Classifying Microblogs For Disasters

Sarvnaz Karimi Jie Yin Cecile Paris
CSIRO Computational Informatics
Marsfield, NSW, Australia
firstname.lastname@csiro.au

ABSTRACT

Monitoring social media in critical disaster situations can potentially assist emergency and media personnel to deal with events as they unfold, and focus their resources where they are most needed. We address the issue of filtering massive amounts of Twitter data to identify high-value messages related to disasters, and to further classify disaster-related messages into those pertaining to particular disaster types, such as earthquake, flooding, fire, or storm. Unlike post-hoc analysis that most previous studies have done, we focus on building a classification model on past incidents to detect tweets about current incidents. Our experimental results demonstrate the feasibility of using classification methods to identify disaster-related tweets. We analyse the effect of different features in classifying tweets and show that using generic features rather than incident-specific ones leads to better generalisation on the effectiveness of classifying unseen incidents.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis

Keywords

Tweet Classification, Social Media Mining, Disaster Management

INTRODUCTION

Microbloggers cover events of all types in their posts. Events can be predictable, such as sport games, festivals, conferences, elections and ceremonies, or unpredictable, such as natural or man-made disasters, including earthquakes and terrorist attacks. Twitter, in particular, provides a popular medium for providing complementary sources of information and rapid communications during times of crisis. We investigate how to identify whether a Twitter message (tweet) refers to a disaster or not, in order to assist in more effective and efficient handling of unpredictable disastrous situations. For example, during the bushfire season in Australia, emergency services look for first-hand information from locals including reports of fires that are ignited due to the heat, or early warnings of such happenings, especially in remote areas. People on the ground can provide invaluable information via Twitter regarding the unfolding of events, critical damage to infrastructure, threats to lives, or any help available or needed. Our goal is to filter the massive amounts of incoming tweets[1] to identify high-value messages related to natural or man-made disasters in order to assist different authorities in dealing with specific types of disastrous events and taking appropriate actions promptly.

Although Twitter potentially offers the advantages noted above, it also poses new challenges. The text in tweets is often quite noisy, phrased using informal language, and potentially containing inaccurate information, such as rumours, mixed with reality. The high volume of colloquial text is also largely dominated by mundane updates that are personal, pointless babbles. From this heterogeneous mix, it is a non-trivial task to identify high-value messages that are relevant and specific to current disastrous events.

Simply using keyword-based search to identify disaster-related tweets leads to the retrieval of a large proportion of *false positives*. Some typical examples are as follows.[2]

*She's a natural disaster: a **tsunami** in her eyes an **earthquake** in her chest a **hurricane** **flooding** her mind she's a traveling catastrophe.*
*Recently, my office has been **flooded** with gifts from some super friends.*
*@M1Jarvis I don't like u! my stomach is on **fire**! Can't move! Hate you ! Regards , Niall*

Although these tweets include mentions of disaster-related keywords such as "*earthquake*", "*flood*", or "*fire*", they are associated with a totally different context, not referring to disastrous events. In our experiments, we use a list of disaster-related keywords to create a data set of 5,747 tweets, more than 50% of which are annotated as non-disaster, although they include these keywords.

Another approach to find disaster-related tweets could be to use hashtags. Hashtags are usually created soon after an event interests a large crowd of users on Twitter. This is particularly the case for disastrous events; for example, the hashtags #eqnz and #qldflood were created after the 2010 Christchurch earthquake, New Zealand, and the 2010

[1]Half a billion tweets per day in late 2012 (http://cnet.co/U3hOUW - accessed 23 Sep 2013).

[2]The search keywords are bolded in these examples to indicate why these tweets match.

ADCS '13, December 05 - 06 2013, Brisbane, QLD, Australia

Disaster Type	Description
Earthquake	Natural disaster, also known as quake or tremor.
Flooding	Only natural disaster, excludes flooding due to pipe breakage.
Fire	Includes natural or man-made fire with substantial effect on a society. All types of fire included, e.g. wildfire, or fire started in a gas station.
Storm	Includes cyclone, typhoon, wind-storm, tornado, or hurricane.
Civil disorder	Including riot, terrorist attack, or protest.
Traffic accident	Any motor vehicle crash, such as car accident, and plane crash.
Other	Any other disaster that is not covered above but is a non-personal event. Anything private is not included.

