Finding out the needy one from Tweets : An analysis using #kerelafloods

Rahul Ranjan
Computer Science & Engineering
IIIT Senapati,Manipur (IN)
Email: rahul@iiitmanipur.ac.in

Himangshu Sarma
Computer Science & Engineering
IIIT Senapati,Manipur (IN)

Navanath Saharia
Computer Science & Engineering
IIIT Senapati,Manipur (IN)

Abstract—Natural disasters are very difficult to predict and when it happens its very difficult for government and different aid agencies to get the real time information about the disaster effected people or properties. To save the life and properties first step to know where it happens and what is the current scenario of that place. To get these information social media plays a key role. In this article, we explores a detailed overview on Kerala flood situation which happened in late July 2018, severe flooding affected the Kerala very badly. In this article, we proposed a system to detection of real-time help needed to people based on different tweets related to Kerala flood. This article provided a complete detailed overview on the situation in Kerala and collects needed information through tweets as twitter is one of the best way of spreading awareness about the worsening situation in Kerala. To alleviate this problem, tweets, which are largely available,can be exploited to extract the required data.

Index Terms—keralaflood, LDA,K-Means Clustering, flood analysis system,Clustering Algorithm, Topic modeling.

I. Introduction

Natural disasters are very difficult to predict and therefore it always cost lots of casualties as well as damages of properties. According to the records from 2010 to 2017 only there were 4983 natural disasters events worldwide, where maximum of those happens in Asia only[1]. Floods and storms are the maximum number of events out of these. In between 2010-2015 only 822 floods and 525 storms are recorded. Every year 139 million people affected from natural disaster and 72,205 people died from it[2].

Technology playing an important role to help the suffering people from natural disasters. In this article, we tried to analyse the crisis communication on Twitter[3] happened during Kerala Floods in early August 2018 where over 445 people died. To alleviate this problem, tweets, which are largely available, can be exploited to extract relevant data. Twitter is one of the fastest communication medium using which people can report different realtime problems/activities happening to the world. It is also very easy to track for different organizations to get the location of the victim or complainer. In case of natural disaster, it's not easy for government or other organizations to get all the information in realtime from the ground zero to take the necessary action. In this kind of environment, if the rescue person will not get the information when it needed it may cost someone's life. Therefore, twits can play a major role of detecting the location of the place as well as what kind of help needed to save humans life.

II. Background Study

Time to time technology helped the people who are suffering from natural disasters. During the Haiti earthquake 2010, another SMS system designed for communication in between disaster effected people and aid agencies named as Trilogy Emergency Relief Application (TERA)[4]. From the short messages of Nepal's earthquake a researcher detected the emotions which may be used for different applications to provide help to the survivors [1]. During the Nepal's earthquake, NASA launched a device named as Finding Individuals for Disaster and Emergency Response (FINDER) to detect human's who are buried under buildings, roads etc. This device is a size of small briefcase which is a heartbeat detection radar designed to detect victims buried in rubble [2]. The next section discusses about our proposed system.

III. Proposed System

In our system, we used tweets of Kerala floods as an input to analyze the different scenario to help the people in the realtime. The twitter platform was widely used for sharing information, awareness,etc. Using twitter data we have analysis the whole crisis properly by processing the data. We have extracted the hashtages from tweets and applied clustering algorithms (e. g K-Means) to group similar messages and information like rescue,actions,supplies,emergency calls, together which took place on twitter.

The analysis is done from the tweets which are extracted during the early August 2018 when the flood happens. The analysis is restricted to 15000 tweets extracted by looking

[1]https://www. statista. com/statistics/510959/number-of-natural-disasters-events-globally/; Access Date: 31st August 2018

[2]https://safetymanagement. eku. edu/resources/infographics/when-disaster-strikes-technologys-role-in-disaster-aid-relief; Access Date: 31st August 2018

[3]https://twitter. com

[4]https://safetymanagement. eku. edu/resources/infographics/when-disaster-strikes-technologys-role-in-disaster-aid-relief; Access Date: 31st August 2018

for the hashtag #keralaflood. Topic analysis of preprocessed tweets is done using Latent Dirichlet Allocation (LDA) [3]. The System extracts the tweets on running a script after every 60 minutes automatically and stores in a file (say output. csv).

Following Steps have been Followed to achieve the required result -

- The data has been obtained from twitter using TweetManager & TweetCriteria and by using a Python-script which extracts username,mentions,tweets,date & time of tweet,tweet from ,tweet till,top-tweets,max-tweets,etc and writes it to a comma-separated value (. csv)file.
- The tweets extracted are saved in a . csv file (eg. output. csv) and later on imported using pandas.
- Preprocessing of the tweets is done which involves removal of stopwords, URLs, punctuations, tokenization, stemming and other techniques were applied.
- The hashtags are extracted from the tweets and frequency of each tweets is counted. To explore the the system in detail the top 15 tweets are plotted using matplotlib.
- Next step was to extract the bi-gram models from tweets for better analysis of data, for this purpose nltk collocations was used.
- To cluster the similar meaning words together, our task was to convert raw tweets(text) into matrix of vectors using tf-idf ,such that it could be utilized in K-Means clustering algorithm [4].
- The clustered data was ordered properly. To find central subject in the sets of documents (tweets) and topic modeling Latent Dirichlet Allocation (LDA) is used.

