Towards an Automated Information Extraction Model from Twitter Threads during Disasters

Kiran Zahra *
University of Zurich †
kiran.zahra@geo.uzh.ch

Rahul Deb Das
University of Zurich, IBM Germany
das.rahuld@gmail.com

Frank O. Ostermann
University of Twente
f.o.ostermann@utwente.nl

Ross S. Purves
University of Zurich
ross.purves@geo.uzh.ch

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 critical information from the threads. Our results reveal that 70 percent of all threads add 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

INTRODUCTION

Social media, particularly Twitter, has become a prevalent source of information for disaster management (Pourebrahim et al. 2019), (Hiltz et al. 2020). The 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, F. O. Ostermann, et al. 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 proper 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, it 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?*

*Corresponding author
†

- *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. This paper relies on this observation, analyzing 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 analyzed and classified as to whether they provide additional, helpful 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 essential to know what types of information are shared in Twitter threads during disasters. Imran, Elbassuoni, 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. Ray Chowdhury et al. 2019 analyzed disaster-related tweet content and noted that tweets posted during disasters have a limited and specific vocabulary compared to tweets discussing general topics. Therefore, we also created a set of thematic lexicons to assist in the categorization of disaster themes.

As a disaster unfolds, a crucial 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 et al. 2017). We developed an information extraction model that used thematic lexicons to build text summaries based on all of the tweets found in a thread. We also compared our lexicon-based model with other models based on the GloVe and word2vec algorithms. As a baseline model, we used a simple bag-of-words (BoW) model to quantify improvements in performance through the three approaches with which we experimented.

The main objectives of this research are to:

- Develop a disaster-related thematic taxonomy capable of classifying important types of information (such as event reports, time, intensity etc.) shared on social media posts during disasters.
- Analyse if a thread adds more information to what is shared in the initiating post using crowdsourcing approach.
- Create thematic lexicons for disaster-related themes.
- Develop an automated information extraction model capable of extracting disaster related information in the form of a summary from social media threads.

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. Anderson et al. 2019 developed an infrastructure using Twitter conversations to add context and location to analyse the useful content.

Other authors have explored more generally the different types of informational content found in tweets during disasters. In early work, F. 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 focused 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 et al. 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 static list of keywords associated with each category to aid classifying of individual tweets specific to Hurricane Sandy incident. However, none of the work categorizes disaster-related information themes solely based on the content and independent of a specific regional disaster event.

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 based on specific crisis event. Chowdhury et al. 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 et al. 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. Nafi et al. 2020 used abstractive text summarization technique to summarise disaster-related documents. 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 et al. 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. Moreover, extractive text summarization techniques are considered simpler and state of the art techniques for text summarization (Jadhav and Rajan 2018);(Wu and Hu 2018).

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.

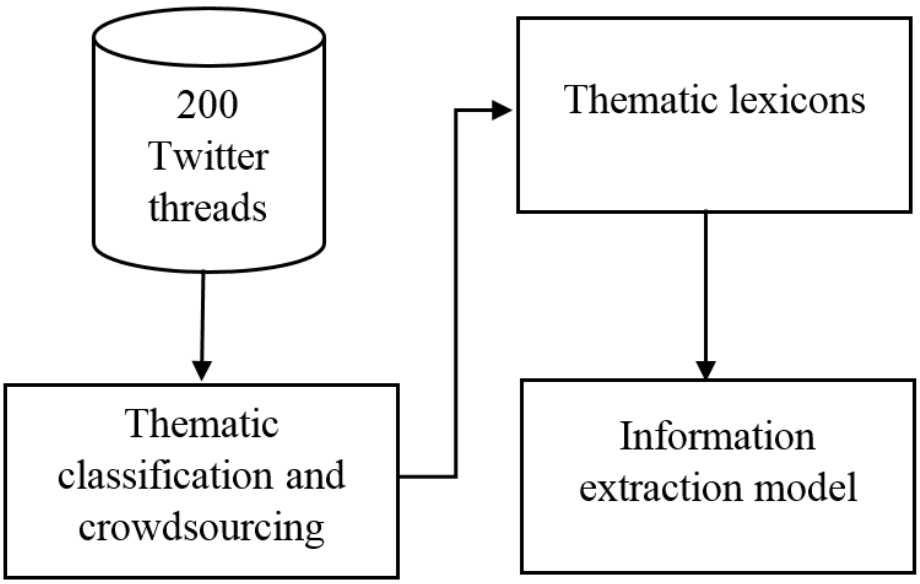

Figure 2. Overall workflow of the methods

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 terms i.e. disaster-related keywords or hashtags 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[1]. Therefore, for the experimental study to collect threads, we manually searched Twitter[2] for tweets posted during an earthquake.