Table 1: Disaster types and their definition.

Queensland floods, Australia. However, hashtags alone can not provide sufficient evidence for distinguishing disaster-related tweets. First, not all disasters have a specific hashtag associated with them, and hashtags are not created until after a crisis takes place. Second, even after hashtags are created, they might change over time, or the same hashtag may be used for vastly different purposes [Bruns and Burgess, 2012]. Finally, a disaster-related hashtag may be used much later after the disaster occurs, for example to mark an anniversary. The following examples illustrate these. Although these tweets contain hashtags "#qldflood" or "#eqnz", they do not indicate that a disastrous event is present somewhere, and therefore should be considered as non-disaster.

Gympie Mayor Ron Dyne says a flood study to consider a levy for the Mary River will start in the next two weeks. ***#qldflood***
Yuss! I win at ***#eqnz*** *quick draw! #competitiveaboutstupidshit.*
A quiet moment of ***#eqnz*** *reflection to mark the second anniversary of Christchurch's Feb 2011 earthquake.*

To deal with these difficulties, we present a machine learning based approach to classify microblogs for signs of a variety of *disaster types*, including earthquake, flooding, fire, and storm (Table 1). The majority of previous work studying events in Twitter data is either focused on extracting information for predictable events, such as sport games [Lanagan and Smeaton, 2011], or if they considered disasters, they mostly performed *post-hoc analysis* on specific incidents that happened in specific geographical locations [Sakaki et al., 2010, Verma et al., 2011, Sreenivasan et al., 2011]. In contrast, in this work we address the general problem of distinguishing a wide variety of natural or man-made disasters, instead of focusing on specific incidents. More importantly, we consider a more realistic setting, in which we train a classification model on past incidents, and recognise tweets about unseen incidents.

Main contributions of our work are: (1) We analyse the discriminative power of different features in classifying tweets. We show that using *generic features* rather than *incident-specific* ones is more useful not only for discriminating disaster-related tweets from non-disaster-related ones, but also for determining a disaster type; (2) We show the potential in cross-disaster-type training of classifiers, which plays a positive role in identifying previously unseen types of disasters; and, (3) We point out the time-dependency of microblog data and that conventional evaluation methods, such as K-fold cross-validation, would generate biased classification results. As an alternative, we propose using *time-split evaluation*, which can eliminate the effect of future knowledge for tweet classification.

RELATED WORK

The idea of using social media content to extract and understand current news has been explored for a variety of applications. Relevant studies can be categorised into two primary areas: first, tweet classification which deals with classifying very short, noisy, and informal text; second, studies that investigate the role of social media during times of crisis.

Tweet Classification

Classification of short and sparse text has been studied in the past in various scenarios, such as search snippets, product reviews, chat, and forum text. Most of these studies have proposed using additional sources of information augmented with a bag-of-words approach in order to compensate for the sparsity and brevity of the data. This problem has also been addressed for Twitter. To build a system called TweetStand that shows worthy news to its users, Sankaranarayanan et al. [2009] classified tweets into two classes: news or junk. They used a Naïve Bayes classifier with bag-of-words features from Tweet content. Sriram et al. [2010] made a step forward and classified tweets into five generic categories, including news, events, opinions, deals, and private messages. They investigated a combination of features including author, presence of shortened words and slang, time-event phrases, opinionated words, emphasis on words, currency and percentage signs, username at the beginning of the tweet, and username within the tweet. For a collection of 5,407 manually annotated tweets, they show that their feature set significantly outperforms bag-of-words only features for all the five classes. In our work, given the different nature of the tweets we are interested in, we use only some of these features. For example, opinions are not the focus of our work.

Nishida et al. [2012] presented a wide range of tweet classification frameworks using a temporally aware Naïve Bayes classifier. Their experiments were conducted on a data set in which classes were defined based on their hashtags. That is, a set of hashtags were considered to belong to a class. Three different generic classes were defined. Our work, however, avoids such simplification for two reasons: first, disaster hashtags are not pre-defined and depending on where they happen and what their types are, they appear and disappear; second, some hashtags are abused by spammers. We therefore opt for a more careful data set and labelling creation.

