Event Classification and Retrieving User's Geographical Location based on Live Tweets on Twitter and Prioritizing them to Alert the Concern Authority

Sarthak Vage[1], Sarvesh Wanode[2], Kunal Sorte[3], Prof. Dipak Gaikar[4]

[1-3]BE, Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India
[4]Department of Computer Engineering, Rajiv Gandhi Institute of Technology, Mumbai, Maharashtra, India

---***---

Abstract - *Social media has become a place where people meet, greet and share vast information and their personal views. It is a very useful source for procuring vital information in emergency situations. One such platform is Twitter which is a micro-blogging and social networking service where a large number of people from celebrities to common people express their views and opinions and share information. We could use this information to mitigate and extract useful information from messages in Twitter called as tweets. We developed a system takes input as tweets from hashtag provided by the user and categories them into relevant and non-relevant using Naïve Bayes algorithm. These relevant tweets contain data that is crucial for any emergency situation management. These classified tweets are further prioritized by XGBoost algorithm, which determines the tweets that are most important during such situations. In the final stage the user can click on the tweet by which they can see all the key details of the tweet such as user name, user location and other such information to take suitable action on the situation.*

***Key Words*: Twitter, Hashtag, Naïve Bayes, XGBoost, NLTK, API, Geolocation, etc.**

1. INTRODUCTION

Social media provides a platform to discuss real-world events and express one's opinions. Micro blogging systems such as Twitter provide such platform where people come together. By better understanding them we can more effectively filter information as the event unfolds. By using Machine Learning various techniques we can unearth key details which can help in better management of an emergency situation. This information can be used alert the concerned authorities to take suitable actions during distress situations.

The rise of mobile internet in past two years has increased the social activity of people to a large context due to which Twitter has seen a large number of new users from urban as well as rural area. People with political importance also have accounts and post to increase popularity and dominance by expressing their views. Almost all the important officials have their account on Twitter to help people and reach to them socially. Twitter act not only as an awareness platform, but also a place where people can ask for help and send advice during disasters. Twitter API and Tweepy library were used to get the tweets. But most of the data generated on Twitter consists of conversational tweets and tweets used for advertisements often termed as spam-bots, which usually hold no value for helping in emergency situation. Users also post hilarious posts for entertaining purposes. For an official to check for informational content among all these tweets would be a great waste of time. To save their time we categorized tweets into relevant and non-relevant tweets where relevant ones hold key information and non-relevant containing useless conversational tweets. To further ease the process the tweets are classified according to their priority levels for ease of user to get key details.

This system uses Naïve Bayes algorithm which is a stochastic model and belong to a family of simple probabilistic classifiers based on applying Bayes theorem. To classify according to the priority various models are tested among which XGBoost gave the most accuracy which is a decision tree-based Machine Learning algorithm that uses gradient boosting framework. An important aspect is geo -location of the tweet as it is important to know the location of the user in some situations. After categorizing the developed system all such vital information related of the user is provided to the concerned authority.

2. RELATED WORK

The Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery [1] paper has classified tweets and location from the tweets where found. In the second paper tweets are classified into informational tweets and conversational tweets and then informational tweets are used for the disaster response. In this paper they worked on a coding schema for separating social media messages into different themes within different disaster stages such as mitigation, preparedness, emergency response and recovery. They have manually examined 10,000 tweets generated during natural disaster phases. They used the Hurricane sandy which struck in northwest US as their training dataset and only used geo-tagged tweets, from these tweets they filtered out the tweets related to storm, hurricane and where categorized into different themes as above. They manually annotated subset of tweets into categories for the training and testing and others through an automatic program. They pre-processed the tweets and a logistic regression classifier is selected to train and automatically categorize the messages in the predefined categories. Ten-fold cross-validation technique was used to test the model. The classification results are necessary and useful for emergency managers to identify the transition between phases of disaster management, the timing of which is usually unknown and varies across disaster events, so that they can take action quickly and efficiently in the impacted communities. Information generated from the classification can also be used by the social science research communities to study various aspects of preparedness, response, impact and recovery.

In Valuable Information from Twitter During Natural Disasters [2] paper the tweets are classified as a binary classification problem, considering the fact that a tweet should be either informative or not. They manually labelled 1086 tweets for the classification task with help from students from our research labs. This set of tweets contains 139 informational and 943 conversational tweets. They used parameters URL ex- traction, Emoticons, Instructional keywords, Phone numbers and Internet slang sentence. They used 10-fold cross-validation for evaluation, using Naive Bayes classifiers as implemented in the Weka toolkit. The results of the designed feature sets were compared with the outcomes of the —bag of word, along with the results of a combined result set of —bag of words and the designed feature set. Precision, recall, and F-measure were reported for each class.

