Data mining Twitter during the UK floods

Investigating the potential use of social media in emergency management

Thomas Spielhofer
The Tavistock Institute of Human Relations
London, United Kingdom

Deborah Markham
Oxford Computer Consultants Ltd
Oxford, United Kingdom

Reynold Greenlaw
Oxford Computer Consultants Ltd
Oxford, United Kingdom

Anna Hahne
The Tavistock Institute of Human Relations
London, United Kingdom

Abstract — **During a large-scale crisis, many people use social media to share information. However, emergency services are unable to use this information easily due to data quality and quantity issues. We use Twitter data generated during a recent flooding crisis to gain insights into this problem, evaluating techniques that could be used in real-time to provide actionable intelligence to emergency services. Our work also includes an exploratory analytical study of the data.**

Keywords—Twitter, data, flooding, emergency services, data mining

I. Introduction

The convergence of social networking and mobile media technology is shifting the way people communicate, and gain or share information, even in emergencies or crisis situations. Social media now represents a continuous feed of an unstructured, unvalidated real-time record of citizens' experiences and when a large scale emergency occurs, such as a flood, earthquake or rioting, a lot of valuable information is shared on sites such as Twitter, Facebook or Instagram. During recent catastrophes throughout the world, like hurricane Sandy in 2012, the European floods 2013/14 and terrorist attacks in Paris and Brussels in 2015, social media have been widely used [1]. Indeed, the terrorist attack in Paris generated around 6.7 million posts with the hashtag #prayforparis on Twitter within 10 hours [2]. Currently the emergency services have no established methods of monitoring this data, filtering it and integrating the results into crisis management. Social media is not routinely used by emergency services *either* to inform their situational awareness and response, *or* to communicate with the citizens they serve (for example to respond to or request further information).

Social Media has the potential to provide actionable intelligence to emergency services during a crisis but its large quantity, poor structure and lack of validation make it a source that is challenging to integrate into the emergency response. Against this background, the main aim of the 'Emergency Management in Social Media Generation' (EmerGent: www.fp7-emergent.eu) project, funded by the European Union's Seventh Framework Programme, is to develop effective information mining techniques to transform these large quantities of low value information into low quantities of high value (*i.e.* actionable) information.

Some platforms for harnessing the power of social media during emergency situations have been developed. For example, Artificial Intelligence for Disaster Response [3] is a platform for classifying tweets during a disaster. Classifications are user-defined, and supervised machine learning techniques are used to categorize the tweets. Another tool Tweedr [4], further extends this idea and uses clustering to group together related Tweets by means of a previously developed ontology of damage types and causality [5]. Both of these tools have the potential to be valuable to emergency services, but generally rely on very specific incidents to be searched for (e.g. #Irene during hurricane Irene). Other platforms developed for emergencies are even more specific. For example, a tool has been developed to detect earthquakes in Australia and New Zealand [6]. This system identifies earthquakes in Australia and New Zealand from tweets, sending an email notification to the Joint Australian Tsunami Warning Centre when an earthquake is detected. These alerts are generated based on relevant word frequency bursts and a classifier. Although their method is fairly successful, it is limited by its application to earthquakes in Australia and New Zealand and it is not clear how easy it would be to scale up to other crises.

The EmerGent project aims to develop a more generally useful tool for emergency services that collects, analyses and presents relevant information in their area. This tool should allow them to search for specific hashtags or manually label messages, but it will not require them to do so. Instead, the software itself will aim to synthesize message in a meaningful way. However, this is a challenging aim. When searching social media for less specific terms, the data is far 'noisier' and the lack of context makes it more challenging to know if an event is taking place in a relevant location to the user. This is the problem we address in the first part of this paper. For this purpose we will describe and assess techniques implemented for extracting and refining information present in social media streams. In the second part of this paper, we then present the findings from the analysis of such data relating to a recent emergency in the UK in December 2015. We intend as part of this paper to inform readers about the details of the algorithms we have chosen, their strengths and weaknesses as demonstrated by their application to sample Twitter data, as

This research has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 608352.

