event reporting, location, time (subthemes: relative and absolute),

[2] https://twitter.com

intensity, casualty and damage reports, and help calls. Table 2 shows the themes and sub-themes, their definitions, and examples of relevant tweet content.

Table 2: Thematic classification with definitions and examples of relevant content (the example tweet is a fictitious example)

No.	Theme	Definition	Sub-theme	Example
1.	**Event reporting**	Report about the event	-	I just **felt an earthquake** in California at 12:00…shook the whole building. I need help….one building collapsed.
2.	**Location**	The location where the event occurs	-	I just felt an earthquake in **California** at 12:00…shook the whole building. I need help….one building collapsed.
3.	**Time**	The time when the event happened	Absolute time	I just felt an earthquake in California at **12:00**…shook the whole building. I need help….one building collapsed.
			Relative time	I **just** felt an earthquake in California at 12:00…shook the whole building. I need help….one building collapsed.
4.	**Intensity**	The intensity of the event	-	I just felt an earthquake in California at 12:00…**shook the whole building**. I need help….one building collapsed.
5.	**Casualty and damage reports**	Includes reports where people are reporting about casualties and damage caused by the event	-	I just felt an earthquake in California at 12:00…shook the whole building. I need help….**one building collapsed**.
6.	**Help calls**	It includes reports where people are asking for help	-	I just felt an earthquake in California at 12:00…shook the whole building. **I need help**….one building collapsed.

Then we used a crowdsourcing platform Figure Eight[3] (now appen[4]) to assess:

1) Which of these themes are present in the initiating tweet.
2) Whether the thread adds additional information to the initiating tweet.
3) Which themes are present in the thread as a whole.

We asked crowdworkers to read the first tweet and choose which if any of the six disaster-related themes were present in the tweet. Then we presented the crowdworkers with the whole thread and asked whether it added more information to the first tweet. In the case of a positive response, the crowdworker had to choose again which themes were present in the thread. In case of a negative response, crowdworkers were redirected to the next thread. The Figure Eight platform provides quality control features in annotation tasks including the number of annotators assigned to a task, the minimum time spent per judgement, and the percentage of correct answers that Figure Eight provides to train the crowdworkers. Furthermore, crowdworkers can be assigned specific training, selected from specific groups and paid different amounts for a task. In our case we:

[3] In May 2019

[4] https://appen.com/

- Employed only “level two” crowdworkers, a smaller group of more experienced and higher accuracy contributors.
- Paid workers 25 cents per judgement.
- Used eight questions to train the crowdworkers.
- Required a minimum accuracy of at 50 percent for training questions.
- Allocated a minimum time of 50 seconds to spend on reading and understanding each thread.
- Collected three judgements per row to achieve sufficient inter-rater agreement.

3.3 Creating thematic lexicons

In the next step, we developed thematic lexicons for four of the six themes (event reporting, intensity, casualty and damage reports, and help calls) and one sub-theme (relative time) using a two-step process. Figure 3 shows the overall workflow of developing thematic lexicons.

1. **Preparation:** We identified seed words in the threads that were annotated by crowdworkers as containing disaster-related themes for three themes and one sub-theme: event reporting, time (relative), casualty and damage reports, and help calls. We combined the terms found in our dataset with other terms usually used to describe the phenomenon. For the intensity theme, we used the Modified Mercalli intensity scale (Wood and Neumann, 1931) to identify seed words in addition to the terms found in threads.
2. **Development:** We used WordNet (Fellbaum, 2012) a lexical database of semantically related words to search for definitions of seed words. Two authors read the definitions and agreed on relevant ones. Based on the selected definitions, set of synonyms (synsets) were retrieved from WordNet and were used to populate the lexicons.

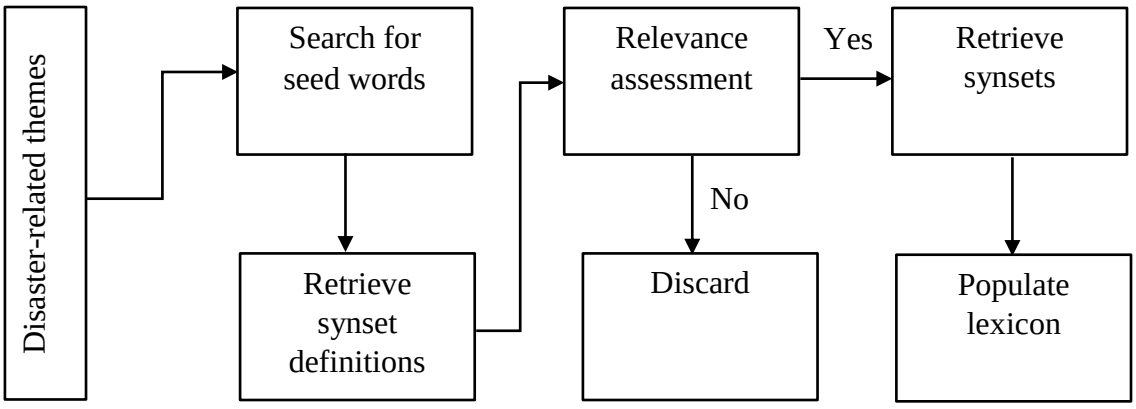


Figure 3: Overall workflow of developing thematic lexicons.

For one theme (location) and one sub-theme (absolute time), we did not prepare lexicons. To identify these themes in our corpus, we used a pre-trained neural network-based NER[5] model implemented in Spacy[6]. This model extracts named entities from unstructured text in the form of personal names, temporal expressions, and location mentions (Nadeau and Sekine, 2007).