The following section discusses the methodology:

A. Data

As discussed earlier, the input data for the system is tweets related to Kerala flood. The “tweet,” or messages used in Twitter, are limited to 140 characters which creates a wonderful practice of being concise with the message you would like to convey. The tweets also contains images, URLs, videos, gif,etc. the best things it contains hashtags, which are words that capture the subject of the tweet and they are prefixed by ’#’ character. The hashtags may carry sentiments or emmotions (#sarcasm), actions (#accomodate), climate (#rain) and many more. Also, there are some other symbols as well like ’’ which symbolizes usernames or handles, retweet (‘rt’) is a tweet by a user X that has been shared by user Y to all of Y’s followers,a heart shape logo represents a ’like’ on twitter. The Tweets are collected from twitter by using its own authentication method an API called OAuth [5], that indexes tweets that match a given search string and writes the obtained tweets to the output file. It is a powerful & opensource tool to extract tweets from twitter. The usernames in the extraction query can be provided as ’#keralafloods’.

B. Data preprocessing

The hashtags were extracted from the tweets stored in (. csv) file and on proper analysis it gives various information about the situation in Kerala like weather forecasts (#rain), relief (#help), support(#chiefministerfund), location affected (#saveErnakulam),etc. To get the accurate information, text Processing is a very important task to obtain less noisy and better result which will lead to a must needed information for flood affected people. It involves -

- Removal of punctuations.
- Removal of stop-words, emoticons, numbers and slangs [6].
- Removal of URLs.
- Extract meaning bearing words.

C. Approach & Implementation

After preprocessing, we have counted, which hashtag was used most frequently, counted its frequency and plotted using matplotlib (Figure 1). From the analysis, we found that some hashtags were occurring frequently, i. e. , keralaflood, kerala, twitter, help, please, people, relief, etc. and some less frequent hashtags, i. e. , work, chief, collect, road, etc.

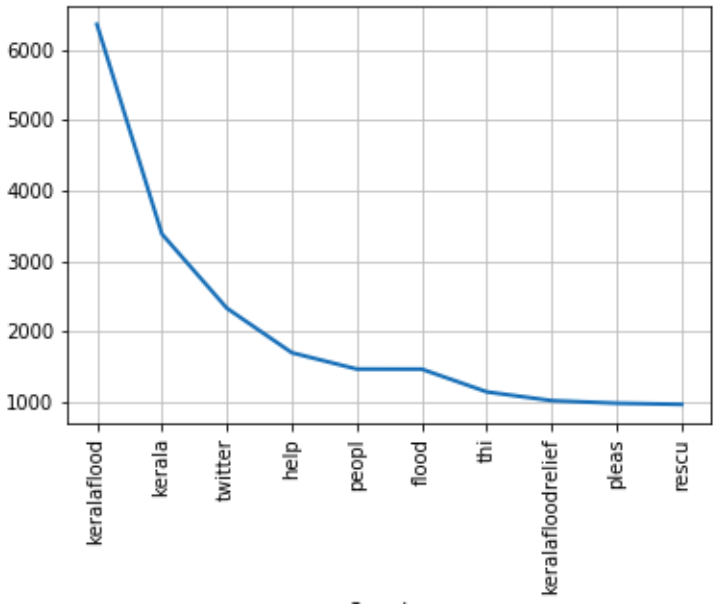

Fig. 1. Top 10 #hashtags used (Count vs hashtags)

The data contained many similar meaning hashtags during flood and it was a tough task to deal with such data, so these data required clustering of similar tweets & topic modeling to reduce the amount of noisy data. The tweets were grouped together based on clusters based on closeness or distance amongst them. Before using vectorizer it is better to add each tweets to the data

	twt1	twt2	twt3	twt4		twt(n)
T1	10	0	1	0	. . .	0
T2	0	2	0	0	. . .	18
T3	0	0	0	0	. . .	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .
T(m)	0	1	8	0	. . .	3

TABLE I
TF-IDF Matrix

frame, and then we have used Term Frequency Inverse Document Frequency (TF-IDF) [7] to vectorize the tweets. The TF-IDF vectorizer converts the tweets into vectors when each tweets are processed and appended to a list and this list was provided to the TF-IDF vectorizer. Each value in the vector depends on count of words or a term appears in the tweet (TF) and on how rare it is amongst all tweets/documents (IDF). In Table Idepicted the working of TF-IDF approach.

D. Experiments & Results

The TF-IDF values are calculated using n-grams till trigrams, meaning phrases with 1,2,3 words are used to compute frequencies. We have also calculated cosine similarity among the tweets/data. Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. Cosine similarity helps in measuring cohesion within clusters. We used K-Means clustering algorithm to group tweets into chosen number (say, here five) of groups. The Output on using five groups cluster we obtained is shown in Table II.

