On Identifying Disaster-Related Tweets: Matching-based or Learning-based?

Hien To, Sumeet Agrawal, Seon Ho Kim, Cyrus Shahabi
Integrated Media Systems Center, University of Southern California, Los Angeles, USA
{hto,sumeetag,seonkim,shahabi}@usc.edu

***Abstract*—Social media such as tweets are emerging as platforms contributing to situational awareness during disasters. Information shared on Twitter by both affected population (e.g., requesting assistance, warning) and those outside the impact zone (e.g., providing assistance) would help first responders, decision makers, and the public to understand the situation first-hand. Effective use of such information requires timely selection and analysis of tweets that are relevant to a particular disaster. Even though abundant tweets are promising as a data source, it is challenging to automatically identify relevant messages since tweet are short and unstructured, resulting to unsatisfactory classification performance of conventional learning-based approaches. Thus, we propose a simple yet effective algorithm to identify relevant messages based on matching keywords and hashtags, and provide a comparison between matching-based and learning-based approaches. To evaluate the two approaches, we put them into a framework specifically proposed for analyzing disaster-related tweets. Analysis results on eleven datasets with various disaster types show that our technique provides relevant tweets of higher quality and more interpretable results of sentiment analysis tasks when compared to learning approach.**

I. Introduction

Enhancing situational awareness is of great importance for disaster response and recovery. Information shared on social media such as Twitter greatly enhance time-critical situational awareness [34], [9] by not only spreading the news about casualties and damages, rescue efforts and alerts but also providing on-topic information for disaster-affected population and first responders who may benefit from such information. Twitter has been one of the most popular means of communication during disasters. Particularly, geotagged tweets have been used to understand the situation in the affected areas, e.g., analyzing the speed and impact of an earthquake [31] as shown in the news about the Napa Earthquake 2014, *"Six Critically Injured, 120 Treated At Napa Hospital Following 6.0 Earthquake"*.

Understanding tweet messages is challenging since they are short (maximum 144 characters) and informal, especially in disasters when identifying timely relevant information is critical. Thus, there has been a large body of work aiming to identify disaster-related tweets. Existing studies use either learning-based or matching-based approaches to select tweets that are relevant to a particular disaster. Learning-based approach builds a model from a set of labeled tweets and uses the model to predict another set of data (e.g., [36], [24], [26], [23]). One challenge of learning-based approaches is that the accuracy of the trained model highly depends on the quality and size of training dataset. However, training datasets in the existing studies are often small because having large labeled data for training was demanding. Conventional matching-based approach enumerates a set of keywords and hashtags that are relevant to a particular disaster, and searches for the tweets containing those words (e.g., [35], [6], [14], [25], [2]). One drawback of this approach is that the set of manually defined keywords and hashtags may not be all-inclusive, i.e., users often use different unique hashtags for the same event.

In this study, we provide a comparative evaluation of both learning-based and matching-based approach in identifying disaster-related tweets. Furthermore, we propose an improved matching-based technique to better identify tweets that are relevant to a particular disaster type such as earthquake, flood. Our technique enhances conventional matching-based approach by effectively enlarging the number of relevant tweets and improving their quality. The technique is twofold. First, it searches for candidate hashtags in a collection of hashtags by matching a small set of core keywords. The core keywords are predefined for each disaster type while the hashtags can be extracted from a collection of tweets (i.e., tweet corpus). Thereafter, the candidate hashtags are refined by ruling out irrelevant ones through crowdsourcing. Both the refined hashtags and the core keywords are used to match relevant tweets. To evaluate our matching-based technique, we use a complementary state-of-art technique that identifies relevant tweets by learning. Our technique applies a set of standard models including *word2vec* for representing each tweet as an embedding vector, TF-IDF for penalizing high-frequency words, latent semantic indexing for dimension reduction, and logistic regression for classifying tweets into relevant and irrelevant ones. In order to have larger training datasets, we aggregate labeled Twitter data from multiple sources and group the tweets of the same disaster type.

Our experiments evaluate the two techniques by putting them into a proposed framework for analyzing geotagged tweets posted by the affected and unaffected population in disasters. Since there is no real ground truth in aggregated large datasets, the set of tweets selected by both methods is considered as a ground truth, from which their recall scores (the fraction of identified tweets that are relevant) are calculated as evaluation measurements. Experimental results on eleven disasters of three different types (earthquake, flood, wildfire) show that the matching-based technique provides a smaller number of relevant tweets but with higher quality

(measured by the recall score) when compared to the learning-based technique. Our contributions are as follows.

1) We identify the specific challenges of the two techniques in classifying relevant tweets: a large number of unique hashtags used for a particular disaster and a lack of big labeled datasets for training classification models.
2) We propose an improved matching-based technique that significantly increases the number of relevant tweets found when compared to traditional matching approaches (by up to 80%).
3) We create eleven new datasets of geotagged tweets, each corresponds to a disaster occurred in 2014-2015. Our datasets (refined hashtags and tweets in affected/unaffected areas) are published with source code on GitHub[1], and open to other researchers.
4) We evaluate the matching-based technique on the datasets, which produces a set of relevant tweets with a higher quality when compared to the learning-based approach. Consequently, in a particular application of sentiment analysis, the matching-based technique also yields more interpretable results.