In a recent study, Duan et al. [2012] proposed a way of adding context to short texts of tweets. They introduced collective classification where tweets sharing the same hashtag or the same URL were considered similar. They classified a sample of tweets from five days into seven categories of entertainment, politics, science and technology, lifestyle, business

and products, sports, and others. This method helped to augment the context for classification improving over a bag-of-words classifier. They reported an accuracy of approximately 69% using their best setting. Our work is built on the content of a single tweet at a time. We show that using bag-of-words can be very effective for our problem. We note that our classes, disaster types, are contextually much more similar to each other than generic classes, such as science and entertainment, which do not share similar vocabulary.

Twitter During Disasters

The importance of social media during emergency situations has been emphasised in previous studies. Starbird et al. [2010] were one of the first to study and analyse Twitter data during disasters. They performed a qualitative analysis on tweets published during a flooding incident to study the behaviour of microbloggers. They discovered that people rely and distribute official information on Twitter more than on other untrusted sources.

By monitoring Twitter in Japan, Sakaki et al. [2010] detected earthquakes. They developed a system that acts as a quick alarm for people to prepare themselves for the coming disaster. To fulfill this, they developed a classifier that works on tweets retrieved with two specific query terms related to earthquakes: *earthquake* and *shaking*. They used three sets of features: tweet length and position of query term in the tweet, tweet words, and the context of the query term in the tweet. They reported an F-measure of around 73% for two queries on a data set of 597 positive examples. In our work we take this work further to detect not only earthquakes, but also other disaster types such as cyclones or bushfires.

Verma et al. [2011] investigated identifying tweets that contained situational awareness information during emergency situations. Using Twitter data (2,000 tweets) from four crisis events in the United States, they showed that tweets contributing to situational awareness are often objective, impersonal, and formal. Experimenting with two types of classifiers, Naïve Bayes and Maximum Entropy, they showed high accuracies (approximately 84-89%) for each event in their collection. In our work, we do not limit ourselves to specific incidents. Our work also looks at disaster tweets in general and is not focused on situational awareness.

Twitcident [Abel et al., 2012] is a web-based framework built on Twitter to filter and analyse tweet streams once a disaster happens. Twitcident offers a classification of tweets for casualties, damage, or users experience in the crisis situation. This classification however is implemented using handcrafted rules.

Overall, our work is different from previous work as it does not rely on data from a specific location or specific type of event. We cover multiple types of disasters from various geographical locations and focus on using classifiers trained on past events to classify unseen incidents.

THE DATASET

Our data set was created with two criteria in mind: it covers a variety of disaster types (see Table 1) and a wide range of locations in the world. We randomly sampled a total of 6,500 tweets published in a range of two years, from December 2010 till November 2012. We did not include retweets.

Tweets from Australia and New Zealand were retrieved using the ESA system [Yin et al., 2012], which works based on the Twitter streaming API[3], and stores a massive database of tweets from 2010 till present. Tweets from other parts of the world were retrieved using Twitter search API, using a list of generic disaster-related keywords, such as fire, flooding, storm, tornado, hurricane, cyclone, and earthquake. The complete list of keywords can be found in [Karimi and Yin, 2012].

Our data set contained tweets about the following disasters, among others: earthquake in Christchurch, New Zealand, 2011, Cyclone Yasi in Queensland, Australia, 2011, the Queensland floods, 2010-2011, severe bushfires in Victoria, Australia, 2011, the earthquake in Melbourne, Australia, 2012, the fire in Diamant Hotel in Canberra, Australia 2011, York the floods, England, 2012, and the Hurricane Sandy, United States, 2012.

	Type	No. Tweets
Disaster	Earthquake	1049
	Fire	667
	Flooding	617
	Storm	351
	Traffic accident	84
	Civil disorder	17
	Other	26
	Don't know/Ambiguous	39
	Multi-type	11
Non-disaster		2897

Table 2: Distribution of disaster types in the annotated tweets.