[1] https://twitter.com/en/tos
[2] https://twitter.com

As eyewitness reports and personal observations are considered a credible source of information (Truelove et al. 2014), our search strings were also comprised of eyewitness features and disaster-related keywords as described in (Zahra, Imran, et al. 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: (i) tweet text must be about an earthquake event, (ii) 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 user name, time of the tweet, and tweet content from all tweets in the 200 threads manually. Although our dataset contains very limited amount of threads to testify our claim about the threads adding more information to the initiating tweet. However, it establishes the ground for more research where automated workflow can be developed to download Twitter threads by linking *in reply to status ids* in tweet json file.

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 to keep it comprehensible for human annotators. This means that 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 (for 200 threads)
Number of tweets	6.9	7	1380
Number of unique users	6.4	7	1288
Length of thread (characters)	409.2	402.5	81837
Length of thread (tokens)	70.6	67	14124

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 relevant literature research (Imran, Castillo, et al. 2014); (Tapia et al. 2011); (Ashktorab et al. 2014) and discussion with an emergency management specialist at National Institutes of Health, USA. 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 defined are event reporting, location, time (relative and absolute), intensity, casualty and damage reports, and help calls. Table 2 shows the themes, their definitions, and examples of relevant tweet content.

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

- Which of these themes are present in the initiating tweet.
- Whether the thread adds additional information to the initiating tweet.
- 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:

- Employed only "level two" crowdworkers, a smaller group of more experienced and higher accuracy contributors.

[3] In May 2019
[4] https://appen.com/

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

No.	Theme	Definition	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 (absolute, relative)	The time when the event happened	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.

- 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.

The crowdsourcing platform provides inter-rater agreement of crowdworkers for each judgement in the form of *confidence score*. The average confidence score for all the threads was 0.89 out of 1. That means in most of the cases all three crowdworkers agreed on the same response. The detailed confidence score with each judgement is available on GitHub[5].

Creating thematic lexicons

In the next step, we developed thematic lexicons for five of the six themes (event reporting, time (relative), intensity, casualty and damage reports, and help calls) 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 four themes: 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. All the seed words have also been uploaded on GitHub.

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

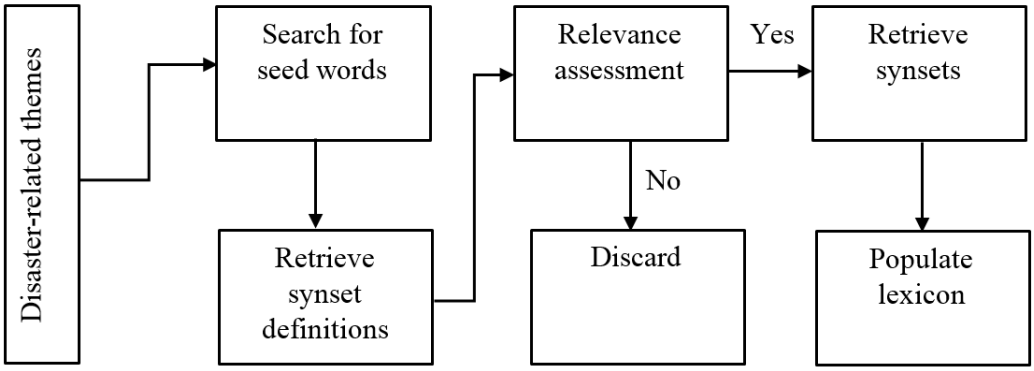

Figure 3. Overall workflow of developing thematic lexicons.

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.

For themes location and time (absolute), we did not prepare lexicons. To identify these themes in our corpus, we used geoparsing and temporal parsing. To do so, we used a pre-trained neural network-based NER[6] model implemented in Spacy[7]. This model extracts named entities from unstructured text in the form of personal names, temporal expressions, and location mentions (Nadeau and Sekine 2007).

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 et al. 2014) and a variant of word2vec (Mikolov et al. 2013); (Imran, Mitra, et al. 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[8] model.