978-1-5090-5234-9/16/$31.00 ©2016 IEEE

well as their context in the field as a whole. Moreover, we intend for readers to realize the benefits of the algorithms we have implemented to emergency services, as well as the potential benefits of our future work.

II. The Dataset

In December 2015 and January 2016, a series of storms, starting with Storm Desmond, caused localized flooding in the north west of England, and other parts of the UK. Cumbria was the worst-hit county [7]: more than a month's rain fell in one day on Saturday (5th December) and main rivers all across Cumbria exceeded the highest levels ever recorded. Further heavy rainfall and severe flooding occurred over Christmas 2015 as a result of Storm Eva. Storm Frank followed at the end of the year, bringing storms and severe gales to western parts of the UK and the north-west of Scotland. Further flooding occurred, leading to many homes being evacuated. The BBC reported that one person died as a result of these floods. In total, the Government has confirmed that around 16,000 properties in England were flooded [8].

We used the 2015-2016 floods in the UK to identify, and take steps towards resolving, key challenges when accessing and analyzing social media data during large-scale incidents. To do this, we captured a subset of Twitter data using the public Twitter API between 25th November 2015 and 13th January 2016 (inclusive). The only inclusion criteria were that the tweets had to be in English, were not retweets, and contained any of the following phrases:

- Flood[1];
- Heavy rain;
- Stormy weather.

We did not select on any other criteria (such as location, the presence of certain metadata or the characteristics of the author). Over this fifty-day period we collected a total of 970,106 tweets. A qualitative assessment of the data indicated a number of issues. Firstly, as expected, many of the tweets were not about flooding, but instead used other meanings of the search terms, such as 'floods of tears' or song titles with 'heavy rain' in them. Secondly, only 1.5% of our dataset had a location given by geotags. We need the location of a tweet both to remove posts from outside the UK (there were posts about floods in other countries), and to identify locations of relevant incidents within the UK. Finally, the data is poorly structured: it is difficult to quickly find specific information, and to find tweets discussing the same incident.

III. Information Extraction

To address the difficulties associated with social media data highlighted in the previous section, we have investigated a number of techniques to make the data more usable to emergency services. In this paper, we focus on three key areas: (i) noise reduction; (ii) location detection; and (iii) consolidation.

[1] This includes any words for which flood are the stem, e.g. flooding or flooded.

'Noise' reduction

For social media data to be useful to emergency services, we need a method to automatically remove irrelevant messages. The problem we are facing is very similar to that of email spam filtering, for which Naïve Bayes Classification [9, 10] is a commonly used technique. Following this approach, we implemented our own Bayesian text classifier: a supervised machine learning technique which uses conditional probability to classify text based on a bag of words.

We used the dataset described above to evaluate the success of our approach, creating two random sets of 1,000 tweets each: (i) a training set and (ii) a testing set. We manually labelled each of the tweets in these two sets as "spam" (not relevant to flooding in the UK) or "not spam" (relevant to flooding in the UK). We used our training set of data to train our Naïve Bayes Classifier. Following this, we used our classifier to automatically label the test set of data as "spam" or "not spam" and then compare these labels to those determined manually. A summary of our results can be seen in Table 1. We find that 76.1% of our tweets have been correctly labelled by the classifier, when compared to our manual labels. These success rates are similar to those of other text classification studies in the literature [3].

TABLE I. Comparing manual and automatic labelling for a set of tweets

Category	Automated label	Manual label	Proportion
True Positive	Not spam	Not spam	31.9%
True Negative	Spam	Spam	44.2%
False Positive	Not spam	Spam	20.6%
False Negative	Spam	Not spam	3.3%