ANNOTATION

For a machine learner such as a classifier to work, we need to present it with a representative set of training data (positive and negative examples). We therefore annotated our tweet data set manually to identify disaster tweets and their types. We annotated our data based on two main questions:

1. Is this tweet talking about a disaster? (Yes or No);
2. What type of disaster is it talking about (as defined in Table 1)? (multiple choice).

Annotations were done by three annotators for each tweet hired through Crowdflower,[4] a crowdsourcing service over Amazon mechanical Turk. After taking a majority vote where at least two out of three annotators have 100% agreement on both of the questions, we ended up with a set of 5,747 annotated tweets, from which 2,850 tweets were identified as disaster-related and 2,897 as non-disaster. The split of the data was therefore almost even between positive and negative samples. The breakdown of tweets annotated for each type of disaster is shown in Table 2. In disaster tweets, 37% were annotated with earthquake, followed by fire, flooding, and storm comprising 23%, 22%, and 6% respectively. A small portion of tweets were labelled for multiple types of disasters (multi-type). For example, the following tweet was annotated with both fire and storm:

> *The #storm meets a #bush #fire in the #suburbs of #newcastle.*

[3] `https://dev.twitter.com/docs/streaming-apis`
[4] `http://crowdflower.com/`

We removed multi-type tweets from our data set because of their low percentage (less than 1%) compared to single-type tweets. Also, since the number of tweets in the categories of traffic accident and civil disorder was small, we merged them with *other* and *ambiguous* categories, creating one class called *other*. This way, our data was divided into six types: earthquake, fire, flooding, storm, other, and non-disaster. In our annotation guidelines, *ambiguous* tweets were disaster-related tweets that the annotators had difficulty in associating them with one of our defined disaster types.

MICROBLOG CLASSIFICATION FOR DISASTERS

In our work, we focus on two tweet classification tasks: (1) *Disaster or not*: a binary classification, where tweets are classified into disaster-related or not; (2) *Disaster type*: a multi-class classification, where tweets are predicted to one of the following six classes: non-disaster, storm, earthquake, flooding, fire, and other (any other disasters), as defined in Table 1.

Classification Methods

For our classification tasks, we experimented with Support Vector Machines (SVMs) [Chang and Lin, 2011], from the category of discriminative classifiers, and multinomial Naïve Bayes [Juan and Ney, 2002] from the category of generative classifiers. Both have been demonstrated to perform well for similar text classification tasks. As SVMs consistently yielded better classification accuracies overall, we only report the classification results based on SVMs.

Feature Extraction

Text classification requires training and testing instances to have a set of representative features. The choice of such features depends on both the text to be classified and the classification task.

Tweets have specific characteristics when compared with traditional text: they are short, limited to 140 characters; Often microbloggers use tweets in reply to others by using *mentions*, Twitter usernames preceded by @, or employ *hashtags*, such as #CycloneEvan, to make the grouping of similar messages easier, or to increase the visibility of their posts to others interested in the same topic. The use of *links* to Web pages, mostly full stories of what is briefly reported in the tweet is also popular. Selecting features for a text classifier built on Twitter data can therefore benefit from both conventional text features, such as n-grams, and Twitter-specific features, such as hashtags and mentions. Vosecky et al. [2012] provide an extensive list of the variety of features that could be considered for Twitter data, including language style and timestamps.

In this study, we investigate the effect of the following features and their combination on the performance of two classification tasks:

N-grams Unigrams and bigrams of tweet text at the word level, excluding any hashtag or mention in the text. To find n-grams, we pre-processed tweets (as detailed in [Karimi et al., 2012]) to remove stopwords and punctuations;

Hashtag Binary features of the hashtags which indicate whether a specific hashtag exists in a tweet or not. Since hashtags are usually created to indicate a topical event, they are incident-specific features;

Hashtag Count The total number of hashtags contained in a tweet;

Mention A binary feature of mentions indicating the absence/presence of a specific mention in a tweet; Similar to hashtags, mentions are incident-specific features because for a particular event, there often exist key players or influential people or organisations who are frequently replied by other people.

Mention Count The total number of user mentions contained in a tweet.

