CHAPTER

9

PRODUCING BETTER DISASTER MANAGEMENT PLAN IN POST-DISASTER SITUATION USING SOCIAL MEDIA MINING

Sounak Sadhukhan[*], Soumya Banerjee[†], Prasun Das[*], Arun Kumar Sangaiah[‡]
[*]Indian Statistical Institute Kolkata, India [†]BIT Meshra, India [‡]School of Computing Science and Engineering, Vellore Institute of Technology, Vellore, India

9.1 INTRODUCTION

Today, disaster is very common phenomenon to the human society. Losses in disasters have shown growing trends in terms of death troll and property damage throughout the world due to urbanization, increasing population and increasing degradation of environment. People become more and more vulnerable to the disasters, including earthquake, flood, tsunami, cyclone, droughts, landslide, terrorist attacks etc. In these cases, not only the direct damage to the society has been huge but also the long term economic impact is more difficult to estimate.

Eliminating the threats of disaster is not possible, but it is possible to reduce the effect of disaster through proper plans and precautions. Disaster management does the same thing. It goes through a number of phases namely prevention, mitigation, preparedness, response, and recovery. In disaster situation, time is very precious. So, lot of processes should be running in parallel to save time, like, evacuating affected people, finding out affected areas and distributing relief items, obtaining relevant information from different areas and distributing those information among the appropriate parties, which as a whole forms a big task. Relevant information are always helpful to make decisions about where to pay attention during these above-mentioned phases. However, attaining reliable, accurate and timely geo-spatial data is a challenge, especially in situation where disaster develops rapidly.

Twitter, a micro-blogging site, is the largest one among the different social media platforms. Twitter has become a popular platform for communication in recent years. Twitter user shares news, pictures, and observations. To overcome the challenges faced during disasters, Twitter data analysis can play major roles. In twitter, average citizen reporting on activities "on-the-ground" during disaster is realized as increasingly valuable. Tweet contains valuable information that can help to improve the quality, speed and efficiency of disaster response. So, mining the tweets is very much effective in case of emergency management which includes rescue operation reallocation, relief material distribution and medical help etc. Responders can collect tweet streams generated by social media at the time of disaster, analyze these unstructured data to find out the status of the affected areas and the devastating nature of the disaster and also predict the expected loss due to the disaster. But, tweet stream not only bears relevant information but also carries irrelevant data and noise.

Computational Intelligence for Multimedia Big Data on the Cloud with Engineering Applications. DOI: 10.1016/B978-0-12-813314-9.00009-8

In our study, we have explained the mechanism to extract helpful information for disaster management from a bulk amount of tweets in the disaster situations. To reach our goal we have completed few tasks, like, processing the Twitter messages, removing stop words and stemming. After that, tweets are labelled to classify information and then we build up models to classify the tweets. We have also developed a voting classifier which is the amalgamation of decision tree, support vector machine, logistic regression, Naïve Bayes, and stochastic gradient descent. Then we compare the classification accuracy of each and individual classifier for bag-of-words vectorization and Tf-Idf vectorization for tweets. After classification, our aim is to extract information from the informative tweets, especially those which are very crucial for the responders.

This paper is organized as follows. In Section 9.2, we explain the background of the work, followed by data description in Section 9.3. Section 9.4 describes the text mining process and explains each of the steps of this process. Section 9.5 illustrates the developed classification algorithms and comparative discussions on the results of the study. The information extraction process is described in Section 9.6 and Section 9.7 draws the conclusion.

9.2 LITERATURE SURVEY

In recent years, research on social media analytics and Twitter data in times of disasters has gained lots of interest. Researcher uses Twitter data in times of disasters to focus light on the role of information creation and sharing on social media platforms to create situation awareness (SA). [1] illustrated, though SA has often been analyzed from a person's point of view but it can also be aggregated at the group level. Twitter is facilitating group level SA that can be used by disaster management agencies to develop an understanding about the situation on the ground. The status updates by an individual lead to collective level activity. In a certain situation, to take any good decision from the accumulated information and understanding about what is going on are important [2], especially in disaster situation which requires immediate action to respond. This collective level information sharing enhances the SA in a broader prospect and helps emergency organizations act as per the situation.