[5] Named Entity Recognition

[6] https://spacy.io/

3.4 Information extraction model and text summarization

In the final step, to extract disaster-related information from Twitter threads, we developed four information extraction models using extractive text summarization. The first two models are based on the commonly used word embedding algorithms, e.g., GloVe (Pennington, Socher and Manning, 2014) and a variant of word2vec (Mikolov *et al.*, 2013; Imran, Mitra and Castillo, 2016) to generate word embeddings and determine the context of each word based on semantic similarities. The third model uses a standard BoW approach, which is based on term frequencies and serves as a baseline model in this work. The fourth model is also based on the notion of the BoW model but takes into account the presence of disaster-related terms from the lexicons, and is called the TBoW[7] model.

We performed the following preprocessing steps to prepare our thread corpus for text summary generation using our information extraction models:

- Segmentation of each tweet into individual sentences.
- Removal of special symbols from sentences and then concatenate them to form a corpus.
- Sentence tokenization and stop word removal using stop words lexicons retrieved from the NLTK library in Python (Loper and Bird, 2002).

To develop the model trained on GloVe, we used embeddings based on a corpus of six billion words containing Wikipedia 2014 articles and the Gigaword 5 dataset. For the model trained on word2vec, we used word embeddings published on CrisisNLP[8]. These embeddings are trained on 52 million tweets posted during various disasters (Imran, Mitra and Castillo, 2016). After creating the word embeddings, we generated a cosine similarity matrix. Then we used the similarity matrix to create a graph for each thread, where a node in the graph represents a sentence and the edge between two nodes represents the similarity value. Following that, we used TextRank (a variant of PageRank) algorithm (Mihalcea and Rarau, 2004) applied to the cosine similarity graph to rank the sentences in a given thread.

To develop our baseline BoW model, we computed the term frequency of each token in our text corpus. Then the term frequencies of all tokens present in a sentence are combined to assign an aggregate score which is used to rank the sentences in a given thread.

For the TBoW model, we performed an additional preprocessing step on our text corpus called lemmatization to find normalized forms of words (Plisson, Lavrac and Mladenić, 2004). To develop the model, we retrieved thematic tokens from each sentence by looking up a given token against five thematic lexicons developed in section 3.3 (i.e. event reporting, time (relative), intensity, casualty and damage reports, and help calls). To retrieve spatial and temporal (absolute) thematic tokens, we used a pre-trained NER model leveraging a shallow neural network. To boost the performance of the NER, particularly for retrieving the location entities, we also used several spatial rules such as location names being proper nouns or common nouns appearing after spatial prepositions, e.g., at, near, or to (Das and Purves, 2020). In doing so, we also extracted location entities with vernacular geographical aspect, which are usually detected by a pre-trained model. Furthermore, we iteratively assigned weights to the disaster-related terms related to six themes. Thus, the TBoW model is essentially a modification of the BoW model where we assign more weight to the thematic terms in an adaptive manner.

[7] thematic bag-of-words

[8] https://crisisnlp.qcri.org/

For every token in a sentence, we assigned its term frequency as its respective weight. If the token is also thematic (related to the six themes), then the weight is assigned by relative thematic magnitude. Here we assume that during a disaster, spatial and temporal aspects, as well as help calls, are more important for disaster responders. Therefore, we assigned a higher weight value of 10 to terms in the three thematic lexica of location, time (absolute, relative), and help calls, compared to the other three themes of event reporting, intensity, and casualty and damage reports to which we assigned a weight value of five.

Non-thematic tokens receive a weight value of one, assuming they are not critical information during a disaster. In many cases, a thematic token can appear more than once in a thread, which may indicate a repetitive or higher emphasis on the given aspect. In this case, we consider all the tokens (even repetitive ones) and assign respective weights. Each weight is then normalized using the maximum weight in a given sentence. Following this, we combined all thematic entities in a sentence and computed an aggregate score for the given sentence. Since the TBoW is based on frequencies, sentences with more thematic tokens in a given thread receive a higher score.

To generate text summaries from all four models, we selected the 30 percent of highest scoring sentences from each thread according to the respective models. As we select full sentences to generate the summary, to round off the value of 30 percent, we took a final value which is the smallest integer value that is bigger than or equal to 30. Table 3 summarizes the feature types and different aspects of all four information extraction models. Whereas, Figure 4 shows the process of TBoW information extraction model for a single thread.

Table 3: Feature types and different aspects of summarization models

Feature types and other aspects	**GloVe**	**Word2vec**	**Bag of Words (BoW)**	**Thematic Bag of Words (TBoW)**
Features	Word embedding	Word embedding	Term frequency of all non-stop word tokens	Relative weights assigned to the thematic tokens
Dimension of word vector	100	300	1	1
Model to generate word vector	Co-occurrence matrix	Feed-forward neural network (skip-gram)	Rule-based	Rule-based
Data set used for training word embedding model	A corpus size of 6B tokens from Wikipedia 2014 and Gigaword 5	A corpus size of 52M tweet messages	No training	No training
Sentence scoring and ranking	TextRank algorithm	TextRank algorithm	Sentences are scored based on aggregate weights of all tokens and ranked in descending order	Sentences are scored based on adaptive aggregate weights of the tokens and ranked in descending order