We also consider two other types of features: **Link**, which is a binary feature specifying whether or not a tweet contains any link to a web page, and **Tweet Length**, which is the total number of words in a tweet, including hashtags, mentions, and links. In all our experiments, adding bigrams, links, and length of tweets led to insignificant improvements of the accuracy over unigram-only features and therefore, we do not include their results in our experiments.

EXPERIMENTS

We run experiments to evaluate how effectively our classifiers can identify tweets that are relevant to a disaster, and whether or not we can identify the type of disaster as defined in Table 1. In our experiments we evaluate the effect of different factors on classification effectiveness as listed below.

- Effect of training on specific past incidents on the classification of current incidents. Do hashtags make training specific to an incident and what is the effect of removing hashtags from the tweet texts? Does training on one incident or type of disaster narrow our classifiers to specific incidents or types of disasters?
- Effect of using incident-specific or generic features in classification accuracy. What are the best features to use for disaster classifiers?
- Effect of training size. What is the impact of different sizes of the training data on the classification accuracy?

We evaluate classification effectiveness using accuracy (percentage of correctly classified tweets).

Time-Split Evaluation

K-fold cross validation, first introduced by [Geisser, 1974], has been the most popular evaluation approach for supervised classification tasks. Therefore, most existing research studies on tweet classification have adopted K-fold cross validation to evaluate the effectiveness of the classification models (*e.g.*, [Sriram et al., 2010, Takemura and Tajima, 2012, Vosecky et al., 2012]). This evaluation approach overlooks the time-dependency among microblog data (*i.e.*, tweets). It is an unrealistic setting, because a classifier during its training makes use of future evidence—including hashtags, mentions or other contextual information such as names of disasters or locations—to predict other related tweets relevant to those events.

To eliminate the effect of future knowledge on classification, we use time-split evaluation as an alternative to evaluate the performance of our classifiers. Specifically, we sorted our data set based on the publication time of the tweets, and then divided the data into ten equal portions based on time. We kept the most recent 10% portion for testing, and the rest for training. The size of the training data was increased incrementally from 10% to 90%, using the earliest portion of the data each time. Evaluation was done for two settings: a binary setting of deciding whether a tweet is disaster related or not, and a multi-class classification setting where a tweet was classified for disaster type.

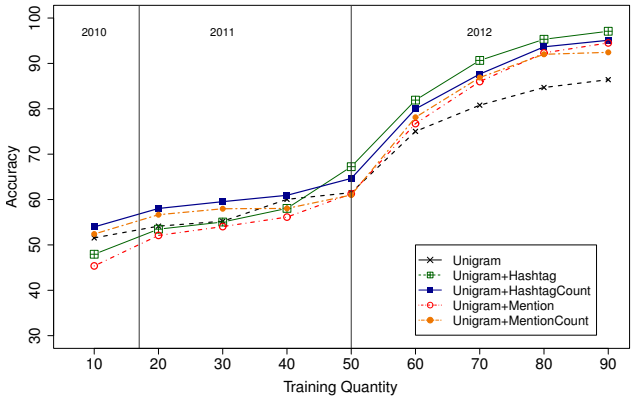
Figure 1 shows the classification results. When only using unigrams as features, the *disaster or not* classifier achieved just over 50% accuracy (51.5%) when only the earliest 10% of the data was used for training. Accuracy increased to 86.4% when 90% of the data was used for training. Incident-specific features, hashtags and mentions, did not help when less than 50% of the data was used for training. In contrast, more generic features (hashtag counts and mention counts) boosted the accuracies over the unigrams only baseline. For example, when we used hashtag counts in addition to unigrams at 10% cut-off, the accuracy was 58.0% (an absolute 7.5% more than unigrams only), while the combination of unigrams and hashtags achieved the accuracy of 43.1%. Using mentions and mention counts had similar effect, except that they were consistently less effective than hashtags and hashtag counts. The reason for hashtags and mentions not being useful is that they are specific to the events that happened in late 2010 and 2011, whereas our testing data belongs to late 2012.