Currently, research is focusing on employing automatic approaches to extract the situational awareness information from Twitter data during disasters. [3] had differentiated in original and secondary information synthesized from prior sources where individuals not only shared information about their situation but also re-tweeted the available and useful information. [4] had discovered that the tweets contain situational news used to create situational awareness. The author explains it through case study of Oklahoma wildfire, 2009. [5] illustrated that the prime concern of social media always seems to share information of the situation rather than sharing opinions and thoughts irrespective of the nature of the disaster. Police shooting case in Seattle revealed this truth. [6] illustrated that Twitter is not only an important information source or a platform which develops situational awareness among its users but it can also be useful for disaster management agencies to actively manage a crisis. When an earthquake struck Yushu in China, 2010, the disaster-affected people not only shared information about the situation on the ground but also made the social media as a platform to seeking help or for showing empathy with those who suffered most losses. [7] had shown, Twitter is not only used by affected people asking for help, but it can also be a helpful medium to local authorities to manage the situation. They use social network analysis to find out few kinds of pattern, like, behavior of the community, effectiveness of the online users, types of information shared during the Australian floods in 2010–2011.

The study used social network measures like global centrality and between-ness for the recognition of individuals and clusters when the flood occurred in 2010–2011 in Queensland. Analyzation of geo-location and specific keywords of tweets have been done by [8] for making the responders aware about the situations in time of disasters. Tweets generated during Hurricane Irene were examined by [9] with the help of classification models using machine learning techniques showing the correlation between the Tweets and the peaks and the location-dependent nature of Tweets during the period when hurricane occurs. Extraction of valuable "information nuggets" from disaster tweets during Tornedo stuck in Joplin in 2011 have been done by [10]. They have used machine learning techniques to classify tweets and then extracted self-contained information items like, "caution and advice", "casualties and damage", "donations of money, goods or services", "people missing, found or seen", and "information source", which are very important for disaster management. In another study [11], the authors have used conditional random fields to extract damages and casualty information from the tweet. [12] also claimed as same as [5] when Twitter messages related to the Boston marathon bombing case were analyzed. [13] developed an automated tool "Tweedr". At the disaster situation, this automated tool extracted relevant information for first responders from tweets. The tool has three phases. In the first phase, the data are classified using the classification algorithm. Next, the classified tweets are clustered to merge tweets that are of same types and then applied information extraction to token and phrases that are related to losses and casualties.

9.3 DATA DESCRIPTION

9.3.1 STUDY EVENT

Due to the annual northeast monsoon, the Coromandel Coast region of the South India including states like Andhra Pradesh, Tamil Nadu, and the union territory of Pondicherry experienced heavy rainfall from November, 2015 to December 2015 badly-affecting Chennai. According to the government record, around 500 people lost their lives and number of displaced was over 1.8 million. Damage and losses due to flood ranged from US$7 billion to US$15 billion. With the rain staying from 8th November to 14th December, 2015, the flood caused by El Nino phenomenon was recognized as the most expensive natural calamity of the year in India. As an impact of the heavy rainfall, the water reservoirs in Chennai started overflowing in an uncontrollable manner, causing inundation in the surroundings, power-cut, disrupted communication and limited access to rudiments [14].

9.3.2 DATA COLLECTION

Tweets were harvested from Twitter using a Twitter Search API to obtain publicly available tweets and python script created for the purpose and using a combination of hashtags (e.g. #chennaiflood, #ChennaiFlood, #ChennaiRains, #chennairains). We have downloaded 7,318 unique tweets for the period from 1st to 6th December 2015. But Twitter has some issues with tweet scrapping rate limit. For registered users, the scrapping rate is fixed to 350 requests per hour and for anonymous users, it is 150 requests. Thus, we here collect only a subset of all the tweets which may or may not be the good sample to represent the population of tweets during the flood period.

Table 9.1 Tweet categories and description

Category	Sub-category	Description
Information to the responder	Casualties and damages	Reports about the death toll and damages of public/private structures.
	Status of weather	Reports the weather forecast related news.
	Current situation of any place	Describes situation of a particular area.
	Water logging, road closed	Information on waterlogging and closed roads.
	People seeking help	Reports if any kind of help needed.
	Complain	Reports if there is any complaint regarding rescue, relief item distribution, unavailability of basic commodity services.
Information to the people	Offering help	People offering food, shelter, clothes.
	Donate money	Speaks about donating money for victims and relief funds.
	Distribute help line numbers	People distribute different emergency help line numbers.
	Rescue operation	Spreads the news of rescue operation.
	Others (traffic news etc.)	Spreads other kinds of helping information.
Personal and emotional		Expresses emotions or the information that do not help the responder and the people beyond his/her immediate circle.
Irrelevant		Messages that are not in English or not understood or not related to flood.