In table II we have shown five clusters, with following number of tweets in respective clusters and this shows that most of the tweets clustered in group Id = 1, followed by group ID-4 and so on. Later we computed the tweets in various cluster ID by using proximity to centroid technique and achieved some interesting result which is shown in Figure III-D for each IDs versus tweets.

0	keralaflood, twitter, kerala, rescue, flood	115
1	help, donate, people, please, need	3925
2	chiefminister, minister, relief fund, crore	72
3	children, feet long bridge, bridge rescue, citizen malampuzha, senior citizen malampuzha	257
4	twitterkeralaflood, status, keralafloodrelief, keralaflood, kerala	1824

TABLE II
Five different groups with their respective hashtags

E. Cluster-based & topic modeling

K-means algorithm was used to find out the similar tweets from the dataset. Being popular, K-means algorithm requires the value of K apriori as opposed to hierarchical clustering algorithms

K-means generates the following clusters for the K-value 5.

- Cluster 0: Words: alert district rain affect these affect these district these district
- Cluster 1: Words: announc person maharashtra serious injur decease
- Cluster 2: Words: keralaflood kerala twitter help people
- Cluster 3: Words: will keralaflood keralafloodrelief train kerala
- Cluster 4: Words: chief minist chief minist announc announc crore

F. Topic Modeling Using LDA

Topic modeling is very important step for text-processing as it deduces the theme of texts (in this case tweets) [8]. A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents (in this case tweets). To compare the previous result, we implemented LDA to identify the topic of the tweets.

We performed the LDA on basically five topics and set of words for each topic which is shown in Table III-F.

Topics	Words(Clustered After Topic Modeling)
0	keralaflood, kerala, nation, need, flood, relief
1	keralaflood, twitter, family, rescue, flood, people
2	keralaflood, twitter, kerala, rescue, water, operation
3	keralaflood, road, twitter, ernakulam, near, please
4	keralaflood, twitter, train, kerala, govern, keralafloodrelief

TABLE III
LDA output for five different topics

The above table of topics and words (in this case tweets) was the output obtained from the LDA after topic modeling. The words are associated with each and every topic and they represent certain meanings. Topic 0 is about Help, Topic 1 is about Action, Topic 2 is somewhat similar to Topic 1,Topic 3 is about Affected Location/Geography, Topic 4 is about Support and so on other can be identified , etc.

IV. Conclusion and Future works

It is clear from the above analysis that how powerful a social media is and it can be used very widely to help the mankind and other species. Social media can be harnessed to great effect in times of crisis. Some of the steps which are taken in this article are also adopted by twitter itself to help surrounding community in fighting against disaster, crisis, etc. As twitter has initiated the practice of creating

hashtags specific to individual crises to index tweets easily. Facebook has also taken steps forward to fight commonly against crisis by launching 'Mark Safe' feature to those who have listed a crisis-hit location as their place of residence.

This system will surely help the Government agencies like NDRF, CRPF , Home Ministry and other aid relief agencies in collecting different data to develop analytics capabilities focused on mining Twitter for real-time updates to take meaningful action during the crisis and disaster like flood in Kerala.

The Future tasks that will be implemented in this system will be analysis the sentiments that which sentiment is hatred or against the situation and which are in favor of situation. We will focus on making it more accurate and useful by applying some Neural Network concepts in this project in future. To make this system more powerful and useful, by implementing some technique that can detect non-hashtag words that are relevant for analysis. Using deep learning, the next task will be to identify short conversation messages tweets like 'plz' as please, 'ppl' as people,etc. Also another area is stopping rumors as during the Kerala floods, quite a number of false 'news reports' and 'alerts' circulated on Facebook, Twitter and the mobile messaging application WhatsApp. Machine learning can be employed to check the veracity of social media by comparing contents from actual news reports.

The power of social media will continue to be researched and newer applications will continue to be built to harness its power. Disasters may strike at any time, While prevention from them may not be possible, but it is advisable and better to be prepared for unfortunate eventualities.

References

[1] N. Saharia, "Detecting emotion from short messages on nepal earthquake," in Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on. IEEE, 2015, pp. 1–5.

[2] S. Kedar, S. Owen, C. Jones, A. Donnelan, M. Glasscoe, and R. Duren, "Select technologies and capabilities to improve earthquake resiliency in california," 2016.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of machine Learning research, vol. 3, no. Jan, pp. 993–1022, 2003.

[4] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, pp. 100–108, 1979.

[5] D. Hardt, "The oauth 2.0 authorization framework," Tech. Rep., 2012.

[6] H.-N. Teodorescu and N. Saharia, "An internet slang annotated dictionary and its use in assessing message attitude and sentiments," in Speech Technology and Human-Computer Dialogue (SpeD), 2015 International Conference on. IEEE, 2015, pp. 1–8.

[7] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.

[8] H. M. Wallach, "Topic modeling: beyond bag-of-words," in Proceedings of the 23rd international conference on Machine learning. ACM, 2006, pp. 977–984.