The remainder of this paper is organised as follows. Section II discusses the related work, followed by our framework in Section III. Experimental results are presented in Section IV. Finally, we conclude the paper and discuss the future work in Section V.

II. Related Work

Social media such as Twitter and Facebook has been widely regarded as active communication channels during emergency events such as disasters caused by natural hazards. The Federal Emergency Management Agency (FEMA) identifies social media as an essential component of future disaster management [9]. Tweets sent during catastrophic events have been known to contain information that contributes to situational awareness [34], and a recent survey of studies for analyzing social media in disaster response can be found in [13]. Therefore, social media data analytics for disaster response has gained extensive interest from the research community. Existing studies focus on extracting disaster-related information from socially-generated content during natural disasters, from which actionable information can be disseminated to disaster relief workers [15]. More recent studies build classifiers for identifying earthquake-relevant tweets [6], classifying tweets based on informative and uninformative tweets [28], [24], [36], [4]. Furthermore, tweets can be categorized by type [33], [16], [3], [1] (i.e., affected individuals, infrastructure and utilities, donations and volunteer, caution and advice, sympathy and condolence) or by information source [32], [22], [8], [7], [33] (i.e., eyewitness, government, NGOs, business, media).

Analyzing public sentiment during disasters is a popular application of social media for disaster response. Although the problem has been extensively studied in other domains, such as product reviews [27], [19], understanding public sentiment from social media is gradually applied in disasters. However, this problem is challenging as the social media content are often short and unstructured [11], [12]. There has been a growing body of work addressing such challenges [5], [20]. In [5], sentiment classification of twitter messages during Hurricane Sandy was performed and the extracted sentiments were visualized on a geographical map centered around the hurricane. In [20], an entropy-based metric was proposed to model sentiment contained in tweet messages. The extracted sentiments were visualized through a map-based interface to reveal interesting patterns in disasters.

These studies typically used machine learning tools to filter disaster-related tweets. An issue with such approaches is the lack of appropriate interpretation of the results, e.g., why a technique works well on some data (high precision and recall) but do not perform as well in others [14], [2]. In addition, the studies tend to examine one particular disaster and imply that the findings are generalizable to others [10]. However, it is known that information shared on Twitter varies considerably from one crisis to another [17], [25], [26]. We aim to narrow the gap by customizing each component of our framework for a particular disaster.

III. Framework

We propose a five-step framework for processing and analyzing disaster-related tweets as shown in Figure 1. Note that we consider only geo-tagged tweets in this study. For each disaster, spam tweets are removed (step 1). Thereafter, the cleaned geo-tagged tweets are mapped to affected and unaffected regions in the vicinity of each disaster (step 2). For each region, tweets are categorized into relevant and irrelevant ones (step 3), in which relevant tweets are analyzed to identify the popularity of sentiments expressed by users during the disasters (step 4). Finally, spatial and temporal patterns and trends of the mapped sentiments are revealed through visualization (step 5). In the following we detail each phase of the framework.

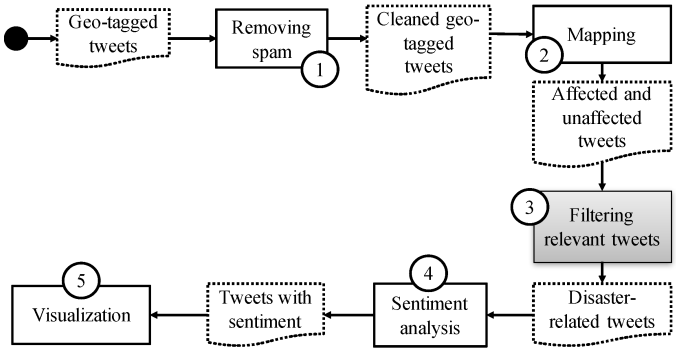


Fig. 1: A framework for analyzing geo-tagged tweets generated during disasters. The focus of this study is highlighted as shaded boxes.

A. Removing spam tweets and mapping tweets into affected and unaffected areas

Many tweets are generated by spammers or non-human bots. Those tweets, if large enough, may falsify the results

[1]https://github.com/infolab-usc/bdr-tweet

of analysis. Therefore, it is important to remove those spam tweets at the first step. We eliminate tweets from a user who generates more than a certain number of tweets (in this study, 15 tweets per day). Note that we do not remove retweets (RTs) as we are interested in the actions of reposting or forwarding existing tweets.

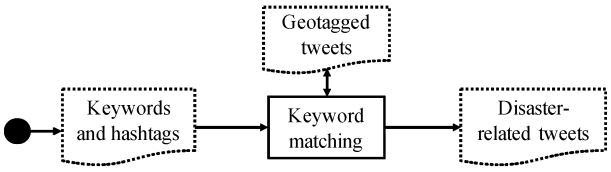
In the second step, we categorise tweets based on geographical regions, i.e., regions affected by a disaster and outside of the impact zone. We are specifically interested in the messages from the affected population as they usually require more and timely attention or assistance. According to FEMA, there are two kinds of assistance: individual assistance (e.g., damage to impacted residences) and public assistance (e.g., repair or replacement of facilities). A declaration for individual or public assistance for counties is requested by the Governor. Affected regions are considered to be the counties that need assistance while unaffected regions are other nearby counties without any assistance. The result is a clean dataset from affected regions.