9.3.3 DISASTER TWEET ONTOLOGY

Messages generated during a disaster are excessively varied. The system needs to find the informative tweets that can help the responder in disaster situation to build up better rescue and relief operations at disaster situations. Tweets generally are separated into three main classes: Informative: If the tweet provides valuable information that is of interest to other people beyond the author's group. Personal and emotional: if the tweet is only of interest to its author or his/her immediate circle and does not provide any useful information beyond the group, who does not know the author and Irrelevant: if the tweet is not in English or not understood properly [10]. Furthermore, we have differentiated two types of informative messages: information to the responder, i.e. the messages which can help authorities to build up a better rescue, reallocation and relief distribution plan for the people or information to the people, which directly help the people to take decisions. [4] produced a comprehensive list of categories with finding from the disaster-literature and government procedures for disaster response, to analyze twitter messages. Totally she had identified 32 categories to situational awareness. The ontology developed in their study forms the basis for the categories below. The categories of messages we have considered in this study are explained in Table 9.1.

Since not all data can be labelled manually, we sample a small subset. 2,949 of the tweets are selected randomly from the data set and labelled by 10 persons individually and made data set labelled as information to the responder, information to the people, personal and emotional and irrelevant (see also Table 9.2).

Table 9.2 Tweet examples with category value

Tweet	Category
RT @anantha: #Kilkattalai lake has breached. 200 ft Radial Road flooded at Eachangadu signal. #ChennaiRains	Information to the responder
RT @galattadotcom: #RaghavaLawrence charitable trust provides food packets to those affected by rain in Chennai! #chennairains	Information to the people
God has also become intolerance #ChennaiFloods	Personal and emotional
RT @owaistshah: Slap to intolerant Indians. #chennairains #ChennaiFloods	Irrelevant

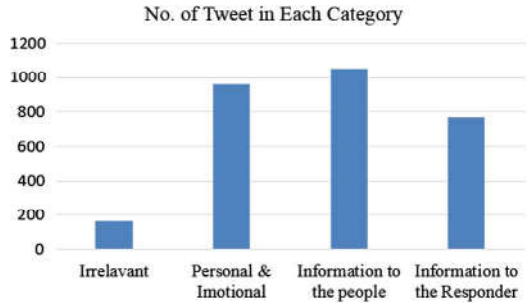

FIGURE 9.1
Quantitative assessment of tweets before merge.

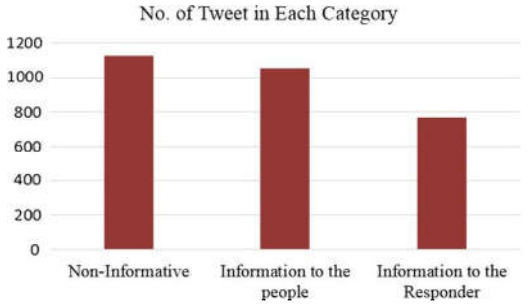

FIGURE 9.2
Quantitative assessment of tweets after merge.

9.3.4 QUANTITATIVE ASSESSMENT

In our data, 26% of the messages were labelled as Information to the responder, 35.77% were labelled as Information to the people, 32.58% labelled as personal and emotional and only 5.62% labelled as Irrelevant. From the plot (Fig. 9.1), it is clearly visible that the data set is not balanced. Our study is concerned about the informative tweets. So, Irrelevant and Personal and emotional tweets are of no use to us. If we merge these two categories, then our intention will not be hampered. We merge these two categories into one new category, Non-informative category. After this activity, the data set becomes quite balanced. The new category contains 38.21% tweets. Fig. 9.2 is showing that the dataset is quite balanced after the merge.

9.4 TWEET CLASSIFICATION PROCESS

Sentiment analysis or opinion mining aims to automatically extract subjective information in the user-generated content. Though, sentiment analysis is normally applied for product review, marketing or business analysis but it can also be applied to help mankind in disaster situations. The objective is the determination of the behavior and point of view of a speaker or a writer about some topics or the contextual schism of a document. The views are based on one's assessment, disturbed mental state