3. PROPOSED SYSTEM

In this paper the proposed system is all about creating software which could identify and filter out the tweets from people who need help or people who want to report important information on Twitter. Since the system is all for reporting these informational tweets, the main part is to distinguish between conversational tweets and informational tweets using Naive Bayes Classifiers. After classification we set the priority of the tweet and relay the message to the concerned authorities with geo-location if needed for immediate appropriate action.

3.1 Design Details

When the software is executed it starts collecting live tweets until the software has good internet connection established. The software through Tweepy library continuously extracts live tweets from Twitter API. The tweets are stored in JSON format with additional information like user id, Geo-location, time, date, retweet count, and number of likes, URLs many such parameters which could be useful.

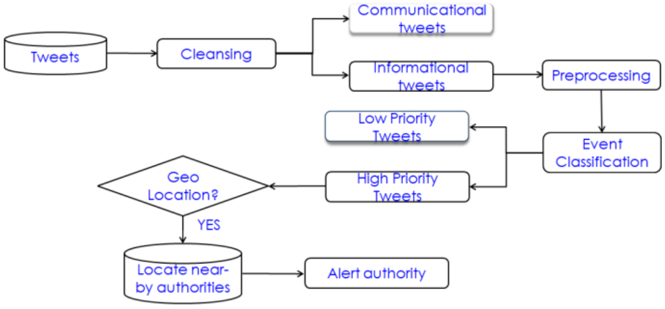


Fig -1: System architecture

The informational tweets are separated from conversational tweets. This segregation is done through Naive Bayes Classifiers. While some elements in tweets like Emoticons, Internet Slang, URLs are factored as some parameters along with important words such as 'help', 'save', fire' etc., are using a bag of words approach are used to classify these tweets are used as factors for classification.

After extracting tweets which have some valuable information we prioritize the tweets according to their necessity. For example some tweets indicating a fire breakout or some cry for help are more important than tweets like reports, small complaints etc.

The location of the person tweeting can be very important in some cases, like knowing where a person is stuck in the case of flood or any such calamity can be crucial. We check whether the tweet user has enabled the Geo-location on the device. If the device has Geo-location switched on, then it will directly give the user location.

Finally the tweet is forwarded to the concerned authorities, relief teams to take necessary action.

3.2 Methodology

Overall structure of the proposed system is given in Fig-1. The system consists of the following modules, which have been described in the subsequent sections.

- Data set
- Live Data Collection
- Data Pre-processing
- Event Classification
- Event Prioritization

3.2.1 Dataset

The first dataset used contains 10876 tweets. This dataset was used for Naive Bayes model in which the dataset with tweets classified into relevant or irrelevant was given as training and testing input.

Table -1: Dataset-1

Dataset-1	
Tweet	**Class**
Just happened a terrible car crash	Relevant
We're shaking...It's an earthquake	Relevant
#raining #flooding #Florida #TampaBay #Tampa 18 or 19 days. I've lost count	Relevant
Beware world ablaze sierra leone & amp; guap.	Not Relevant
on the outside you're ablaze and alive	Not Relevant

The second dataset was used for prioritization of tweets using XGBoost and Random Forest classifier. This dataset contains 4430 tweets for various important hashtags which acts as a query word for the text input.

Both datasets contained tweets which were unfiltered i.e. they contained the features which were not vital such as emoji's, URLs, user name mentions and other such things to keep the things as realistic as possible for better relevancy of the tweets.

Table -2: Dataset-2

Dataset-2		
Tweet	**Keyword**	**Priority**
Typhoon Soudelor kills 28 in China and Taiwan	flood	1
INEC Office in Abia Set Ablaze - http://t.co/3ImaomknnA	fire	1
I waited 2.5 hours to get a cab my feet are bleeding	bleeding	1
BigRigRadio Live Accident Awareness	accident	0
ACCIDENT PROPERTY DAMAGE; PINER RD/HORNDALE DR	accident	0

3.2.2 Live Data Collection

In order to train and validate the model, sufficient tweets related to an event are needed, which should relate to the realistic scenario of that event. Twitter API was used to capture live tweets. Twitter provides different access rights and credentials after applying for its API. The data collection is done using streaming API of Twitter with Tweepy python library. This study concentrates only on tweets in English language. One of the major problems with data collection from Twitter is that it may contain a lot of irrelevant tweets such as advertisements.