B. Identifying relevant tweets

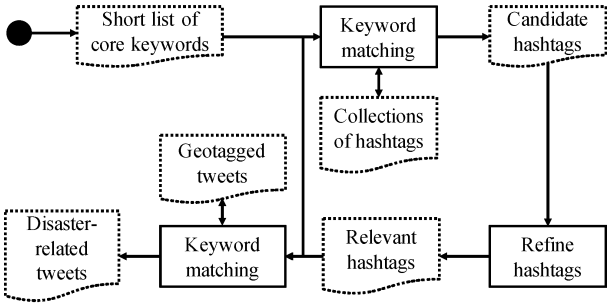
Typically, there are two approaches in identifying tweets that are relevant to a specific disaster: matching-based [6], [14], [25], [2], [16], [15] and learning-based [36], [24], [25], [26], [23], [11], [3].

1) Matching-based approach: The studies in this group typically use a set of keywords or hashtags to determine relevant tweets by identifying them in messages. An issue with the existing studies in this approach is that they typically use a small set of predefined hashtags such as combining disaster name/type with the name of affected area (e.g., *#napaearthquake*) or official name of the disaster (e.g., *#hurricanesandy*). However, such a simple approach may miss many relevant hashtags generated and used by users, for example, *#3amearthquake, #staysafenapa, #fearoftheearthquake*, which are diverse in terms of word choices. Also, many such hashtags are misspelled, such as *#eathquake, #eartquake, #earrhquake*, which cannot be detected by the existing simple solution. The performance of the matching-based approach relies on the completeness of the used keywords and hashtags. Our approach is to systematically construct a complete set of keywords and hashtags.

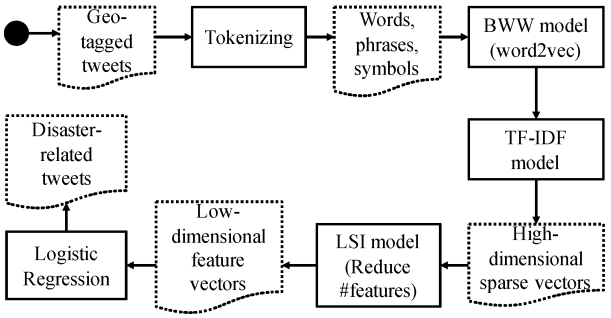
Observing that each disaster type has a small set of core keywords (see column 3 of Table I) like existing studies, we use these core keywords to search for more relevant hashtags on a dictionary of all hashtags, retrieved from the collection of tweets. For each disaster, the dictionary of hashtags can be retrieved by a linear scan through tweet collection to increase the number of disaster-related hashtags. However, simple word matching may include keywords irrelevant to disasters, for example, candidate hashtag *"#fireworks"* contains keyword *"#fire"* but it is unlikely related to wildfires. Such automatic semantic analysis is very challenging and still hard to achieve in practice. Hence, we refine all candidate hashtags to improve the quality of the hashtag collection through crowdsourcing. Crowd reviewers discard relevant hashtags, producing a set of



(a) Conventional matching-based method



(b) Matching-based method



(c) Learning-based method

Fig. 2: Methods to select disaster-relevant tweets.

related hashtags only. The number of candidate hashtags and refined hashtags for each disaster in our study is shown in the last two columns of Table I. On the other hand, the refined hashtags may not cover all the keywords corresponding to a particular disaster type (when the keywords occur in the tweet messages except the hashtags). Therefore, we combine the refined hashtags with the keywords, creating the final list of keywords and refined hashtags for each disaster. These terms are used to search tweets relevant to a given particular disaster. The matching-based approach is depicted in Figure 2b.

An advantage of our solution over conventional matching-based approaches is the ability to systematically generate a comprehensive list of relevant hashtags starting from a small set of core keywords. The number of relevant hashtags can be as large as a few hundred (see Table I) while existing studies [6], [14], [2], [16], [15] often use less than ten hashtags. Nevertheless, there may be still missing relevant tweets that do not contain the keywords or hashtags but are semantically relevant to a disaster. An example of such type of tweet is "*@Securb Safe and Sound! Just a lot of shaking in SF but*

DisasterId	Type	Core Keywords	Candidate Hashtags	Refined Hashtags
napa_earthquake	earthquake	quake, tremor, foreshock, aftershock	300	108
michigan_storm	flood	flood, storm, typhoon, tornado, hurricane, mudslide, strong wind, high water	89	87
newyork_storm			832	705
texas_storm			166	125
iowa_stf			160	124
iowa_stf2			168	116
iowa_storm			20	13
washington_storm			65	33
jersey_storm			20	18
california_fire	wildfire	fire, firing, burn, buring, blaze, blazing, flame, framing	318	149
washington_mudslide	flood + wildfire	*keywords of both flood and wildfire*	108	64

TABLE I: Summary of used keywords and hashtags.

no damage! Thanks@for checking!." This tweet is relevant to Napa earthquake although it does not contain any core keywords or refined hashtags of the disaster.