We performed the following pre-processing 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[9]. These embeddings are trained on 52 million tweets posted during various disasters (Imran, Mitra, et al. 2016). After creating the word embeddings, we generated a cosine 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 Tarau 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.

[6]Named Entity Recognition
[7]https://spacy.io/
[8]thematic bag-of-words
[9]https://crisisnlp.qcri.org/

For the TBoW model, we performed an additional preprocessing step on our text corpus called lemmatization to find normalized forms of words (Plisson et al. 2004). To create the model, we retrieved thematic tokens from each sentence by looking up a given token against five thematic lexicons developed in previous section (i.e., event reporting, time (relative), intensity, casualty and damage reports, and help calls). We used a pre-trained NER model to leverage a shallow neural network to retrieve spatial and temporal (absolute) thematic tokens. 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 2019). We also extracted location entities with vernacular geographical aspects, which a pre-trained model usually detects. 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.

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 given by relative thematic magnitude. Here we assume that spatial and temporal aspects and help calls are more critical for disaster responders during a disaster. 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

RESULTS

Crowdsourced assessment of thread information potential

Our results reveal that 70.5 percent 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

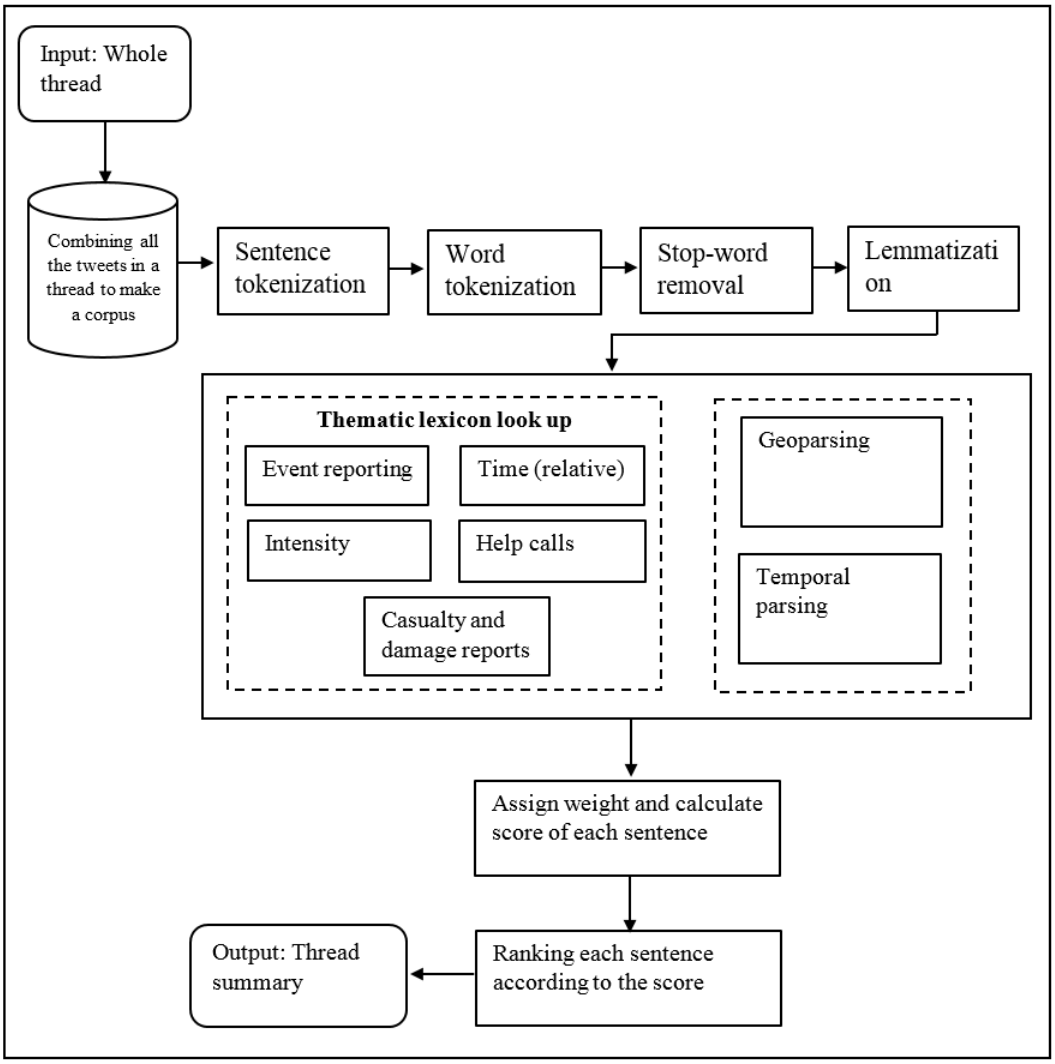

Figure 4. Overall workflow of TBoW information extraction model for a single thread

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.