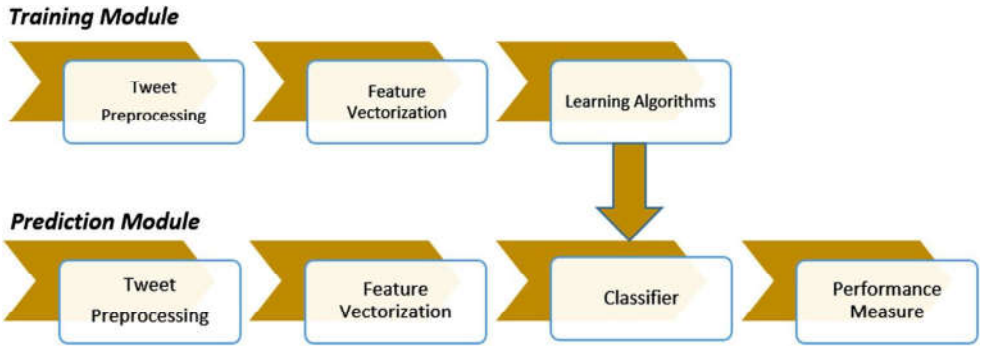

FIGURE 9.3

Tweet classification process.

or the anticipated emotional interaction. But few natural language processing (NLP) challenges pose obstacles in accomplishing this objective.

Tweets are short, heterogeneous, and noisy. A tweet, which is only 140 character long, creates several challenges to traditional NLP, like, (i) Tweets are not always written properly and often full of grammatical mistakes and punctuation errors. Sometimes it can be written in some local languages. Tweets frequently omit articles and auxiliary verbs. They frequently contain abbreviations, short hand messages and local slang etc. Traditional NLP parsers are often unable to process this type of text data. (ii) Tweets contain very little lexical redundancy. The steps of tweet classification are explained in Fig. 9.3.

9.4.1 PREPROCESSING OF TWEETS

Tweet sentiment analysis is a text classification problem. But training a classifier using tweets, is difficult due to large amount of noise like, shortness, emoji, marks, irregular words involved in tweets. So, it needs a bit of preprocessing before it is going to train a classifier.

Removal of Stop Words

Stop words are usually defined as the most commonly used words present in a language, but they are only meaningful in sentences. There is not available a complete list of stop words used by NLP tools. Some of stop words are: ‘a’, ‘the’, ‘is’, ‘are’, ‘which’ etc.

Stemming Word

A grammatically correct sentence is built from different forms of words. Related words with same meaning forms families, which make easier for NLP to search for one of the words and to return documents containing another word in the set. Stemming is a pure heuristic process which cuts down word-ends to achieve this goal efficiently in most of the cases. Stemming also often removes the derivational affixes.

9.4.2 FEATURE VECTOR REPRESENTATION

There are few algorithms available to represent text as a vector form. In our experiment, we have used two approaches, bag-of-words representation and TF-IDF representation.

9.4.2.1 Bag-of-Words Representation

It is one of the easiest ways to represent text as a vector. It ignores the ordering between words in document representation—only the count of words is important. The bag-of-words representation does not bother about the sentence structure or grammatical construction. It builds a dictionary using words available in the text and only counts the frequency of words.

9.4.2.2 Term Frequency-Inverse Document Frequency Representation

Term Frequency-Inverse Document Frequency (Tf-Idf) is a weighted statistical measure to evaluate the importance of a word in a corpus. TF-IDF is not a single method but a class of techniques: term frequency (TF) and inverse term frequency (IDF). The similarity between documents is measured by TF and term importance is measured by IDF. So, the multiplication of Tf and IDF of a word produces the frequency of this word is in the document multiplied by uniqueness of the word. If term frequency is $Tf(w_i, D)$ and document frequency is $Df(w_i)$. Then, from $Df(w_i)$, inverse document frequency $Idf(w_i)$ can be calculated as follows:

$$Tf\ (w_{i,}\ D) = \frac{(\text{Number of times term w appears in a document D})}{(\text{Total number of terms in the document D})}$$

$$Idf\ (w_i) = \log_e \frac{(\text{Total number of documents})}{Df\ (w_i)}$$

The Tf-Idf of feature w_i for document D is then calculated as the product: $Tf\ (w_{i,}\ D)\ .\ Idf\ (w_i)$. The words with high Tf-Idf scored in a document are frequently occurred in that document and deliver the most important facts about the document.

9.4.3 LEARNING ALGORITHMS

Tweet categorization is a text classification problem and can be performed in supervised method, semi-supervised method and unsupervised method. Researchers mainly use Bayesian classifiers, decision tree, K-nearest neighbor algorithms, support vector machine etc. to classify text. In our study, we have studied supervised learning algorithms to classify tweets. Support vector machine (SVM), Naïve Bayes (NB), stochastic gradient descent, logistic regression, decision tree and few ensemble methods like, random forest, extra trees, ada boost, bagging classifier have been used.