2) Learning-based approach: Learning-based approach tends to include more tweets into relevant sets when compared to the matching-based method. The accuracy of learning-based techniques often highly depends on the quantity and quality of training datasets. However, to best our knowledge there is no large labeled dataset available for classifying disaster-related tweets. Thus, we prepare our own training datasets from multiple data sources. In the following, we present our five-step solution (see Figure 2c) to classify tweets into relevant and irrelevant ones: 1) tokenizing each tweet into words, phrases and symbols called tokens, 2) representing tweets using bag-of-words model, 3) penalizing high-frequency words using TF-IDF model, 4) projecting tweets to fixed-size low-dimensional feature vectors, and 5) using logistic regression for classification.

a) Preparing training datasets: Existing studies [25], [26] filter tweets relevant to a specific disaster by building a separate model (or classifier) from a specific training dataset for each disaster (e.g., one for Colorado floods and another for Alberta floods). A challenge of this approach is to have large labeled data for training the models. However, labeled datasets used in these studies are often small, e.g., each training dataset in [25] has only about 1000 tweets.

To mitigate this issue, unlike [25], [26], we build the same model for all disasters of the same type. The reason for this is twofold. First, we improve the models by enlarging the sizes of the training datasets. Second, most tweets related to a disaster type share the same set of core keywords. Consequently, for every kind of the disaster (earthquake, flood, wildfire), we create different training data from publicly available data sources, including CrisisLexT26 [25] and CrowdFlower10K[2]. Note that we also train one hybrid model of flood and wildfire for *washington_mudslide*. A summary of the two data sources and the combined training datasets for each disaster type are shown in Table II. CrisisLexT26 includes tweets collected during 26 large crisis events in 2012 and 2013, with about 1,000 labeled tweets per crisis. CrowdFlower10K contains 10,876 tweets relevant to various kinds of disaster, from major ones that cause damaged structures to minor ones such as car accidents. This dataset was created in two steps: 1) automatically searching for tweets with keywords such as *"ablaze"* and *"quarantine"*, and then 2) tweets are manually classified by CrowdFlower's workers into one of the three labels: *"Relevant"*, *"Not Relevant"*, and *"Can't Decide"*. In order to enhance the quality of the training data, we discarded the tweets with *"Can't Decide"* label and the ones with confidence scores less than one (i.e., only keep tweets with 100% confidence). The confidence score is determined by CrowdFlower's workers. Finally, we use the keywords and hashtags in Table I to find tweets relevant to each disaster.

b) Tokenizing tweets: For each tweet, we use our customized tokenizer to remove non-ASCII characters, HTML tags and replace emojis with corresponding words (e.g., *";)"* by *"happy"*). We also attempt to split hashtags into meaningful elements if possible (e.g., *"californiaearthquake"* to *"california"* and *"earthquake"*).

c) Representing tweets using bag-of-words model: We use a bag-of-words (BOW) model to represent each tweet as a vector where each dimension denotes a particular word and a value which specifies the number word appearance in a particular tweet. Specifically, we use a powerful word embedding technique named *word2vec* [21], which was also used in many recent studies [36], [24], to create the BOW model. We use an implementation of *word2vec* provided by Gensim library [30]. The input of *word2vec* is a tweet corpus. It outputs a set of high-dimensional *sparse* vectors (one for each tweet) as most words in our collection of tweets do not appear in a particular tweet. In this study our tweet corpus contains every word that appears at least twice in the whole tweet corpus. With *word2vec*, the vectors of similar words are grouped together in the vector space. As a result, the word vectors capture many linguistic regularities [21]. For example, the cosine similarity of *vector("earthquake")* and *vector("aftershock")* would be higher than that for *vector("earthquake")* and *vector("flood")*, or vector operations *vector("earthquake")-vector("geophysical")+vector("tornado")* would result in a vector that is very close to *vector("meteorological")*.

d) Penalizing high-frequency words using TF-IDF model: We apply a well-known statistic in information retrieval and text mining, named term frequency-inverse document frequency (TF-IDF), to factor the importance of a word in the corpus. This model diminishes the weight of words that occur very frequently in the tweet corpus such as "the"

[2]http://www.crowdflower.com/data-for-everyone

Disaster type	CrisisLexT26		CrowdFlower10K		Combined training datasets
	Related	Not Related	Related	Not Related	
Earthquake	2,334	2,132	57	35	4,561
Flood	5,121	3,135	795	382	9,447
Wildfire	2,015	1,387	122	33	3,557

TABLE II: Summary of labeled datasets for training.

and increases the weight of terms that occur rarely such as "earthquake". Toward that end, each word count in the BOW model is penalized by multiplying with an inverse function of the number of tweets containing the word.

e) Projecting tweets to fixed-size low-dimensional feature vectors: Previous studies [6], [23], [11], [3] have used dimensionality reduction techniques to reduce the number of features of classifiers. Similarly, we use a popular topic-modeling technique, termed Latent Semantic Indexing (LSI), to transform the high-dimensional BOW space to a lower dimensional latent space. LSI uses Principal Component Analysis (PCA), which is a popular approach to map high-dimensional data into low-dimensional latent space while preserving as much variance as in the high-dimensional data as possible with the rapidly diminishing return for each new dimension. The output of this phase is a set of fixed-size low-dimensional feature vectors, which is fed into the logistic regression classifier. We use a popular implementation of logistic regression supported by the *scikit-learn* library in Python.