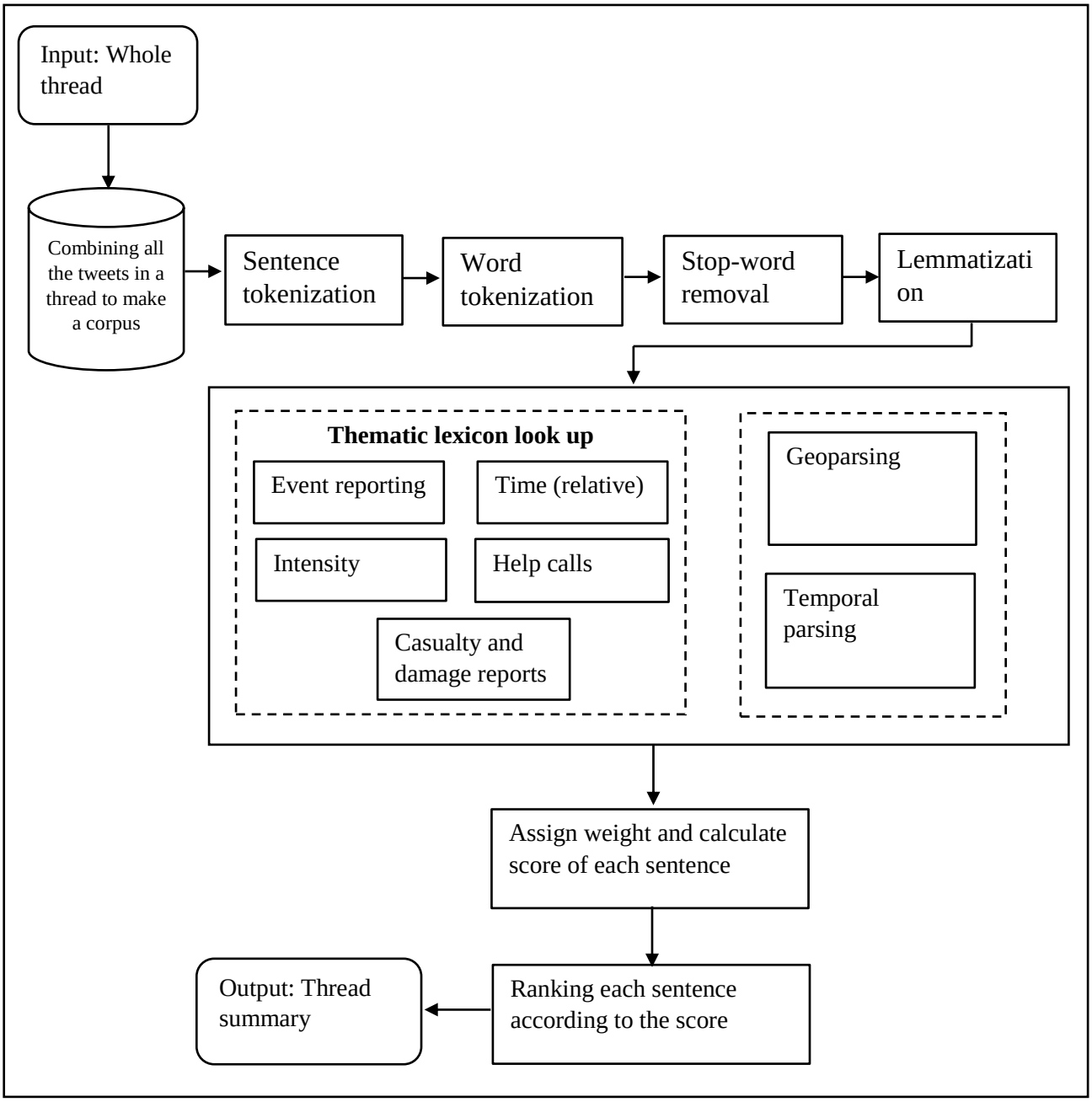


Figure 4: Overall workflow of TBoW information extraction model for a single thread

4. Results

4.1 Crowdsourced assessment of thread information potential

Our results reveal that 70.5% of threads add information to that already contained in the initial tweet. Furthermore, we compared the number of themes present in the initial tweet with the number of themes present in the whole thread (Figure 5). The themes of event reporting, location, and time are mentioned more often in the initial tweet than the rest of the thread. By contrast, the intensity theme was mentioned more often in the remainder of threads than the initial tweet. For the themes of casualty and damage reports, and help calls, overall very few instances are present in our data: only one casualty and damage report is found in the threads, and two help calls are found in the first tweet as well as in the rest of the thread. One possible explanation for these low numbers is that none of the earthquake events in our dataset caused

mass destruction or casualties. Therefore, we observe relatively high numbers of mentions or event reports, location, time, and intensity, but few others.