Only after we used 50% of the data for training, use of the hashtags and mentions in addition to unigrams became more effective than using their respective counts. At 50% cut-off using unigrams led to 61.6% accuracy, the use of unigrams and hashtags gained 67.2% accuracy, and the use of unigrams and hashtag counts was 64.7% accurate. The use of hashtag or mention counts, however, consistently improved the results over unigrams only, suggesting that they are reliable features to use for both of the classification tasks. This is when we saw fluctuations in accuracies when using hashtags and mentions themselves. The reason is obvious: when a hashtag is known for a disaster or the key tweeters during a disaster are known, classification becomes easier. For example, @CDCemergency almost always tweets about a disastrous event. Once it is known, no matter what point in time, classifier is more confident to classify tweets to a disaster class.

Disaster type classifier followed a similar trend. However, the accuracies were smaller than that *disaster or not* classifier. The accuracies were lower by between 3 to 11 percent in each cut-off (10% to 90%) and all feature combinations. This is due to the higher complexity of classifying into six classes compared to two.

To shed light on why the features we chose were useful, we calculated the average number of hashtags in tweets of our data set. Tweets annotated as disaster-related had 1.17 hashtags on average, whereas non-disaster tweets had only 0.37 hashtags. On the other hand, the number of unique hashtags in non-disaster tweets was 1.5 times higher than that in disaster-related tweets. This number was 2.3 for mentions. We also calculated lexical diversity between these two sets and the same pattern holds. Disaster-related tweets were

(a) Disaster and Non-Disaster



(b) Disaster Type

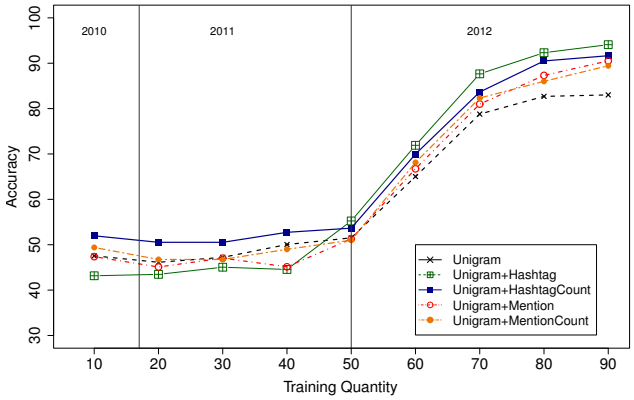


Figure 1: Classification accuracy for different training cut-off and different feature combinations. Breakdown of tweets based on year is also shown. Dotted lines are added to show the trend and do not represent any data point.

less diverse in their vocabulary. This explains why the features we presented above are effective.

Discussion

Our specific task of identifying disastrous events requires to learn the patterns that imply a type of natural or man-made disaster so as to detect new instances of the disaster. However, for a machine learning approach, we can only provide historic instances of specific disasters to train a classifier. Built upon the seen instances, a classifier could easily bias towards specific events or specific locations associated with events as seen in our experiments. Hashtags and mentions increasingly became effective when more of them were revealed. Generally, this situation even deteriorates due to the brevity and sparsity of tweets.

For example, the following tweet dating back to December 2012 is specific to bushfire in New South Wales, Australia:

> *A massive cloud of smoke can be seen in south-west Lake Macquarie from the Wyee bushfire #nswfires #wyeefire @NewcastleHerald*

A classifier trained on the tweets specific to this bushfire can easily associate its location to bushfire. Therefore, if it later

on encounters with the following tweet, it may associate it with bushfire.

Lake Macquarie is big & beautiful `http://lockerz.com/s/257143427`

It is therefore crucial for the training of the classifiers to be exposed to a variety of tweets from different locations both during a disaster, and other time periods. Also, it is very important to train the classifiers using generic features, such as hashtag count and mention count, rather than those specific to individual disasters (specific hashtags).

Cross-Disaster Classification

As natural and man-made disasters are unexpected events, different parts of the world can be affected by different or even new *types* of disasters, for some of which we may not have annotated data on which to train our classifiers. Even for the same type of disaster, one can only train a classifier on past incidents of disasters while any new incident has its own specific hashtags and affected locations. Therefore, we also conduct experiments to evaluate the ability of our classifiers to identify previously unseen disaster types.

In our experiments, we excluded all the data for one type of disaster in the training set, and then tested the classifier on the tweets of the excluded type. Since a classifier can only assign labels that are seen in its training phase, we only used our *disaster or not* classifier to determine whether or not a tweet is talking about a disaster.