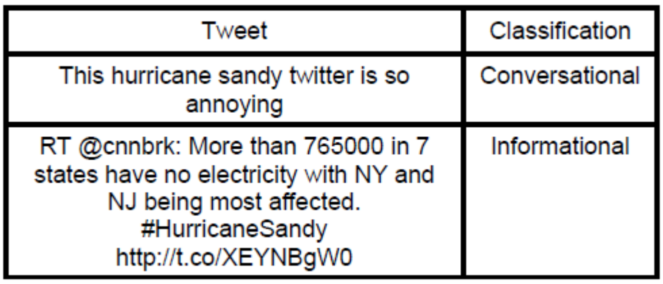
3.3.3 Data Pre-processing

Tweets mostly consists of internet slang, hashtag and links, for training a model we need to remove these things and get a proper English sentence out of the tweet. This is performed with following steps:

- Pre-processing: Here with help of NLTK stopwords and regular expression we remove the punctuation's, special characters, links etc.
- Tokenization: Using NLTK Tokenizer we tokenize the sentence in list of words of a sentence, duplicates are also removed from the sentence.
  Tweet for Pre-processing:
  "@person1 retweeted @person2: Corn has got to be the most dellllicious crop in the world!!!! #corn #thoughts..."
  Pre-processed Sentence:
  "AT USER rt AT USER corn has got to be the most delicious crop in the world corn thoughts"
  Tokenized Sentence:
  {"corn", "most", "delicious", "crop", "world", "corn", "thoughts"}

3.3.4 Event Classification

Hashtags and keywords in tweets help us extract tweets related to a target event. Here we classify informational tweets from conversational tweets. For event classification evaluation 10-fold cross-validation for Naïve Bayes classifiers is used. Naive Bayes classifiers, a family of classifiers that are based on the popular Bayes' probability theorem, are known for creating simple yet well performing models, especially in the fields of document classification and disease prediction. The results of the designed feature sets are compared with the outcomes of the "bag of word". The following result was found in previous researches.



Tweet	Classification
This hurricane sandy twitter is so annoying	Conversational
RT @cnnbrk: More than 765000 in 7 states have no electricity with NY and NJ being most affected. #HurricaneSandy http://t.co/XEYNBgW0	Informational

Fig-2: Informative and conversation

- Building the vocabulary: A vocabulary in Natural Language Processing includes all the words resident in the Training set we have, as the model can make use of all of them relatively equally. This is just creating a list of 'all words' we have in the Training set, breaking it into word features. Those 'word features' are basically a list of distinct words, each of which has its frequency (number of occurrences in the set) as a key.

- Matching tweets against our vocabulary: In this step go through all the words in our Training set (i.e. our word features list), comparing every word against the tweet at hand, associating a number with the word following:
  Label 1 (true): if word in vocabulary is resident in tweet
  Label 0 (false): if word in vocabulary is not resident in tweet

- Building our feature vector: Using NTLK the actual feature extraction from the lists is done.

Now with help of NLTK we train the Naïve Bayes model.

3.3.5 Event Prioritization

Classified tweets need to be prioritizing further to get actually tweets of people who need help in any situation. The prioritization of tweets depends on 5 to 6 major factors. The factors are following:

1. Total no of verbs in the sentence
2. No. of words before verb
3. No. of words after verb
4. Main verb
5. Hashtag

Prioritization is done with XGBoost and Random Forest classifier. For passing the verbs and hashtag word encoding is done for passing it to the classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. Both the models performed well for giving the desirable results, but XGBoost performed better than random forest giving the maximum accuracy.

3.3.6 Alert Message

Once the tweet is prioritize, the twitter API gives the meta data about the tweet which include the Username, UserID, Tweet count, Retweet count, location, Geolocation (Co-ordinates). From this information the exact co-ordinates of the tweet is found. This location is used to get all the authorities and facilities present from the database. A database of all the authorities, health care facilities, NGO, etc with their location is maintained. This system has ability to alert all the relevant nearby authorities with the exact location co-ordinate of the tweet if the user has GPS enabled on the device and if not the approximate location of the user is provided.

4. RESULT

The Naive Bayes classification used for distinguishing between relevant and irrelevant tweets recorded 91% accuracy for training dataset and 90% on testing dataset where 80% of the dataset was reserved as training dataset and rest 20% used as the testing dataset for the model. In the prioritisation stage different models were tested. The models used were SVM, Random Forest and XGBoost. SVM gave very low accuracy for any relevancy and hence is not mentioned. Random Forest gave 68.51% accuracy whereas XGBoost provided accuracy of 79.8% from confusion matrix. Both were given same five inputs as stated in the methodology of event prioritisation. Currently the developed system is capable of finding out all the relevant and prioritised tweets and provides all the important information related to that specific tweet. There also exist a notify button for those tweets to inform the concerned authority which sends a text message containing the tweet to them. For now due to Twitter API restrictions the sending information to such organisations is not allowed. In future more dialog with Twitter we can gain further access to API and also more association with such organisations will help in emergency situations.