C. Analyzing sentiment of relevant tweets

In this section, we present sentiment analysis as a particular application of the disaster-related tweets identified by the prior phase (see Section III-B). We use this application to compare the results of the two approaches for identifying relevant tweets. To identify sentiment of individual tweets, we adopt a popular word embedding technique, named *doc2vec*, which has been shown to achieve state-of-the-art results for sentiment analysis tasks [18]. In [18], each document (or tweet in this case) is represented by a fixed-length vector.

We use an implementation of the *doc2vec* algorithm provided by the Gensim library[3]. We feed into the *doc2vec* model three kinds of data 1) a publicly available set of 1.6 millions labeled tweets (Sentiment140[4]) with an equal number of positive and negative ones for *training*, 2) a small set of labeled tweets for *testing*, and 3) our disaster-relevant tweets for *prediction*. The model outputs three sets of corresponding vectors. We use the training vectors to train a scikit-learn logistic regression classifier and the testing vectors to evaluate the accuracy of the sentiment classifier. Thereafter, we use the classifier to predict labels for the prediction vectors.

Using the Gensim tool, we trained the *doc2vec* model using 2,450,000 tweets with word vector dimension of 100 and vocabulary size of 2,178,060. We run *doc2vec* with a negative sampling rate of 10^{-4}, context window size of 10, and we do not ignore words with low total frequency.

[3]https://radimrehurek.com/gensim/models/doc2vec.html

[4]http://help.sentiment140.com/for-students/

IV. Results

A. Experimental Setup

We used a publicly available dataset in [29], which consists of IDs of geotagged tweets within the U.S during two time periods: June to November in 2014 and 2015. From the Tweet IDs, we created eleven datasets using Twitter streaming API, each corresponds to a particular disaster that was officially declared by FEMA during the time periods. We considered three specific types of disasters in our experiments: flood, earthquake and wildfire. Flood is the most frequent type of disaster in our datasets, and also the most common natural hazard in the world. A summary of these datasets is given in Table III, including various statistics such as the number of counties in the vicinity of the disasters and the number of affected counties. This data can be obtained from FEMA website, from which we computed the total number of geotagged tweets in the vicinity area and the number of those in the affected area. By default, we used the disaster-related tweets in the affected counties for evaluation.

B. Experimental Results

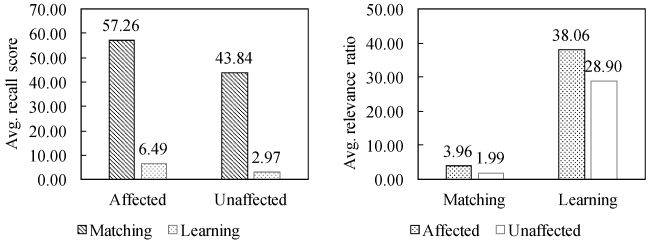
1) Comparison of the matching and learning approaches in identifying relevant tweets: We compute the number of disaster relevant tweets that are selected using both classifiers described in Section III-B. The results are shown in Table IV. We observe that the number of relevant tweets found by the learning-based classifier is much higher than that by the matching-based classifier. The reason for this might be twofold: the learning technique may include many irrelevant tweets and the matching technique would miss many relevant tweets. To elaborate this issue, as demonstrated in [20], [5], we assume that the set of agreed tweets by both classifiers has a higher accuracy than the results from an individual classifier and thus can be considered as a ground truth. Given such assumption, the fraction of relevant tweets that are retrieved (precision score) is always 1. As another measurement, we compute the *recall* of each classifier, which is the fraction of retrieved tweets that are relevant. The recall scores of each classifier in various disaster cases are shown in Table IV. We observe that the recall of the matching-based approach is an order of magnitude higher than that of the learning-based approach for all cases. This indicates that the matching-based technique outputs a smaller number of relevant tweets of higher quality when compared to the learning-based technique which tends to include more tweets with low quality. To summarize, we compute the average recall score of all disasters as presented in Figure 3a. We also observe that the recall scores for the affected regions are significantly higher than that in the

DisasterId	FEMA Code	Start Date	Duration (days)	Tweets	Counties	Affected Tweets	Affected Counties
napa_earthquake	4193	08-24-2014	16	1,868,964	58	374,782	2
michigan_storm	4195	08-11-2014	3	399,293	83	1,90,394	3
newyork_storm	4204	11-17-2014	11	227,073	62	143,505	9
texas_storm	4245	10-22-2015	10	231,808	254	72,088	22
iowa_stf	4184	06-14-2014	11	239,588	99	41,471	26
iowa_stf2	4187	06-26-2014	13	274,954	99	74,355	24
iowa_storm	4234	06-20-2015	6	32,286	99	5,721	19
washington_storm	4242	08-29-2015	1	79,381	39	9,217	6
jersey_storm	4231	06-23-2015	1	114,925	21	20,406	4
california_fire	4240	09-09-2015	52	430,253	58	1,916	2
washington_mudslide	4,243	08-09-2015	33	80,188	39	6,260	8

TABLE III: Statistics of 11 datasets.

unaffected areas. To explain this result, we show the number of disaster-related tweets found by our filtering techniques in both affected and unaffected regions.