Given that the number of tweets that we had annotated for each major type of disaster (earthquake, fire, flooding, and storm) was imbalanced, we used *under-sampling* [Liu et al., 2008] to create training and testing data. In under-sampling, the size of instances of different classes is balanced by random reduction of data from the larger classes. For training, in addition to a set of randomly selected 1,050 non-disaster tweets, we selected 350 tweets from each disaster type (size of the smallest class, *i.e.*, storm). The final size of the training set was 2,100 tweets. For testing, we had 700 tweets, where 350 tweets were from the excluded disaster type, and the rest from randomly picked non-disaster tweets that did not overlap with the training set.

Table 3 shows the classification results at five different feature settings. For all four types of disasters, using unigrams only worked better than random, and it was particularly effective for fire and storm. Adding hashtags or mentions as bag-of-words degraded classification accuracies as compared to using unigrams only. For example, using hashtags for the earthquake and flooding disaster types led to an accuracy of around 50% which is equivalent to a random guess. This indicates that hashtags and mentions are incident-specific features for particular types of disasters. Thus, classifiers built on these features were biased and could not generalise well to classify other types of disasters. In contrast, the addition of more generic features, such as hashtag count and mention count, improved the results. The most effective setting was using unigrams plus number of hashtags.

This experiment has two important implications: first, it emphasises that using disaster related tweets, regardless of their type, is potentially helpful for identifying previously unseen disasters. Therefore, a smaller amount of annotations will be needed per disaster type if data from other disasters is available.

Second, it shows that incident-specific features, such as hashtags and mentions in tweets, are highly relevant to specific incidents and disaster types. If a hashtag is included in the training data on an incident of interest, classification becomes accurate (as seen in the previous section) but specific to that type of disaster or even that specific incident. However, if a classifier trained on past incidents or even different types of disasters is used to predict current occurrences of disasters, using incident-specific features results in poor generalisation performance of making predictions on unseen data. In such cases, generic features are more powerful for classifying disastrous events.

CONCLUSIONS AND FUTURE WORK

In this work, we focused on extracting relevant and useful information from Twitter during times of crisis to assist in better handling of critical disaster situations. We explored the problem of filtering tweets, which are short text messages on the popular social media service, Twitter. We presented a classification framework for filtering microblogs into disaster and non-disaster messages. We also investigated classifying microblog messages by disaster type. We compared using features of more generic nature with features that are incident-specific and their role as more data becomes available and more information on one incident, or similar incidents become accessible. Using more generic features of tweets such as how many hashtags they contain or how many user mentions they contain were consistently helping, while more incident-specific features were only useful only after some specific information about an incident was already revealed to the classifier. Both these feature types have their own applications. The former is useful for early detection of content on disasters, whereas the latter is more useful for tracking Twitter content on a known incident.

The number of location mentions in a microblog has the potential for differentiating disaster tweets from non-disasters. Lingad et al. [2013] showed that simple re-trained named-entity recognisers are very effective in recognising location mentions in microblogs. In the future, we will also experiment using this feature in our classifiers.

We highlighted an important evaluation concern for studies similar to ours. Given the strong temporal aspect of tweets, it is unfair to use conventional evaluation methods, such as cross-validation, that use the entire available data set as a mixed bag. Instead, we recommended time-split evaluation for tweet classification, where only older tweets are used for training.

We investigated whether or not the data on historic disastrous events can provide enough information for identification of disasters of a different type. Our results showed that an SVM classifier trained using simple features such as bag-of-words and number of hashtags in a tweet can identify disasters of other types. Classification of fire-related tweets was particularly promising with approximately 73% accuracy when the classifier was trained on earthquake, storm, and flooding data only. These results are important because manual annotation of data is expensive.