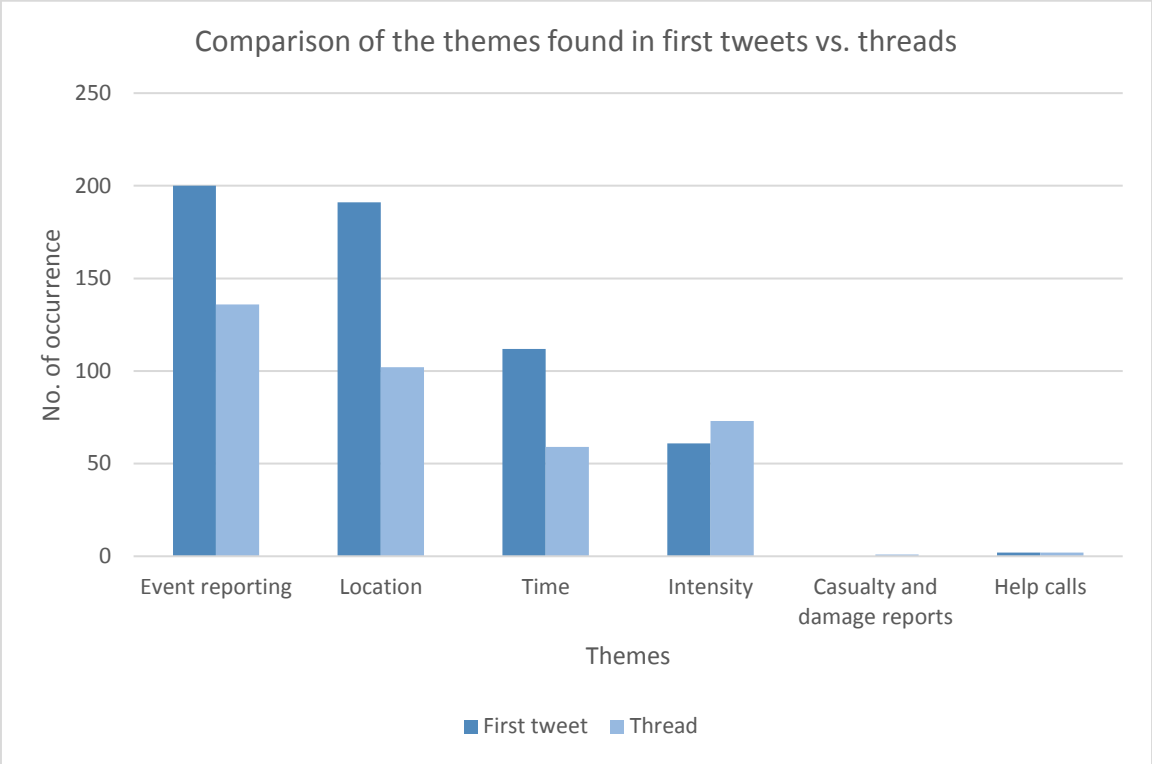


Figure 5: Comparison of the number of themes found in the first tweet and the thread

4.2 Thematic lexicons

Table 4 summarises the characteristics of the thematic lexicons we created. For the event reporting theme, we chose four seed words generally describing an earthquake event. These seed words retrieved 61 initial synset definitions out of which 21 definitions were selected to retrieve further 61 synsets. Therefore, the total number of words in the event reporting theme lexicon is 65. For time (relative) theme, we choose 11 seed words that retrieved 20 initial synset definitions. Only seven synset definitions were found relevant and therefore, selected to retrieve further 18 synsets. The time (relative) theme lexicon thus consists of 29 words in total.

For intensity, we used 56 seed words to describe the various levels of intensity of an earthquake event, for which 371 initial synset definitions were retrieved. Of these, we selected 105 relevant synset definitions that retrieved 393 synsets leading to the 449 words for the intensity theme lexicon. Similarly, for casualty and damage reports theme, nine seed words were selected that retrieved 227 initial synset definitions. We chose 80 relevant synset definitions that retrieved 225 synsets. Therefore, the total number of words in casualty and damage reports lexicon is 234. Finally, for help call themes, four seed words were selected that retrieved 81 initial synset definitions. We selected only 21 relevant synset definitions and retrieved 101 synsets with 105 total number of words in help calls theme.

Table 4: Characteristics of thematic lexicons

Theme	Seed words	Initial synset definitions	Selected synset definitions	Retrieved synsets	Total number of terms in the lexicon
Event reporting	4	61	21	61	65
Time (relative)	11	20	7	18	29
Intensity	56	371	105	393	449
Casualty and damage reports	9	227	80	225	234
Help calls	4	81	21	101	105

Figure 6 shows the word cloud of all thematic lexicons with the words found in our threads dataset[9]. For event reporting theme (figure 6a), *felt* is the most frequently used word with 427 occurrences followed by *earthquake* with 377 occurrences.

For time (relative) theme (figure 6b) *just* is the most frequently used term with 264 occurrences. This result supports previous work (Zahra, Imran and Ostermann, 2020) where we suggested that the presence of term *just* is a strong indication of a personal observation of an earthquake event. Moreover, compared to other natural disasters, for earthquakes social media users tend to report their observations immediately, therefore, use of such temporal markers is common.

For the intensity theme (figure 6c), *good* is the most frequently used term with 38 occurrences followed by various instances of some obvious terms such as *big, strong, small,* etc. By exploring individual tweets we found that users frequently use expressions such as "*felt a good jolt"* to report an earthquake event.

For the relatively rare casualty and damage theme (figure 6d), irrelevant terms such as *last* (17 times) and *go* (15 times) were the most frequent (due to their high frequency in language). However, we also found the terms such as *damage* (11 times), *dead* (5 times), and *fall* (3 times) in our dataset.

For terms related to help calls (figure 6e), words like *stay* (39 times), *get* (26 times), and *take* (23 times) occurred most frequently followed by some obvious terms such as *need* (5 times), and *help* (4 times). Although an actual help calls occurred only once (figure 5) the word *help* occurred a few more times in other contexts, for example, "*Stay safe and keep in touch. Let us know if you need help with anything.*".

[9] The complete lexicons developed in this research are available on GitHub at: https://github.com/rddspatial/text-summarization

Figure 6: Lexicon words that occur in Twitter threads – the size of the word corresponds to the number of times it occurred in the data

4.3 Evaluation of the extracted text summaries

To evaluate which model performs best in preserving maximum information in the thread summary, we selected 50 random threads to compare summaries generated by the four approaches we took. We evaluated the summaries on the presence of disaster-related words from the lexicons. Our analysis is based on context and semantics, i.e. the simple presence of a term is not sufficient. The meaningful presence of a word belonging to one of the six themes increases that model's evaluation score by 1. Two authors of the paper performed this evaluation to measure inter-rater agreement. We then ranked the summaries according to descending scores, i.e. the highest (best performing) model would achieve rank 1, and the model that performs worst (with the lowest score) ranks 4. Our results reveal that the TBoW model has the highest average rank of 1.6 followed by word2vec model with an average rank of 1.7. The GloVe model has an average rank of 2.2, and the baseline BoW model has the lowest average rank of 2.6.