Examining the results, it is particularly striking that we see a large number of false positives, but a low number of false negatives – where a Twitter message of relevance to the floods was falsely classified as irrelevant. One of the main reasons we are getting a large number of false positives is that many messages are about flooding outside the UK, thus have a lot of similarities with messages about flooding in the UK. If we only use posts with geotags, we could eradicate messages from outside the UK easily (but would only use 1.5% of our total sample – as only a very small proportion of tweets have geolocation enabled, which provides the latitude and longitude of the user.). The other reason we get more false positives is that if none of the words in our tweet (aside from the keyword) were encountered in the training set, the tweet is automatically classified as "not spam" to avoid losing potentially relevant information. This is because it is more important to us that potentially useful tweets are not missed than that some irrelevant tweets are present. This emphasis on reducing false negatives at the expense of including false positives is specific to critical scenarios; for example, this approach would not be considered with brand management tools. Even with the large number of false positives included, we have reduced the number of tweets we need to examine by 44%. Moreover, we can reduce the number of false positives over time by providing feedback to our classifier, in the same way as we can improve spam detection in our email by manually labelling wrongly identified emails. As a simple demonstration of this, we move ten false positives from our testing set to our training set and relabel them correctly. This increases our training set to 1,010 tweets and decreases our testing set to 990. We retrain our classifier with this new training set, and once again compare our automated and manual labels. We find that 77.2% of our tweets in the testing set have now been labelled correctly by the classifier. The number of false positives has been reduced from 20.6% to 19.3%.

When using Bayesian classification, there is a confidence score associated with each classification label. Thus the number of false negatives can be reduced by setting a minimum threshold on this confidence score when deciding which messages to mark as spam. For example, if we require a confidence score above 20% when the classifier has labelled a message as "spam", we reduce the number of false negatives in Table 1 from 33 to 23 (a reduction of 30.3%). This number can be decreased further still by increasing the threshold for the confidence score. Of course, increasing the threshold may also lead to more false positive and less correct negative results.

Location detection

As discussed, a lot of the false positives were the result of including tweets from outside the UK. This highlights the importance of identifying the correct origin of tweets. For this reason, we explored whether it is possible to automatically deduce the location a social media post refers to from its textual contents. We use a gazetteer, GeoNames[2], to provide coordinates (latitude and longitude) for place names in the UK (note that the GeoNames gazetteer is not limited to the UK and we have since performed location detection for German tweets). Our method involves checking for matches with words and phrases in a tweet and assigning a primary location. If we obtain multiple locations for a single tweet, we eliminate unigrams where a bigram is present (e.g. remove Yorkshire if West Yorkshire is present), and choose the most frequently mentioned location as the primary location.

To examine the results of our location deduction, we used a random set of 1,000 tweets from the sample described above. We manually labelled these with any UK locations we could deduce from the text. Out of this random set, we manually labelled 144 (14.4%) with UK locations. The majority of these locations were not very specific, for example: Cumbria or North Wales. We then ran the same set of tweets through our automated location detector and compared our results. These are shown in Table 2.

TABLE II. DETECTING LOCATION IN TEXT

Category	Automated result	Manual result	Proportion
True positive	Same location detected	Location detected	10.9%
True negative	No location detected	No location detected	80.5%
False positive	Location detected	No location detected	5.8%
False negative	No location detected	Location detected	2.8%

We see that 91.4% of our automated results match our manual labels (both detected the same location, or both did not detect a location). On the whole, the results appear promising when you consider that 91.4% of tweets are correctly labelled. However, in other respects the results do not seem so good from a user's perspective. If, for this dataset, a user chooses to select posts about the UK only, they will view 16.7% of the data (the false positives and true positives). Of this subset, 34.7% will show falsely detected locations. The false positive results have two major causes:

i. Overlapping place names between the UK and other countries (e.g. Boston in the USA and Boston, Lincolnshire in the UK);

ii. Place names matching regular words or phrases (e.g. reading a book, as opposed to a post about an event in Reading, Berkshire).

Combining location detection with noise removal will hopefully remove some of the former. To reduce the latter, we are considering using part-of-speech tagging to ensure the phrases being matched are likely to be locations. The false negative results tend to be because of misspellings (Cmbria instead of Cumbria) or derivations of place names that are obvious to us when labelling manually, but that our algorithm does not pick up (e.g. Cumbrian vs Cumbria). To reduce the number of false negatives, we need to correct spelling errors. Moreover, we need to be able to extract locations from their derivatives.