Thematic lexicons

Table 4 summarises the characteristics of the thematic lexicons. 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.

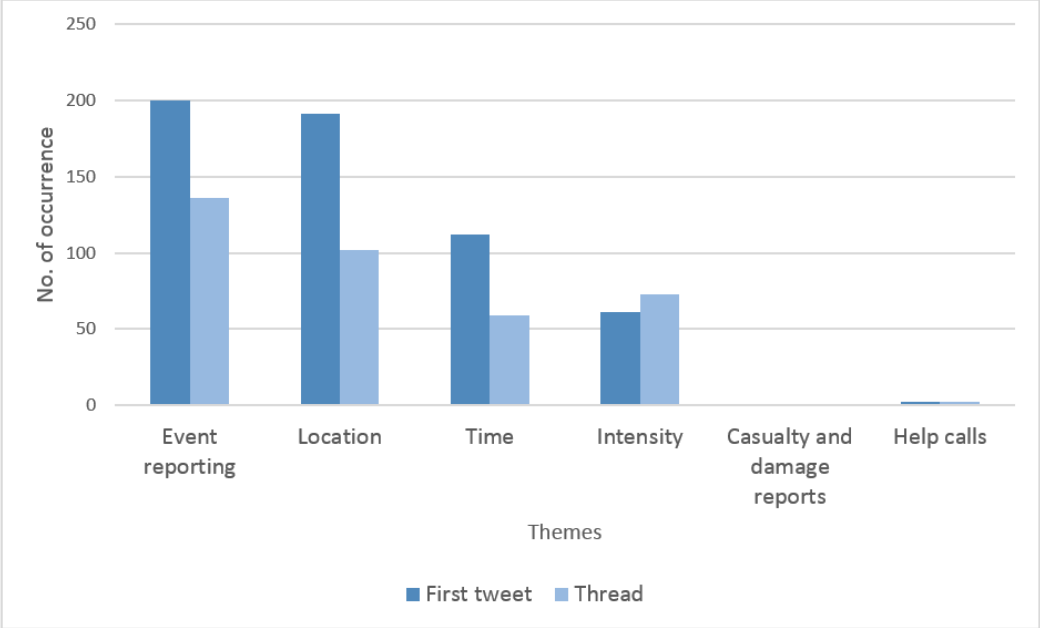

Figure 5. Comparison of the number of themes found in the first tweet and the thread

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[10]. 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, et al. 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

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

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.".

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

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. Whereas, 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. Both the annotators fully agreed on the evaluation score. 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. The initial tweet of the thread is *Just felt an Earthquake here in SantaMonica*. 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[11].