9.4.4 PERFORMANCE EVALUATION

After the classification of tweets, performance evaluation is important. In performance evaluation of classifier, the importance is given to the capability of taking the right categorization decisions. Many measures are taken to check the effectiveness of the classifiers. Some of them are precision, recall, f_1

score, accuracy etc.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$f_1 = 2.\frac{\text{Precision} \,.\, \text{Recall}}{\text{Precision} + \text{Recall}}$$

9.5 TWEET CLASSIFICATION ALGORITHMS

In our study, we have done two experiments, one with bag-of-words vector representation and another with Tf-Idf vector representation and made text classification models using machine learning algorithms, like, OneVsRest Classifier, SVM with Linear kernel, multinomial NB, Bernoulli NB, Decision Tree, Random Forest, Extra Trees, Linear SVC, Logistic Regression and then compared the result of these two cases. For each classifier, the data set is randomly split into 70:30 ratio. 70% of data are used for training purpose while 30% of data are used to test the classifier performance. We have done the above step ten times for every classifier and measured the precision, recall, f_1 and accuracy for every time. After that, mean of precision, recall, f_1 score and accuracy are calculated. The algorithm is illustrated in below.

Tweet_CLassi cation Algorithm:

Step 1: Clean the tweets using the following steps:

1. Remove all the inconsistencies between letters and make sure that all the letters are in lower case.
2. All the non-ASCII characters (emojis) are removed.
3. Website Links are removed from the tweet.
4. 'rt', new line character ("\n") are removed.
5. Particularly "#chennaiflood", "#chennairains" are removed, because these are very common hash tags among these tweets. But, other hash tags remain the same, as hash tags represent emotion.
6. "@ user name " are removed
7. Only English letters are taken, rest of the symbol and numbers are removed from the tweets.

Step 2: Preprocessing has been done by removal of stop words and applied Stemming.

Step 3: ClassifierSet = OneVsRest Classifier, SVM (Kernel = Linear), Multinomial NB, Bernoulli NB, Decision Tree, Random Forest, Extra Trees, Linear SVC, Logistic Regression .

Transform the tweets into vector representation (using either bag-of-words or Tf-Idf representation).

Count = 0

Step 4: For each Classifier in ClassifierSet

While Count 10

Randomly split the data set into 70:30 ratio as training set and testing set.

Table 9.3 Classification results Bag-of-Words vector representation

Classifier Name	Bag of Words Vector Representation						
	Precision		Recall		F1-Score		Accuracy (%)
	Mean	Std.	Mean	Std.	Mean	Std.	
OneVsRest (SVM)	0.707	0.08	0.711	0.04	0.708	0.06	71.93
SVM (Linear Kernel)	0.703	0.06	0.703	0.06	0.702	0.06	71.28
Multinomial NB	0.73	0.08	0.735	0.06	0.729	0.04	73.03
Bernoulli NB	0.706	0.15	0.743	0.05	0.71	0.08	73.03
Decision Tree	0.629	0.08	0.633	0.05	0.63	0.06	64.23
Random Forest	0.673	0.09	0.68	0.04	0.674	0.06	68.73
Extra Trees	0.686	0.07	0.688	0.05	0.686	0.05	69.58
Ada Boost	0.662	0.11	0.675	0.05	0.665	0.07	68.09
Bagging Classifier	0.541	0.3	0.659	0.11	0.518	0.19	59.27
SGD Classifier	0.698	0.08	0.703	0.07	0.697	0.05	70.62
Linear SVC	0.706	0.07	0.709	0.05	0.707	0.06	71.77
Logistic Regression	0.731	0.09	0.741	0.03	0.734	0.06	74.58

Train the Classifier using training set.

Test the Classifier Using test set and measure precision, recall, f_1 score and accuracy.

Count = Count + 1

Count = 0

Measure mean of precision, recall, f_1 score and accuracy and standard deviation of precision, recall and f_1 score.

Step 5: End

Results of classification algorithms including SVM with linear kernel, multinomial NB, Bernoulli NB, decision tree, random forest, logistic regression and stochastic gradient descent are compared. The algorithm is implemented in Python environment using the scikit-learn library. Among all the classifiers, logistic regression, SVM, OneVsRest Classifier, stochastic gradient descent, multinomial NB, Bernoulli NB come up as the most reliable across several accuracy measurements. Our results are shown in Table 9.3 and Table 9.4. Also, from the tabular representation of result, it is clearly visible that, accuracy of all the classifiers in Bag-of-Word vector representation is always less than Tf-Idf vector representation except two cases of Multinomial NB and Logistic regression. Decision tree, bagging classifier and ada boost are the most inaccurate among them.