Figure 7 shows the frequency of the themes present in all summaries. The comparison between the models shows that the TBoW model extracts the highest number of event reporting, location, intensity, and casualty and damage reports themes. For the time theme, word2vec model outperforms the rest.

Table 5 shows an example of the text summaries generated by each of the four information extraction models for one thread. In this particular case, the BoW summary achieves the score of five with two instances of event reporting (e.g. one positive and one negative report), two locations (e.g. Santa Monica and Simi), and one relative timestamp (e.g. just). The GloVe summary preserves more information with a total score of six where three instances are found for event reporting (e.g. two negatives and one positive reports) and three are locations (e.g. Santa Monica, Simi, Hollywood). The word2vec-trained model preserves the lowest amount of information with a score of four where two instances are event reports (e.g. two negative reports) and two are locations (e.g. Simi, Santa Monica). Finally, the TBoW summary that preserves the highest number of thematic instances with a total score of nine with four instances of event reports (e.g. two positives and two negative reports), four locations (e.g. Santa Monica 2x, Simi, Hollywood), and one relative timestamp (e.g. just). The complete set of results generated in the form of text summaries for each model are available online.[10]

[10] https://github.com/rddspatial/text-summarization

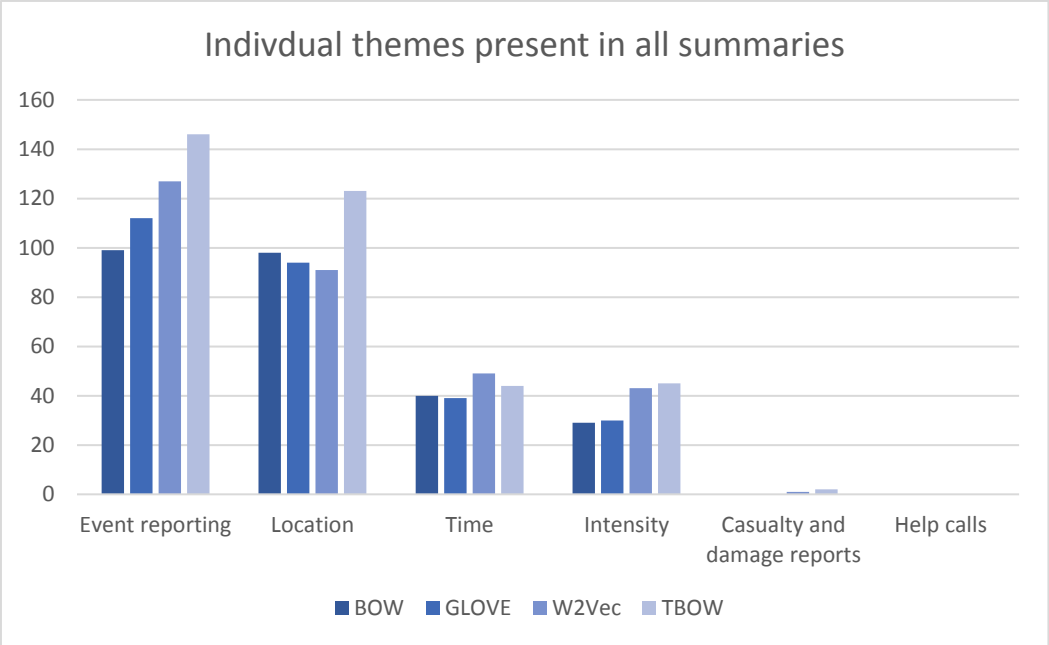


Figure 7: Total number of themes present in all information extraction models

Table 5: Example of the extractive text summaries for one Twitter thread

Model	**Summary**	**Score**
BoW	Just felt an #Earthquake here in #SantaMonica Didn't feel a thing in Simi - not that far away. That sounded far less corny in my head. It's just the earth celebrating the new year a few days late.	5
GloVe	Didn't feel a thing in Santa Monica. Didn't feel a thing in Simi - not that far away. But, i don't have wine. You're in California ... you have lots of earthquakes. I felt it in #hollywood. Going to find flashlights and shoes just in case... Hope all is well.	6
Word2Vec	Didn't feel a thing in Simi - not that far away. That sounded far less corny in my head. It's just the earth celebrating the new year a few days late. Hope all is well. Please keep us updated. i got nothin. didn't feel a thing in Santa Monica.	4
TBoW	Just felt an #Earthquake here in #SantaMonica.Didn't feel a thing in Simi - not that far away..no kidding! i got nothin..I felt it in #hollywood..It's just the earth celebrating the new year a few days late. That sounded far less corny in my head..You're in California ... you have lots of earthquakes..didn't feel a thing in Santa Monica.	9

5. Concluding discussion

Our aim in this research was to explore the potential of social media threads as a source of adding context and additional information to individual posts shared during disaster events. We tested our hypothesis using Twitter threads. We used crowdsourced micro-tasking to determine whether a thread adds information to the content shared in the initial post. For this purpose, we defined *information* shared during disasters into six disaster-related themes. After collecting crowdworkers judgements, we extracted seed terms for thematic lexicons from threads positively annotated for the presence of disaster-related themes. The seed words coupled with WordNet synsets were then used to populate thematic lexicons. In the final step, the thematic lexicons were used to extract and summarize disaster-related information from Twitter threads and results were compared with other word embedding and BoW models.

Our first research question explores the content of Twitter threads for two types of information. First, we want to assess whether a thread contains additional information to the content shared in the first tweet, and second, what type of information is present in the threads. To achieve this objective, we define the term *information* into various disaster-related themes. The crowdworkers annotate a thread positive if any of the themes are present in the thread. The results are compared in Figure 5. We observed that the theme 'event reporting' was most frequently present in the threads. From a disaster responder's perspective, this might not be an actionable piece of information, however, from an information processing perspective, the confirmation of an event in a thread adds credibility to the shared information in a tweet that can be an important source of assessing the credibility of shared information during disasters. That is why, while evaluating the summaries (section 4.3), we counted positive as well as negative event reports every time they appear in the summary.