However, the current results are already much more useful than only viewing a small proportion of tweets (1.5% in our test dataset) which have the location of the user available from geotagging.

Consolidation

We need to be able to condense the available information to enable emergency services to quickly get an overview of events. To enable us to perform sophisticated queries and determine relationships between social media posts, we index our data using Apache Solr[3]. A major strength of Solr is its scalability: as the amount of data we need to process grows we can easily distribute our index over multiple servers, avoiding any performance issues. This is an essential requirement as the

[2] http://www.geonames.org/

[3] http://lucene.apache.org/solr/

use of social media is continually increasing [11] and we need to be able to quickly process data in a critical situation.

We can use Solr to group posts by their location. Currently, this is done by using the location, district or county name, depending on the scale we require. However, there is functionality within Solr to group by geographical coordinate. We are able to use the “faceting” feature of Solr to identify the most frequent words and phrases within a specific location. For example, in Figure 1, we show the top twenty-five token frequency results (excluding search terms and the specified location) for all the data, as well as for three specific locations, over the whole time-range of our dataset in Section 4.2. The most frequent words and phrases give an indication of the situation in a specific place, without needing to read multiple tweets.

Fig. 1. The most frequent tokens for specific data partitions

We can also group the data by time ranges, choosing intervals depending on the requirements of a particular user. For some emergencies, only posts over a few hours will be related, whereas for other events we expect days of posts to contribute to the overall view of an incident.

Additionally, we investigate grouping results by the textual similarity between posts. To group by text, we use cosine similarity [12] to create a similarity matrix for use by a hierarchical clustering algorithm [13]. We anticipated that clustering our posts by their textual similarity would help with the lack of location information: if a number of posts were identified as discussing the same topic and one had a location, we could infer that the others were nearby. However, we found this did not work well for our flooding dataset. This is because tweets about different incidents across the UK often had very high similarity scores. For example, somebody in York might post 'the main road is flooded!', and a similar message could be posted by a user in Carlisle. This leads to false groupings and inaccurate deductions. It may be worth considering using text clustering within a location to break down incidents, and we may find text clustering useful for other types of emergencies.

IV. Analysis of Data: the Cumbria 2015 Floods

In order to explore the usefulness of data extracted using the methods described above, the Tavistock Institute conducted some exploratory analysis of a sub-set of the full data. This included all tweets deemed to be of relevance to floods in the Cumbria region – one of the regions most badly affected. Analyzing Twitter data can help to understand the use of social media in emergencies in many ways, including when it is used, who uses it, and what purpose it is used for. In many previously conducted studies, social media data generated during emergencies have been analyzed, such as for example Twitter data in relation to the Thai Flood in 2011 [14] or Twitter data linked to the flooding in different parts of Australia [15]. Cheong and Cheong identified key players on Twitter and their role in distributing information during the flooding [15]. For the aim of this study, data were analyzed with regards to the scale of information available on Twitter during the floods, on the accounts and the content of the tweets for Cumbria.

The research team used a combination of hashtags and temporal specifications to restrict the data-set collected and filtered as part of the data mining techniques described above to data of relevance to the Cumbria flooding. This resulted in a dataset consisting of around 7,000 tweets related to the flood (excluding retweets) between 1st December and 18th December 2015, containing at least one of the following hashtags: #floodaware, #Cumbria, #SpiritofCumbria, #FloodAppeal, #cumbriafloods, #cumbriaisopen, #carlislefloods, #openforbusiness, #StormEva, #StormGertude, #StormFrank, #StormDesmond, #CumbriaFloodAppeal, #CumbriaStorm.