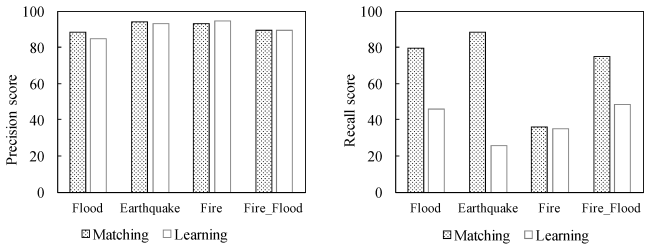
Particularly, we are interested in the percentage of related tweets retrieved, the number of relevant tweets divided by the total number of tweets, and refer to this value as *relevance ratio*. Table IV shows the ratio for both affected and unaffected regions. As expected, the relevance ratio in the affected regions is significantly higher than that in the *nearby* unaffected regions. This observation is true for all cases in our experiments showing that the population in affected areas are more likely to post disaster-related tweets. This result also explains the prior observation that the recall scores are higher in the affected areas when compared to the unaffected area. We also show average relevance ratio over all disasters (see Figure 3b), which shows that, with the matching-based approach, the number of relevant tweets found in the affected areas is twice of that in the unaffected areas.



(a) Accuracy of retrieved tweets (b) Percentage of related tweets

Fig. 3: Comparison of matching vs. learning-based approaches (average over 11 datasets).

Thus far, we assumed that the set of agreed tweets is the ground truth, which may not be true. Hence, we evaluate the matching-based and learning-based approaches in terms of precision and recall scores using the CrisisLexT26 dataset, which has ground truth labeled by the human (see Table II). We split CrisisLexT26 into two equal parts for training and testing. Noting that the matching-based approach does not have the concept of training data; thus, we create hashtags from the entire dataset. The results on the testing data are shown in Figure 4. As predicted, two approaches provide high precision scores, between 85% and 95% (Figure 4a). Importantly, when compared to the learning-based approach, the matching-based approach yields much higher recall scores (Figure 4b), e.g., 248% in the earthquake case. The reason for this is that the learning-based approach would include tweets that contain no relevant keywords or hashtags, which may reduce recall.



(a) Precision score (b) Recall score

Fig. 4: Comparison of matching vs. learning-based approaches on labeled dataset.

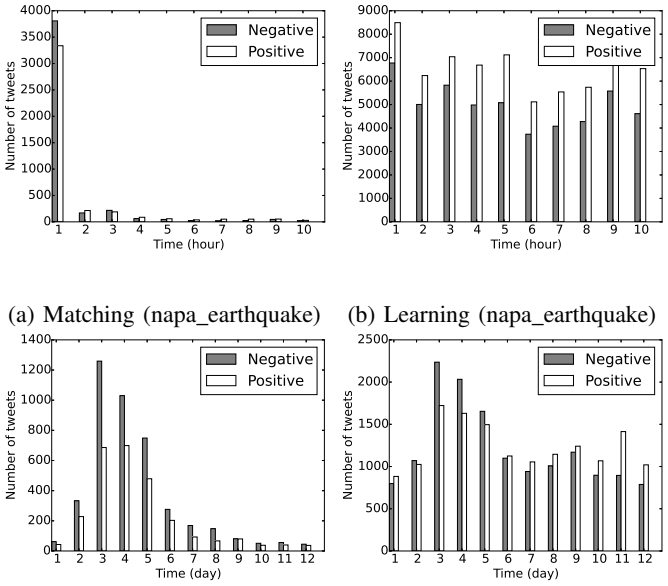
2) The impact of identifying relevant tweets to sentiment analysis application: In this section, we show the importance of accurately obtaining relevant tweets in a popular application of sentiment analysis. Figure 5 presents the results of two datasets: the Napa earthquake and the New York Storm. Each figure shows the number of positive and negative tweets in affected area over time (hourly or daily). Figures 5a and 5c show that, during the peak time periods of the disasters (the first hour in earthquake case and days 3,4,5 in storm case), far more negative tweets than positive tweets were posted when people suffered the most. Then, the number of relevant tweets diminishes quickly over time, especially in the case of the Napa earthquake when the duration of disaster was short. However, this trend is not clearly shown in Figure 5b and 5d where the learning-based approach was used. This is because the learning approach adds excessive irrelevant tweets that may have higher positive ratio than disaster-related tweets. This result confirms our prior finding that the matching-based technique is better at selecting relevant tweets.

Since the matching-based approach provides the sets of disaster-related tweets of higher quality, we will use this technique from now on.