The second most frequently occurring theme, location, is a critical piece of information for disaster responders as knowing the precise location of a disaster event or a help call is crucial to emergency response. When we further analysed the location theme, we observed that in many cases users not only share precise geographic locations but also various locations depending on the extent of the event in the thread. For example, the following is the snippet from a thread showing how users share geographic locations during a disaster:

Tweeter 1: Up feeding the baby and just felt an earthquake. In East Tennessee?!

Tweeter 2: I just felt it in ATL

Tweeter 3: Felt it here too! Glad to know I'm not crazy!

Tweeter 4: In Knoxville and felt it!

Tweeter 5: I thought I must've been dreaming, but something woke me up, and I thought it was the whole house shaking.

Tweeter 6: Just felt here in S. Riane County

Tweeter 7: Yup me too in Oak Ridge

Tweeter 8: Looked it up too see if i was crazy, im glad im not alone.

Tweeter 9: Felt it in Georgia too

Tweeter 10: Yes!! I felt it shake. It woke me up! Just outside Knoxville!

This example shows that threads can potentially be a rich source of location information which would be otherwise very difficult to extract from single posts especially when Twitter has removed its conventional geotagging feature[11]since 2019 that further limits the possibility of collecting precise location from individual posts. However, we did not further explore and report location theme in this research as toponym matching using a gazetteer is another research strand that is out of the scope of this study.

The third and the fourth most frequently occurring themes were intensity and time respectively. In the event of an earthquake, people usually describe its strength with commonly used terms such as "the whole building is shaking" or "heard the bang". For the time theme, relative timestamps such as "just" and "now" were more frequently used compared to absolute time stamps. In case of a different disaster type, intensity lexicon might not fully capture the information because of different terms used to describe the phenomenon.

For the final set of themes i.e. casualty and damage reports and help calls, almost no instances were found. This phenomenon can be explained in two ways: first, the events reported in the threads were mild and did not cause any damage and casualties. As a result, there were no help calls. Second, people did not use this platform to report such events. However, the results reported in (Mihunov *et al.*, 2020) state that 75% of the respondents in their survey stated that they find social media – Twitter easier to use for disseminating help calls than traditional sources.

The second research question explores the potential of using thematic lexicons to extract information from Twitter threads. Social media threads contain more text than single posts, and some are of considerable length. Therefore, it is important to extract only relevant information from the threads. To achieve this objective, we developed four information extraction models using an extractive text summarization technique. Two models were trained on word embeddings i.e. GloVe and word2vec. The third model BoW (a baseline model), was developed on a bag-of-words approach that used frequency of terms to determine the most important information from the thread. Besides, the fourth model TBoW was developed using disaster-related lexicons.

The rationale behind developing such a lexicon-based approach is twofold. First, it is simple and fast to implement during a disaster as compared to word embedding models that require a comparatively big (disaster relevant) dataset to train the model and create the word embeddings that can capture the contexts and semantics. This is time-consuming. Second, depending on the idiosyncrasies in the training corpus, it does not guarantee that the word embedding will capture the context of the disaster-related terms, which are critical to emergency response operations in case if the word embedding models are trained on a general text – GloVe. As the word2vec embeddings used in this research are trained on tweets posted during natural disasters, the information extraction model trained on word2vec also shows promising results. The summaries generated by TBoW model however earned the highest score. This elaborates the effectiveness of our methods that are based on a lexicon-based approach but produce high-quality results outperforming word embedding algorithms. The TBoW information extraction model can easily be adapted for other types of natural disasters such as hurricanes, floods, forest fires etc. by only modifying a new event reporting and intensity lexicons for each disaster.

[11] https://twitter.com/TwitterSupport/status/1141039841993355264

The third research question analyses how disaster-related information in a thread can be summarized using extractive text summarization. We generated text summaries for every thread using all four models and evaluated the results based on the meaningful presence of disaster-related terms. We further analysed the sample of 50 summaries used to evaluate the extractive text summarization (section 4.3). To do so, we compared the number of words in all summaries generated by four models with the number of words in their respective full threads. The analysis revealed that all four models substantially reduced the number of words with an average of 39, 43, 44 and 44 for BoW, GloVe, word2vec, and TBoW respectively compared to an average of 75 in full threads. The highest number of words in TBoW summaries also support our results that TBoW model preserves the maximum information present in threads.

However, we also observed a few outliers. In one of such examples, the full thread contains 123 words and summaries generated by BoW, GloVe, and word2vec contain 86, 77, and 81 words. Whereas, TBoW summary contained the full thread with 123 words by only shuffling the sentences. This means that in this particular case, extractive text summarization did not serve the purpose of condensing the amount of text to preserve the information. This limitation can be addressed by using the abstractive text summarization approach to shorten the length of the extractive summaries while preserving as much information as possible.

We also observed the presence of several irrelevant terms particularly for casualty and damage reports and help calls thematic lexicons developed in this research (section 4.2). Although while developing these lexicons, we filtered irrelevant synset definitions, however, a relevant synset definition does not guarantee a completely relevant set of terms. Nevertheless, these irrelevant terms were not frequently present in the data and therefore did not affect the results.

Despite the limitations of this work and social media data in general, we conclude that social media threads are a useful source to get context and additional information about various aspects of a disaster as compared to a single post. This information can help reduce disaster risk by increasing situational updates that can improve the allocation of resources for various disaster relief operations. The methodology of our research is reproducible and replicable as well with other social media platforms such as Facebook.