Number and frequency of tweets

Analysis of this dataset showed that the number of tweets per day varied immensely within this period of time (see Figure 2). There was a significant increase in the number of tweets on the 5th December 2015. This coincided with the start of the

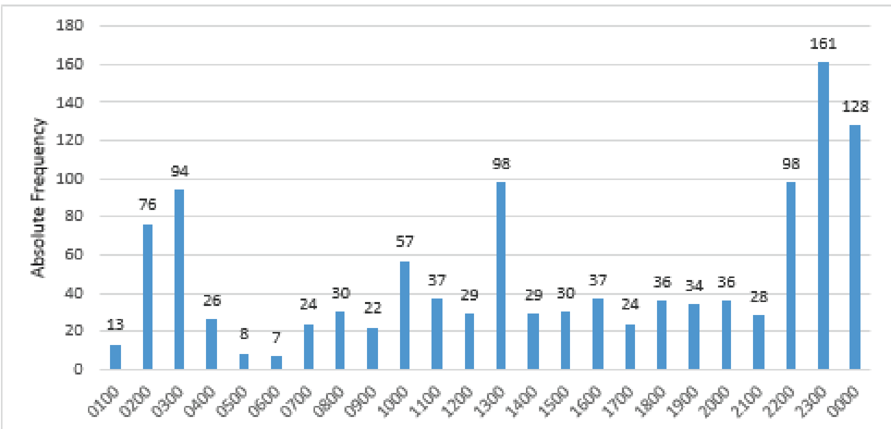

Fig.2. Number of tweets per day between December 1st and December 18th 2015

flood, after heavy rain fall from the 4th December 2015 onwards (see: http://www.metoffice.gov.uk/climate/uk/interesting/december2015). More than 1000 tweets per day were posted on 5th to 7th, and 9th December 2015. The peak in the number of tweets was on the 9th December 2015. The number of tweets decreased dramatically after the 9th December 2015 aligned with the recovery period.

Furthermore, the variations in the quantity of tweets throughout a day was explored. The research team chose the peak day for Twitter activity. Most tweets were posted between 10pm and 12am (see Figure 3). More than 70 status updates per hour were tweeted between 1am and 3am, 12pm and 1pm, and 9pm and 10pm. During the remaining time periods 50 or less tweets per hour were noted.

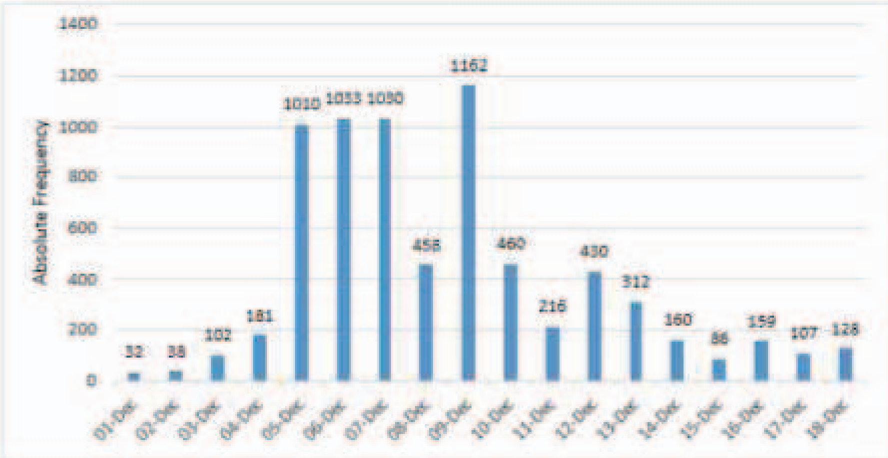

Fig.3. Number of tweets per hour on 9th December 2015

Accounts and content of tweets of Cumbria floods Twitter data

In total, 4322 users tweeted at least one Tweet in relation to the flood between 1st December and 18th December 2015. The majority of these users only tweeted once (see Figure 4). Less than 10% of the users tweeted three times or more, and only 0.9% tweeted more than 9 relevant status updates. In the following, the focus is on accounts that posted at least 10 relevant tweets.