Classification of noisy short text without context is difficult. In the future, we will consider providing the contextual information from other related tweets, and potentially the knowledge from other resources to improve the classification accuracy. We will also consider not only relevancy, but also recency using temporal models. Another critical question

Training	Testing	Accuracy				
		Unigram	Unigram+Hashtag	Unigram+HashtagCount	Unigram+Mention	Unigram+MentionCount
Flooding+Fire+Storm	Earthquake	55.4	50.4	**60.5**	50.3	57.0
Earthquake+Fire+Storm	Flooding	56.2	50.6	**59.6**	52.2	57.2
Earthquake+Flooding+Storm	Fire	69.8	62.5	**72.9**	61.2	71.7
Earthquake+Flooding+Fire	Storm	62.9	52.3	**64.8**	54.1	63.6

Table 3: Accuracy of our *Disaster or Not* classifier when the training data (first column) does not contain any disaster tweet from the testing type of disaster (second column).

to answer is what useful information our identified disaster tweets can provide before, during and after disasters. We have further annotated our data set for four main types of information: announcing a disaster, asking for help, offering help, and reporting damage. This rich data set will give us an opportunity to further classify tweets for assessing their impact on people and infrastructure.

References

F. Abel, C. Hauff, G. Houben, R. Stronkman, and K. Tao. Semantics + filtering + search = Twitcident. Exploring information in social web streams. In *Proceedings of the 23rd ACM conference on Hypertext and social media*, pages 285–294, Milwaukee, Wisconsin, 2012.

A. Bruns and Jean E. Burgess. Researching new discussion on Twitter: New methodologies. *Journalism Studies*, pages 801–814, 2012.

C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

Y. Duan, F. Wei, M. Zhou, and H. Shum. Graph-based collective classification for tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2323–2326, Maui, Hawaii, 2012.

S. Geisser. A predictive approach to the random effect mode. *Biometrika Trust*, 61(9):101–107, 1974.

A. Juan and H. Ney. Reversing and smoothing the multinomial Naïve Bayes text classifier. In *Proceedings of the 2nd Int. Workshop on Pattern Recognition in Information Systems*, pages 200–212, Ciudad Real, Spain, 2002.

S. Karimi and J. Yin. Microtext annotation. Technical Report EP13703, CSIRO, 2012.

S. Karimi, Jie Yin, and P. Thomas. Searching and filtering tweets: CSIRO at the TREC 2012 microblog track. In *Text Retrieval Conference*, 2012.

J. Lanagan and A.F. Smeaton. Using Twitter to detect and tag important events in sports media. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, pages 542–545, Barcelona, Spain, 2011.

J. Lingad, S. Karimi, and J. Yin. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 1017–1020, Rio de Janeiro, Brazil, 2013.

X. Liu, J. Wu, and Z. Zhou. Exploratory under-sampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(2):965–969, 2008.

K. Nishida, T. Hoshide, and K. Fujimura. Improving tweet stream classification by detecting changes in word probability. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 971–980, Portland, Oregon, 2012.

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, Raleigh, North Carolina, 2010.

J. Sankaranarayanan, H. Samet, B.E. Teitler, M.D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51, Seattle, Washington, 2009.

N.D. Sreenivasan, C.S. Lee, and D.H.L. Goh. Tweet me home: exploring information use on twitter in crisis situations. In *Proceedings of the 4th international conference on Online communities and social computing*, pages 120–129, Orlando, FL, 2011.

B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in Twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 841–842, Geneva, Switzerland, 2010.

K. Starbird, L. Palen, A.L. Hughes, and S. Vieweg. Chatter on the red: what hazards threat reveals about the social life of microblogged information. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 241–250, Savannah, Georgia, 2010.

H. Takemura and K. Tajima. Tweet classification based on their lifetime duration. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2367–2370, Maui, Hawaii, 2012.

S. Verma, S. Vieweg, W.J. Corvey, L. Palen, J.H. Martin, M. Palmer, A. Schram, and K.M. Anderson. Natural language processing to the rescue? extracting "situational awareness" Tweets during mass emergency. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 2011.

J. Vosecky, K. Wai-Ting Leung, and W. Ng. Searching for quality microblog posts: Filtering and ranking based on

content analysis and implicit links. In *17th International Conference on Database Systems for Advanced Applications*, pages 397–413, Busan, South Korea, 2012.

J. Yin, S. Karimi, B. Robinson, and M. Cameron. ESA: emergency situation awareness via microbloggers. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2701–2703, Maui, Hawaii, 2012.