Acknowledgements

The authors of this research would like to acknowledge two students at the Department of Geography, University of Zurich **Vanessa Reiser** and **Reetta Vaelimaeki** for their help in collecting the data for this research. We also acknowledge project funding from the Swiss National Science Foundation (200021E-166788).

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6. References

Alam, F., Ofli, F. and Imran, M. (2018) CrisisMMD: Multimodal twitter datasets from natural disasters. In: *12th International AAAI Conference on Web and Social Media, ICWSM*. 2018 pp. 465–473.

Ashktorab, Z., Brown, C., Nandi, M. and Culotta, A. (2014) Tweedr: Mining Twitter to Inform Disaster Response. *ISCRAM 2014 Conference Proceedings - 11th International Conference on Information Systems for Crisis Response and Management*. [Online] (May), 841–842. Available from: doi: 10.1145/1835449.1835643.

Aung, E., and Whittaker, M. (2013). Preparing routine health information systems for immediate health responses to disasters. *Health policy and planning*, 28(5), 495-507.

Chowdhury, J., Caragea, C. and Caragea, D. (2020) On Identifying Hashtags in Disaster Twitter Data. *Proceedings of the AAAI Conference on Artificial Intelligence*. [Online] 34 (01), 498–506. Available from: doi:10.1609/aaai.v34i01.5387.

Chowdhury, J.R., Caragea, C. and Caragea, D. (2019) Keyphrase extraction from disaster-related tweets. In: *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. [Online]. 2019 pp. 1555–1566. Available from: doi:10.1145/3308558.3313696 [Accessed: 21 September 2020].

Das, RD. and Purves, RS. (2020) Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India, IEEE *Transactions on Intelligent Transportation Systems*, vol. 21, no. 12, pp. 5213-5222, doi: 10.1109/TITS.2019.2950782.

Fellbaum, C. (2012) WordNet. *The Encyclopedia of Applied Linguistics*. [Online] Available from: doi:10.1002/9781405198431.wbeal1285.

Ferreira, R., De Souza Cabral, L., Lins, R.D., Pereira E Silva, G., et al. (2013) Assessing sentence scoring techniques for extractive text summarization. *Expert Systems with Applications*. [Online] 40 (14), 5755–5764. Available from: doi:10.1016/j.eswa.2013.04.023.

Gupta, A., Lamba, H., Kumaraguru, P. and Joshi, A. (2013) Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In: *Proceedings of the 22nd …*. [Online]. 2013 pp. 729–736. Available from: doi:10.1145/2487788.2488033.

Hiltz, S. R., Hughes, A. L., Imran, M., Plotnick, L., Power, R., and Turoff, M. (2020). Exploring the usefulness and feasibility of software requirements for social media use in emergency management. *International journal of disaster risk reduction*, 42, 101367.

Hodas, N.O., Ver Steeg, G., Harrison, J., Chikkagoudar, S., et al. (2015) Disentangling the lexicons of disaster response in twitter. In: *WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web*. [Online]. 2015 pp. 1201–1204. Available from: doi:10.1145/2740908.2741728 [Accessed: 30 September 2020].

Hu, M., Liu, S., Wei, F., Wu, Y., et al. (2012) Breaking News on Twitter. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012*. 2012 pp. 2751–2754.

Huang, Q. and Xiao, Y. (2015) Geographic Situational Awareness : Mining Tweets for Disaster. *Geo-Information*. [Online] 41549–1568. Available from: doi:10.3390/ijgi4031549.

Imran, M., Castillo, C., Lucas, J., Meier, P., et al. (2014) AIDR: Artificial intelligence for disaster response. In: *Proceedings of the companion publication of the 23rd international conference on World wide web companion*. [Online]. 2014 pp. 159–162. Available from: doi:10.1145/2567948.2577034.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., et al. (2013) Extracting information nuggets from disaster- Related messages in social media. In: *ISCRAM 2013 Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management*. 2013 pp. 791–801.

Imran, M., Mitra, P. and Castillo, C. (2016) Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. [Online] 1638–1643. Available from: http://arxiv.org/abs/1605.05894.

Jain, A., Bhatia, D. and Thakur, M.K. (2017) Extractive Text Summarization Using Word Vector Embedding. *Proceedings - 2017 International Conference on Machine Learning and Data Science, MLDS 2017*. [Online] 51–55. Available from: doi:10.1109/MLDS.2017.12.

Kaigo, M. (2012) Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake. *Keio Communication Review*. [Online] (34), 19–35. Available from: http://www.mediacom.keio.ac.jp/publication/pdf2012/KCR34_02KAIGO.pdf.

Kankanamge, N., Yigitcanlar, T., Goonetilleke, A., and Kamruzzaman, M. (2020). Determining disaster severity through social media analysis: Testing the methodology with South East Queensland Flood tweets. *International journal of disaster risk reduction*, 42, 101360.

Kwok, A. H., Doyle, E. E., Becker, J., Johnston, D., and Paton, D. (2016). What is 'social resilience'? Perspectives of disaster researchers, emergency management practitioners, and policymakers in New Zealand. *International Journal of Disaster Risk Reduction*, 19, 197-211.

Ledeneva, Y., Gelbukh, A. and García-Hernández, R.A. (2008) Terms derived from frequent sequences for extractive text summarization. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [Online]. 2008 pp. 593–604. Available from: doi:10.1007/978-3-540-78135-6_51 [Accessed: 24 February 2020].

Liu, L., Lu, Y., Yang, M., Qu, Q., et al. (2018) Generative adversarial network for abstractive text summarization. In: *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. [Online]. 2018 pp. 8109–8110. Available from: www.aaai.org [Accessed: 21 October 2020].