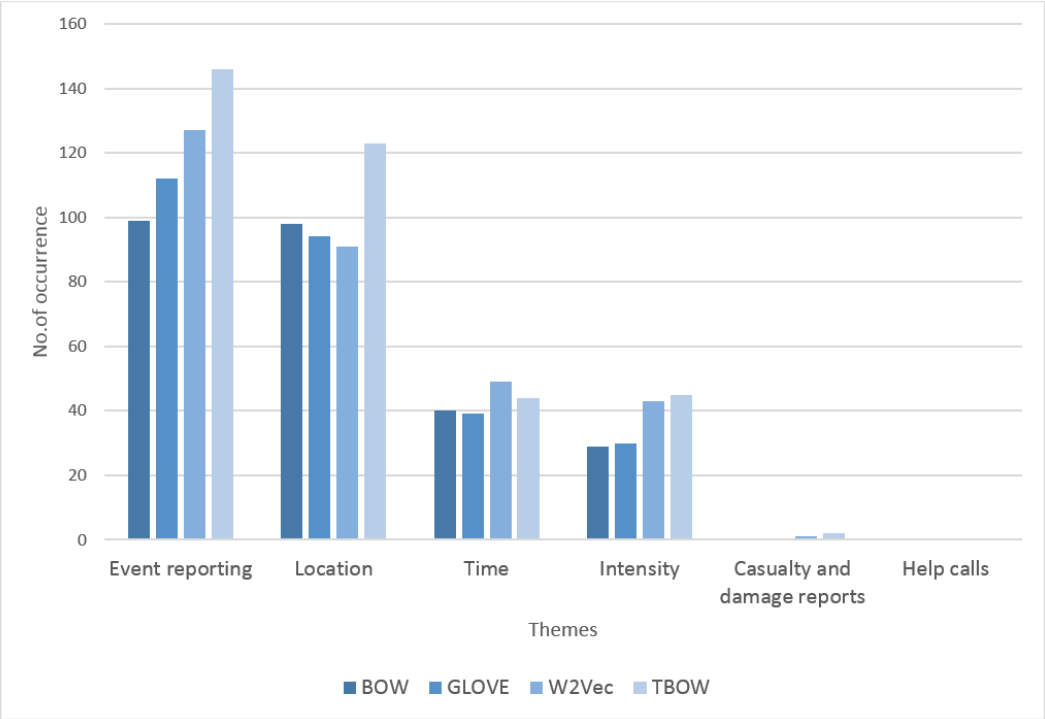

Figure 7. Total number of themes present in all information extraction models

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,

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

Table 5. Example of an extractive text summary

Model	Summary	Total 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

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, 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[12] since 2019 that further limits the possibility of collecting precise location from individual posts. However, we did

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

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 percent 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.

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. 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. 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.

REFERENCES

Alam, F., Ofli, F., and Imran, M. (2018). "Crisismmd: Multimodal twitter datasets from natural disasters". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 12. 1.

Anderson, J., Casas Saez, G., Anderson, K., Palen, L., and Morss, R. (2019). "Incorporating context and location into social media analysis: A scalable, cloud-based approach for more powerful data science". In.

Ashktorab, Z., Brown, C., Nandi, M., and Culotta, A. (2014). "Tweedr: Mining twitter to inform disaster response." In: *ISCRAM*. Citeseer, pp. 269–272.

Aung, E. and Whittaker, M. (2013). "Preparing routine health information systems for immediate health responses to disasters". In: *Health policy and planning* 28.5, pp. 495–507.

Chowdhury, J. R., Caragea, C., and Caragea, D. (2020). "On identifying hashtags in disaster twitter data". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01, pp. 498–506.

Das, R. D. and Purves, R. S. (2019). "Exploring the Potential of Twitter to Understand Traffic Events and Their Locations in Greater Mumbai, India". In: *IEEE Transactions on Intelligent Transportation Systems* 21.12, pp. 5213–5222.

Fellbaum, C. (2012). *Wordnet. the encyclopedia of applied linguistics*.

Ferreira, R., Souza Cabral, L. de, Lins, R. D., Silva, G. P. e, Freitas, F., Cavalcanti, G. D., Lima, R., Simske, S. J., and Favaro, L. (2013). "Assessing sentence scoring techniques for extractive text summarization". In: *Expert systems with applications* 40.14, pp. 5755–5764.

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". In: *International journal of disaster risk reduction* 42, p. 101367.

Hodas, N. O., Ver Steeg, G., Harrison, J., Chikkagoudar, S., Bell, E., and Corley, C. D. (2015). "Disentangling the lexicons of disaster response in twitter". In: *Proceedings of the 24th International Conference on World Wide Web*, pp. 1201–1204.

Huang, Q. and Xiao, Y. (2015). "Geographic situational awareness: mining tweets for disaster preparedness, emergency response, impact, and recovery". In: *ISPRS International Journal of Geo-Information* 4.3, pp. 1549–1568.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. (2014). "AIDR: Artificial intelligence for disaster response". In: *Proceedings of the 23rd international conference on world wide web*, pp. 159–162.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P. (2013). "Extracting information nuggets from disaster-related messages in social media." In: *Iscram*.

Imran, M., Mitra, P., and Castillo, C. (2016). "Twitter as a lifeline: Human-annotated twitter corpora for NLP of crisis-related messages". In: *arXiv preprint arXiv:1605.05894*.

Jadhav, A. and Rajan, V. (2018). "Extractive summarization with swap-net: Sentences and words from alternating pointer networks". In: *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 142–151.