3) Other results:

DisasterId	Spam Ratio (%)	Area	Number of Disaster-Related Tweets			Recall Score		Relevance Ratio	
			Matching	Learning	Agreement	Matching	Learning	Matching	Learning
napa_earthquake	26.00	Affected	8,548	116,187	3,948	46.19	3.40	2.92	39.75
		Unaffected	851	55,678	430	50.53	0.77	0.08	5.10
michigan_storm	28.81	Affected	2,638	31,129	1,183	44.84	3.80	2.07	24.40
		Unaffected	1,767	38,811	689	38.99	1.78	1.13	24.77
newyork_storm	24.69	Affected	6,952	29,412	3,786	54.46	12.87	6.73	28.47
		Unaffected	1,611	19,154	793	49.22	4.14	2.38	28.30
texas_storm	20.82	Affected	2,871	37,044	2,237	77.92	6.04	4.74	61.18
		Unaffected	4,251	37,921	1,561	36.72	4.12	3.46	30.83
iowa_stf	19.61	Affected	1,756	8,031	933	53.13	11.62	4.87	22.26
		Unaffected	3,782	37,304	1,702	45.00	4.56	2.42	23.83
iowa_stf2	20.11	Affected	2,010	17,937	1,193	59.35	6.65	3.18	28.38
		Unaffected	4,145	36,501	1,821	43.93	4.99	2.65	23.33
iowa_storm	17.87	Affected	192	1,112	61	31.77	5.49	3.65	21.12
		Unaffected	442	1,926	57	12.90	2.96	2.08	9.06
washington_storm	13.95	Affected	283	4,657	179	63.25	3.84	3.27	53.75
		Unaffected	1,873	26,976	980	52.32	3.63	3.14	45.23
jersey_storm	16.29	Affected	382	9,088	278	72.77	3.06	2.15	51.19
		Unaffected	1,307	38,093	862	65.95	2.26	1.67	48.56
california_fire	17.02	Affected	107	656	71	66.36	10.82	7.16	43.88
		Unaffected	643	130,494	233	36.24	0.18	0.18	36.71
washington_mudslide	14.75	Affected	174	2,774	104	59.77	3.75	2.78	44.31
		Unaffected	1,707	26,193	860	50.38	3.28	2.75	42.18

TABLE IV: Summary of results.



(a) Matching (napa_earthquake) (b) Learning (napa_earthquake)

(c) Matching (newyork_storm) (d) Learning (newyork_storm)

Fig. 5: Impact of matching vs. learning-based approaches to sentiment analysis.

a) The impact of extending the set of hashtags: In this section, we compare our matching-based approach and the conventional matching-based approach. Figure 6 shows the improvement of our technique over the conventional method in terms of selecting the number of relevant tweets. We observe that our approach produced a significantly bigger dataset of relevant tweets when compared to conventional approaches. The reason for the increase is that our matching-based technique is able to include a more diverse set of hashtags (e.g., *#3amearthquake, #quakeinsf*). This improvement in terms of quality is even more important given the high recall score of our matching-based approach. We also observe that the improvement is higher in populated areas such as New York because there are more unique hashtags being used.

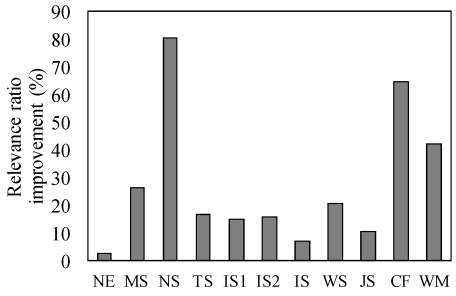
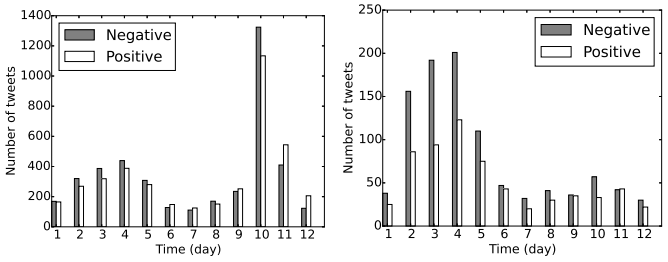


Fig. 6: The improvement of our matching-based technique over the conventional method. NS stands for newyork_storm, etc.

b) The impact of removing spam tweets on the number of relevant tweets: Table IV shows the percentage of spam tweets for each disaster, referred to as *spam ratio*. The spam ratio ranges from 14% to 29%. In removing the spam tweets, we observe that although the percentage of the spam users is only 0.8% (1,937 spammers over 144,297 unique users), they generate 23.33% of the total tweets (928,174 spam tweets over 3,978,713 total tweets). These statistics show that eliminating spam tweets is an important step in real applications. To illustrate, Figure 7 shows the impact of removing spam tweets on the number of positive/negative tweets obtained. Figure 7a shows an unexpected peak in day 10, which disappears after removing spam tweets (see Figure 7b). We further identify a set of 50 active spammers, each posted a large number of tweets (331 to 1121) on that day; most tweets from a spammer have the same content.