Loper, E. and Bird, S. (2002) NLTK: The Natural Language Toolkit. In: *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*. [Online]. 2002 p. Available from: doi:10.3115/1118108.1118117.

Mihalcea, R. and Rarau, P. (2004) TextRank: Bringing Order into Texts. In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. [Online]. 2004 p. Available from: doi:10.1016/0305-0491(73)90144-2.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J (2013) Distributed representation of words and phrases and their compositionality, Advances in Neural Information Processing Systems (NIPS'13).

Moawad, I.F. and Aref, M. (2012) Semantic graph reduction approach for abstractive Text Summarization. *Proceedings - ICCES 2012: 2012 International Conference on Computer Engineering and Systems*. [Online] 132–138. Available from: doi:10.1109/ICCES.2012.6408498.

Mihunov, V. V., Lam, N. S., Zou, L., Wang, Z., & Wang, K. (2020). Use of Twitter in disaster rescue: lessons learned from Hurricane Harvey. *International Journal of Digital Earth*, 1-13.

Nadeau, D. and Sekine, S. (2007) A Survey of named entity recognition and classificatio. *Lingvisticae Investigationes*. [Online] 30 (1), 469–510. Available from: doi:10.1162/COLI_a_00178 [Accessed: 29 September 2020].

Nazer, T.H., Morstatter, F., Dani, H. and Liu, H. (2016) Finding Requests in Social Media for Disaster Relief. In: *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 2016 pp. 2–5.

Nichols, J., Mahmud, J. and Drews, C. (2012) Summarizing sporting events using Twitter. In: *International Conference on Intelligent User Interfaces, Proceedings IUI*. [Online]. 2012 pp. 189–198. Available from: doi:10.1145/2166966.2166999 [Accessed: 29 September 2020].

Olteanu, A., Castillo, C., Diaz, F. and Vieweg, S. (2014) CrisisLex: A lexicon for collecting and filtering Microblogged communications in crises. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*. 376–385.

Ostermann, Frank, and Laura Spinsanti. "Context analysis of volunteered geographic information from social media networks to support disaster management: A case study on forest fires." International Journal of Information Systems for Crisis Response and Management (IJISCRAM) 4.4 (2012): 16-37.

Paulussen, S., Harder, R.A., Paulussen, S. and Harder, R.A. (2014) Social Media References in Newspapers Facebook , Twitter and YouTube as sources in newspaper journalism. *Journalism Practice*. [Online] 8 (5), 542–551. Available from: doi:10.1080/17512786.2014.894327.

Pennington, J., Socher, R. and Manning, C.D. (2014) GloVe: Global Vectors for Word Representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [Online]. 2014 pp. 1532–1543. Available from: http://nlp. [Accessed: 24 February 2020].

Plisson, J., Lavrac, N. and Mladenić, D.D. (2004) A rule based approach to word lemmatization. In: *Proceedings of the 7th International Multiconference Information Society (IS'04)*. [Online]. 2004 pp. 83–86. Available from: http://eprints.pascal-network.org/archive/00000715/ [Accessed: 6 November 2020].

Pourebrahim, N., Sultana, S., Edwards, J., Gochanour, A., and Mohanty, S. (2019). Understanding communication dynamics on Twitter during natural disasters: A case study of Hurricane Sandy. *International journal of disaster risk reduction*, 37, 101176.

Purohit, H., Castillo, C., Imran, M. and Pandey, R. (2018) Social-EOC : Serviceability Model to Rank Social Media Requests for Emergency Operation Centers. In: *International*

conference on advances in social network analysis and mining. 2018 p.

Ritter, A., Sam, C., Mausam and Etzioni, O. (2011) Named entity recognition in tweets: An experimental study. In: *EMNLP 2011 - Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2011 Association for Computational Linguistics. pp. 1524–1534.

Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D., et al. (2009) TwitterStand: News in tweets. In: *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems*. [Online]. 2009 pp. 42–51. Available from: doi:10.1145/1653771.1653781.

Spence, P.R., Lachlan, K.A., Lin, X. and del Greco, M. (2015) Variability in Twitter Content Across the Stages of a Natural Disaster: Implications for Crisis Communication. *Communication Quarterly*. [Online] 63 (2), 171–186. Available from: doi:10.1080/01463373.2015.1012219.

Tapia, A.H., Bajpai, K., Jansen, B.J. and Yen, J. (2011) Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations. In: *8th International Conference on Information Systems for Crisis Response and Management: From Early-Warning Systems to Preparedness and Training, ISCRAM 2011*. 2011 p.

Truelove, M., Vasardani, M. & Winter, S. (2014) Testing a model of witness accounts in social media. In: *Proceedings of the 8th Workshop on Geographic Information Retrieval - GIR '14*. [Online]. 2014 pp. 1–8. Available from: doi:10.1145/2675354.2675699.

Vieweg, S., Hughes, A.L., Starbird, K. and Palen, L. (2010) Microblogging during two natural hazards events. In: *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. [Online]. 2010 p. 1079. Available from: doi:10.1145/1753326.1753486.

Wood, H.O. and Neumann, F. (1931) Modified Mercalli intensity scale of 1931. *Bulletin of the Seismological Society of America*. 21 (4), 277–283.

Zahra, K., Imran, M. and Ostermann, F.O. (2020) Automatic identification of eyewitness messages on twitter during disasters. *Information Processing and Management*. [Online] 57 (1). Available from: doi:10.1016/j.ipm.2019.102107.

Zahra, K., Ostermann, F.O. & Purves, R.S. (2017) Geographic variability of Twitter usage characteristics during disaster events. *Geo-Spatial Information Science*. [Online] 20 (3), 231–240. Available from: doi:10.1080/10095020.2017.1371903