Jain, A., Bhatia, D., and Thakur, M. K. (2017). "Extractive text summarization using word vector embedding". In: *2017 International Conference on machine learning and data science (MLDS)*. IEEE, pp. 51–55.

Kaigo, M. (2012). "Social media usage during disasters and social capital: Twitter and the Great East Japan earthquake". In: *Keio Communication Review* 34.1, pp. 19–35.

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". In: *International journal of disaster risk reduction* 42, p. 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". In: *International Journal of Disaster Risk Reduction* 19, pp. 197–211.

Ledeneva, Y., Gelbukh, A., and Garcıa-Hernández, R. A. (2008). "Terms derived from frequent sequences for extractive text summarization". In: *International conference on intelligent text processing and computational linguistics*. Springer, pp. 593–604.

Liu, L., Lu, Y., Yang, M., Qu, Q., Zhu, J., and Li, H. (2018). "Generative adversarial network for abstractive text summarization". In: *Thirty-second AAAI conference on artificial intelligence*.

Loper, E. and Bird, S. (2002). "Nltk: The natural language toolkit". In: *arXiv preprint cs/0205028*.

Mihalcea, R. and Tarau, P. (2004). "TextRank: Bringing order into texts In Proceedings of EMNLP". In.

Mihunov, V. V., Lam, N. S., Zou, L., Wang, Z., and Wang, K. (2020). "Use of Twitter in disaster rescue: lessons learned from Hurricane Harvey". In: *International Journal of Digital Earth* 13.12, pp. 1454–1466.

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

Moawad, I. F. and Aref, M. (2012). "Semantic graph reduction approach for abstractive Text Summarization". In: *2012 Seventh International Conference on Computer Engineering & Systems (ICCES)*. IEEE, pp. 132–138.

Nadeau, D. and Sekine, S. (2007). "A survey of named entity recognition and classification". In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Nafi, N. M., Bose, A., Khanal, S., Caragea, D., and Hsu, W. H. (2020). "Abstractive Text Summarization of Disaster-Related Document". In: *ISCRAM 2020 Conference Proceedings–17th International Conference on Information Systems for Crisis Response and Management*.

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)*. IEEE, pp. 1410–1413.

Nichols, J., Mahmud, J., and Drews, C. (2012). "Summarizing sporting events using twitter". In: *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, pp. 189–198.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (2014). "Crisislex: A lexicon for collecting and filtering microblogged communications in crises". In: *Eighth international AAAI conference on weblogs and social media*.

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

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)*, pp. 1532–1543.

Plisson, J., Lavrac, N., Mladenic, D., et al. (2004). "A rule based approach to word lemmatization". In: *Proceedings of IS*. Vol. 3, pp. 83–86.

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". In: *International journal of disaster risk reduction* 37, p. 101176.

Ray Chowdhury, J., Caragea, C., and Caragea, D. (2019). "Keyphrase extraction from disaster-related tweets". In: *The world wide web conference*, pp. 1555–1566.

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

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). "Twitterstand: news in tweets". In: *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pp. 42–51.

Spence, P. R., Lachlan, K. A., Lin, X., and Greco, M. del (2015). "Variability in Twitter content across the stages of a natural disaster: Implications for crisis communication". In: *Communication Quarterly* 63.2, pp. 171–186.

Tapia, A. H., Bajpai, K., Jansen, B. J., Yen, J., and Giles, L. (2011). "Seeking the trustworthy tweet: Can microblogged data fit the information needs of disaster response and humanitarian relief organizations." In: *ISCRAM*.

Truelove, M., Vasardani, M., and Winter, S. (2014). "Testing a model of witness accounts in social media". In: *Proceedings of the 8th workshop on geographic information retrieval*, pp. 1–8.

Wood, H. O. and Neumann, F. (1931). "Modified Mercalli intensity scale of 1931". In: *Bulletin of the Seismological Society of America* 21.4, pp. 277–283.

Wu, Y. and Hu, B. (2018). "Learning to extract coherent summary via deep reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1.

Zahra, K., Imran, M., and Ostermann, F. O. (2020). "Automatic identification of eyewitness messages on twitter during disasters". In: *Information processing & management* 57.1, p. 102107.

Zahra, K., Ostermann, F. O., and Purves, R. S. (2017). "Geographic variability of Twitter usage characteristics during disaster events". In: *Geo-spatial information science* 20.3, pp. 231–240.