V. Conclusion and Future Works

In this paper we introduced a five-step framework for analyzing tweets during disasters to enhance situational aware-



(a) With spam tweets (b) Without spam tweets

Fig. 7: Impact of removing spam tweets to the number of tweets on unaffected region of the newyork_storm dataset.

ness. We proposed two different techniques for identifying tweets relevant to a particular disaster type. We observed that the matching-based technique includes less relevant tweets but with higher quality when compared to the learning-based approach. We confirmed our finding by conducting various experiments on three types of disasters (earthquake, flood, wildfire). As a future work, we aim to further evaluate trade-offs between the two techniques by obtaining the ground truth of the datasets through crowdsourcing using Amazon Mechanical Turk. We will also extend our framework to include man-made disasters such as terrorism. Another challenging problem to address is real-time classification of disaster-related tweets for timely analysis of the data. Finally, we aim to use the proposed framework to construct sentiment maps of the affected areas for real-time situational awareness during disasters. This is a step toward using the classified and sentimented information in real-world situation, which is often overlooked in the existing studies.

ACKNOWLEDGEMENT

This research has been funded by NSF grants IIS-1320149, CNS-1461963 and the USC Integrated Media Systems Center. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the sponsors such as NSF.

REFERENCES

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

[2] Z. Ashktorab, C. Brown, M. Nandi, and A. Culotta. Tweedr: Mining twitter to inform disaster response. *Proc. of ISCRAM*, 2014.

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

[4] C. Caragea, A. Silvescu, and A. H. Tapia. Identifying informative messages in disaster events using convolutional neural networks. In *International Conference on Information Systems for Crisis Response and Management*, 2016.

[5] C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, and A. Tapia. Mapping moods: geo-mapped sentiment analysis during hurricane sandy. *Proc. of ISCRAM*, 2014.

[6] A. Cobo, 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*, pages 1189–1194. ACM, 2015.

[7] M. De Choudhury, N. Diakopoulos, and M. Naaman. Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 241–244. ACM, 2012.

[8] N. Diakopoulos, M. De Choudhury, and M. Naaman. Finding and assessing social media information sources in the context of journalism. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2451–2460. ACM, 2012.

[9] FEMA. Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications. 2012.

[10] J. D. Fraustino, B. Liu, and Y. Jin. Social media use during disasters: a review of the knowledge base and gaps. 2012.

[11] X. Hu, J. Tang, H. Gao, and H. Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. ACM, 2013.

[12] X. Hu, L. Tang, J. Tang, and H. Liu. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 537–546. ACM, 2013.

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

[14] M. Imran, 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*, pages 159–162. ACM, 2014.

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

[16] M. Imran, S. M. Elbassuoni, C. Castillo, F. Diaz, and P. Meier. Extracting information nuggets from disaster-related messages in social media. *Proc. of ISCRAM, Baden-Baden, Germany*, 2013.

[17] N. Kanhabua and W. Nejdl. Understanding the diversity of tweets in the time of outbreaks. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1335–1342. ACM, 2013.

[18] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[19] B. Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[20] Y. Lu, X. Hu, F. Wang, S. Kumar, H. Liu, and R. Maciejewski. Visualizing social media sentiment in disaster scenarios. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1211–1215. ACM, 2015.

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

[22] A. Monroy-Hernández, E. Kiciman, M. De Choudhury, S. Counts, et al. The new war correspondents: The rise of civic media curation in urban warfare. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1443–1452. ACM, 2013.

[23] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *arXiv preprint arXiv:1306.5204*, 2013.

[24] D. T. Nguyen, S. Joty, M. Imran, H. Sajjad, and P. Mitra. Applications of online deep learning for crisis response using social media information. *arXiv preprint arXiv:1610.01030*, 2016.

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

[26] A. Olteanu, S. Vieweg, and C. Castillo. What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 994–1009. ACM, 2015.

[27] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[28] B. E. Parilla-Ferrer, P. L. Fernandez, and J. T. Ballena. Automatic Classification of Disaster-Related Tweets. In *Proc. International conference on Innovative Engineering Technologies (ICIET)*, pages 62+, Dec. 2014.

[29] J. Pfeffer and F. Morstatter. Geotagged twitter posts from the united states: A tweet collection to investigate representativeness. *Version: 1. GESIS Data Archive. Dataset. http://doi.org/10.7802/1166*, 2016.

[30] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[31] T. Sakaki, 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*, pages 851–860. ACM, 2010.

[32] K. Starbird and L. Palen. Voluntweeters: Self-organizing by digital volunteers in times of crisis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1071–1080. ACM, 2011.

[33] S. Vieweg, 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*, pages 1079–1088. ACM, 2010.

[34] S. E. Vieweg. Situational awareness in mass emergency: A behavioral and linguistic analysis of microblogged communications. 2012.

[35] D. D. Vu, H. To, W.-Y. Shin, and C. Shahabi. Geosocialbound: an efficient framework for estimating social poi boundaries using spatio–textual information. In *Proceedings of the Third International ACM SIGMOD Workshop on Managing and Mining Enriched Geo-Spatial Data*, page 3. ACM, 2016.

[36] S. Zhang and S. Vucetic. Semi-supervised discovery of informative tweets during the emerging disasters. *arXiv preprint arXiv:1610.03750*, 2016.