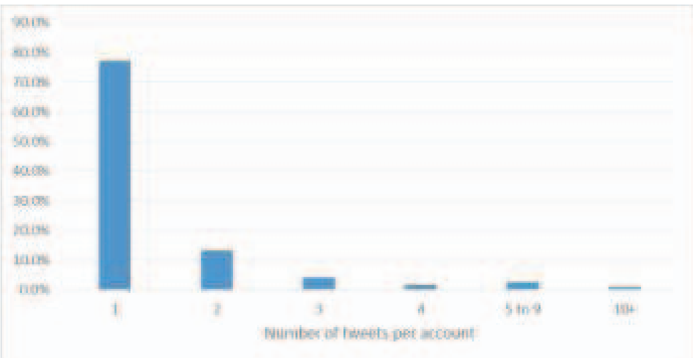

Fig. 4. Percentage of users tweeting a certain number of times

With 397 tweets, this was the highest number of relevant tweets associated to an account called “FloodAlerts”. “FloodAlerts” is an application that continuously posts flood warnings and river levels in England, Wales, and Scotland. The second most frequent Tweeter during the flood in Cumbria is “DailyCUMBRIA” with 42 tweets. In total, 41 accounts were found and classified into: news agencies, citizens, “streaming accounts”, relevant organizations (Environment agencies, Council, Police, Emergency Services, Natural Resources Wales, Emergency Service, and Lifeboats crew), and other organizations. The majority of these accounts were owned by relevant organizations (11), news agencies (10), and “streaming accounts” (10).

After exploring the accounts of the most frequent tweeters during the flood in Cumbria in 2015, the aim was to study the content of the tweets. Not only can a Naïve Bayes Classifier be used to filter for relevant tweets as seen before, but it can also be used to categorize relevant tweets into different types of content. This process is similar to a sentiment analysis, but instead of focusing on sentiments as positive, neutral, and negative emotions expressed in tweets, the focus is on the content of tweets. The classification using Naïve Bayes was performed using the statistical analysis programme R[4] and served as an explorative initial approach to detect benefits and potential issues. The first step of the analysis was a classification of the types of contents. Reading through the tweets led to the definition of three types; ‘safety warning’, ‘flood warning’ and ‘update’. Following this, 100 tweets were manually classified as ‘safety warning’, ‘flood warning’ and ‘update’. 90 of those classified tweets were used to train the Naïve Bayes Classifier. Afterwards, the trained classifier was used to predict the content of the 10 tweets of the test data set. The prediction accuracy was high, however, the classifier predicted all tweets to be ‘Flood Warnings’, which was the predominant type of content. So in this particular case, the Naïve Bayes Classifier was not applicable due to the high prior probability of one type of content, which has been observed before [16]. For this reason another commonly used machine learning algorithm, Maximum Entropy classifier (Maxent), was applied. Using the same manually classified tweets resulted in a prediction accuracy of .75. Applying Maxent to the total data set revealed that 6433 of the tweets were related to ‘Flood warnings’, 140 to ‘Safety warnings’ and 531 to ‘Updates’.

V. CONCLUSIONS

This paper aimed to describe several techniques to make Twitter data more usable to emergency services and to explore what information can be extracted from the tweets. This is done using the recent example of a severe flood in the UK in December 2015.

[4] http://www.R-project.org

We first demonstrated some techniques to address challenges when using social media data to quickly glean information from large-scale crises. There are a lot of data in social media and it is difficult and time-consuming to pick out the relevant information. We have addressed this issue by showing the success of a Naïve Bayes Classifier at automatically removing irrelevant data and dramatically reducing the amount of information we need to process manually. During a crisis, it is essential to emergency personnel to know the locations of specific incidents. Unfortunately, this information is rarely present on tweets. We have made some progress towards picking out the location a tweet is discussing. However, this is not ideal and we are not able to determine whether these are eye-witness reports or someone commenting on the crisis from an entirely different location. This problem is difficult to resolve, and the best solution is for Twitter users to be made aware of the need for them to share their location (either through the metadata, or by including a postcode in their message) when tweeting information in a critical situation. Finally, we also demonstrated ways we can provide consolidated information about an event. Doing so enables us to quickly get an idea of an ongoing event when there is a lot of information, and we can then delve into individual tweets if necessary.