5. FUTURE SCOPE

Geo-Location:
Getting the exact location of the user at the time sending the tweet depends on whether the user has enabled GPS in the device or tagging the location information manually while sending the tweet. For procuring the exact latitude and longitude the user needs to opt in for that feature in the app itself. But many users according to our observation haven't opted for such features and we thus get only the name of the place/city rather than the exact location. However, developing such a feature which could predict the location of the user by using his or her previous history or the predicting by analyzing the location of the friends/followers of the user the can be developed. This would further help to increase the effectiveness of the system.

Access to more resources:
Currently we have access to only two twitter API's. The system latches onto one twitter server and transmits those tweets to the software. Having more access to API's would help us to latch on to the many different servers and increase the flow of incoming tweets.

Collaboration with government bodies:
Collaborating with the government organizations would add relevancy to the system as such information could prove to be useful during emergency situations. Also collab- oration with hospitals and NGO's would be helpful.

6. CONCLUSIONS

Social media messages are rich in content, capturing and reflecting many aspects of individual lives, experiences, behaviors, and reactions to a specific topic or event. Therefore, these messages can be used to monitor and track geopolitical and disaster events, support emergency response and coordination, and serve as a measure of public interest or concern about events. In our work, we designed novel features for use in the classification of tweets, during emergency situation in order to develop a system through which informational data can be filtered from conversation on Twitter. The informational data is extracted from user through Naive Bayes classifiers and given to the prioritizing algorithms implemented for creating the priority of tweets according to their importance and necessity, if the geo-location is available then we pass this information to the concerned authorities. The web-based application provides with a user interface where user can enter keyword related to the situation, also the applications runs live to show the priority and classified tweets dynamically. The user gets an option to view the details of the tweet. The application is capable of sending alert signals via different mediums. This application can be used by the authorities such as police, hospitals, local government offices, NDRF teams, relief camps and other such organizations. The use of Twitter and such social networking platforms is bound to increase in the future and usefulness of the project is certainly going to increase too.

REFERENCES

[1] Qunying Huang and Yu Xiao, “Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery” ISPRS International Journal of Geo Information, USA, 24 August 2015.

[2] Brandon Truong, Cornelia Caragea, Anna Squicciarini, Andrea H.,Identifying “Valuable Information from Twitter During Natural Disasters”, 2014 Proceedings of the 77th ASIST Annual Meeting, Seattle, WA, USA.,2014.

[3] Jyoti Prakash Singh, Yogesh Kumar Dwivedi, Nripendra P. Rana, Abhinav Kumar, Kawaljeet Kaur Kapoor, “Event classification and location prediction from tweets during disasters” 2017 SpringerApplications of OR in Disaster Relief Operations, 19 May 2017.

[4] David M.Neal, “Reconsidering the phases of disaster International journal of Mass Emergencies and Disaster”, Vol 15 No.2, pp 239-264 August 1997.

[5] Adedoyin-Olowe, M., Gaber, M. M., Dancausa, C. M., Stahl, F., Gomes, J. B., “A rule dynamics approach to event detection in twitter with its application to sports and politics. Expert Systems with Applications”, Volume 55, 15 August 2016.

[6] Ajao, O., Hong, J., Liu, W. (2015), “A survey of location inference techniques on twitter”, Journal of Information Science, pp 855–864 2015.

[7] Al-Saggaf, Y., Simmons, P. (2015), Social media in Saudi Arabia: “Exploring its use during two natural disasters Technological Forecasting and Social Change”, Volume 95, June 2015, Pages 3-15.

BIOGRAPHIES

SARTHAK VAGE
Currently pursuing under graduation final year in the branch of Computer Engineering at Rajiv Gandhi Institute of Technology, Mumbai.

SARVESH WANODE
Currently pursuing under graduation final year in the branch of Computer Engineering at Rajiv Gandhi Institute of Technology, Mumbai.

KUNAL SORTE
Currently pursuing under graduation final year in the branch of Computer Engineering at Rajiv Gandhi Institute of Technology, Mumbai.

PROF. DIPAK GAIKAR
Assistant Professor (Computer Department), Rajiv Gandhi Institute of Technology, Mumbai