As a next step, we performed some exploratory analysis on a subset of the data, focusing on a specific area for a shorter time period – the crisis point in that particular area of the UK (Cumbria). We investigated when Twitter is used to share information and by whom. This showed that not only emergency services and citizens are using Twitter, but also that applications were created to post up-to-date information. Furthermore, we explored the content of the tweets with a Naïve Bayes Classifier. Even though in this particular case (i.e., chosen types of contents) the classifier was not applicable, the results of our analysis are useful in determining steps to move forward with our techniques for real-time processing of the data for emergency services. Our classifier, currently used to determine irrelevant tweets, can also be used to categorize information within its limitations.

Bringing together our knowledge from different perspectives has provided useful insights into the flooding events occurring in the UK in the 2015-2016 winter period. Moreover, it has provided inspiration for future work when developing techniques to make social media data more usable in a crisis. Moving forward, we plan to apply our techniques and analysis to different types of critical situations, for example the 2016 bombings in Brussels.

ACKNOWLEDGMENT

This research has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 608352.

REFERENCES

[1] A.L. Hughes, L.A. St Denis, L. Palen, and K.M. Leysia, "Online Public Communications by Police & Fire Services during the 2012 Hurricane Sandy," Proceedings of the Conference on Human Factors in Computing Systems (CHI), 2014.

[2] V. Goel, and S. Ember, As Paris Terror Attacks Unfolded, Social Media Tools Offered Help in Crisis. The New York Times, 14 November 2015, available at: http://www.nytimes.com/2015/11/15/technology/as-paris-terror-attacks-unfolded-social-media-tools-offered-help-in-crisis.html?_r=0

[3] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "AIDR: artificial intelligence for disaster response," International World Wide Web Conference (IW3C2). Seoul, 2014.

[4] Z. Ashktorab, C. Brown, M. Nandi and A. Culotta, "Tweedr: Mining Twitter to Inform Disaster Response," ISCRAM 11th. Pennsylvania, 2014.

[5] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, "Practical extraction of disaster- relevant information from social media," WWW 13 Companion, Switzerland, 2013.

[6] B. Robinson, R. Power, and M. Cameron, "An evidence based earthquake detector using Twitter," Proceedings of the Workshop on Language Processing and Crisis Information, 2013, pp.1–9.

[7] Storm Desmond brings flooding and disruption to parts of UK**:** *BBC News. 5 December 2015*, available at: http://www.bbc.co.uk/news/uk-35014745

[8] S. Priestley, Winter floods 2015/16. House of Commons Library BRIEFING PAPER Number CBP7427, 21 January 2016, available at: http://researchbriefings.parliament.uk/ResearchBriefing/Summary/CBP-7424

[9] C.D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[10] T. M. Mitchell, Machine Learning. New Delhi: McGraw Hill, 1997.

[11] J. van Djick, The Culture of Connectivity: A Critical History of Social Media. Oxford University Press, 2013.

[12] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," in AAAI vol. 6, 2006, pp.775–780.

[13] Y. Zhao, G. Kartypis, and U. Fayyad, "Hierarchical Clustering Algorithms for Document Datasets," in Data Mining and Knowledge Discovery vol. 10, no 2, 2005, pp.141–168.

[14] A. Kongthon, Ch. Haruechaiyasak, J. Pailai and S. Kongyoung, "The role of Twitter during a Natural Disaster: Case stady of 2011 Thai Flood," Int. J. Innovation Technol. Management, vol. 11, No 2, 2014.

[15] F. Cheong and C. Cheong, "Social media data mining: A social network analysis of tweets during the 2010-2011 Australian floods," PACIS Proceedings, 2011.

[16] R. Hsu, B. See, and A. Wu, "Machine Learning For Sentiment Analysis on the Experience Project," unpublished.