9.5.1 VOTING CLASSIFIER

Voting classifier [15–19] is one type of ensemble classifiers, developed through amalgamation of similar or conceptually different classifiers to categorize data with more accuracy at the expense of increased complexity. This type of classifier can be suitable for similar models in terms of performance to diminish the effect of their weak points. The Voting classifier is implemented as “hard voting” and “soft voting”. In case of “hard voting”, categorization is done depending on the majority of votes of classifiers. In “soft voting”, prediction is done depending on weighted class-probabilities. In our study

Table 9.4 Classi cation results for TF-IDF vector representation

Classi er Name	TF-IDF Vector Representation						
	Precision		Recall		F1-Score		Accuracy (%)
	Mean	Std.	Mean	Std.	Mean	Std.	
OneVsRest (SVM)	0.733	0.08	0.743	0.03	0.736	0.05	74.66
SVM (Linear Kernel)	0.732	0.09	0.742	0.03	0.735	0.05	74.56
Multinomial NB	0.708	0.14	0.742	0.06	0.712	0.06	72.71
Bernoulli NB	0.706	0.15	0.743	0.05	0.71	0.08	73.03
Decision Tree	0.605	0.09	0.608	0.06	0.606	0.07	62.01
Random Forest	0.668	0.1	0.681	0.04	0.671	0.07	68.51
Extra Trees	0.694	0.08	0.699	0.04	0.695	0.06	70.67
Ada Boost	0.645	0.11	0.659	0.05	0.648	0.07	66.37
Bagging Classifier	0.546	0.29	0.649	0.09	0.526	0.18	59.4
SGD Classifier	0.716	0.07	0.721	0.03	0.717	0.05	72.71
Linear SVC	0.726	0.08	0.733	0.03	0.728	0.05	73.88
Logistic Regression	0.717	0.12	0.743	0.04	0.723	0.06	73.67

we have acquired “hard voting” scheme using OneVsRest Classifier, SVM with linear kernel, multinomial NB, Bernoulli NB, decision tree, linear SVC, and logistic regression.

If $C_1(x)$, $C_2(x)$, ..., $C_n(x)$ be the n types of classification rules, then the combination of these n classifiers C(x) will be, mode $C_1(x)$, $C_2(x)$, ..., $C_n(x)$

Voting_Algorithm_for_Tweet_Classi cation
Step 1: Clean the tweets using the following steps:

1. Remove all the inconsistencies between letters and make sure that all the letters are in lower case.
2. All the non-ASCII characters (emojis) are removed.
3. Website Links are removed.
4. ‘rt’, new line character (“\n”) are removed.
5. Particularly “#chennaiflood”, “#chennairains” are removed, because these are very common hash tags among these tweets. But, other hast tags remain same, as hash tags represent emotion.
6. “@ user name ” are removed.

Only English letters are taken, rest of the symbols and numbers are removed from the tweets.
Step 2: Preprocessing has been done by removal of stop words and applied Stemming.
Step 3: (i) Make a Voting Classifier VC, whose estimators are OneVsRest Classifier, SVM(Kernel = Linear), multinomial NB, Bernoulli NB, decision tree, linear SVC, logistic regression .
(ii) Transform the tweets into vector representation using feature vectorization algorithm.
Count = 0
Step 4: While Count 10
Randomly split the data set into 70:30 ratio as training set and testing set. Train voting classifier VC using training set.

Test the classifier VC using test set and measure accuracy.

Table 9.5 Classification results for Bag-of-Words and TF-IDF vector representation

Classifier Name	Bag-of-words Vector Representation		TF-IDF Vector Representation	
	Accuracy (%)		Accuracy (%)	
	Mean	Std.	Mean	Std.
OneVsRest Classifier	71.77	0.91	74.62	1.13
SVM (Linear Kernel)	71.01	1.15	74.64	0.91
Multinomial NB	73.71	1.03	72.96	1.13
Bernoulli NB	72.9	1.15	72.9	1.15
Decision Tree	64.63	1.35	62.53	1.61
Random Forest	69.14	1.38	69.34	1.34
Extra Trees	69.34	1.02	70.66	1.06
SGD Classifier	70.86	1.6	72.55	0.99
Linear SVC	71.46	1.27	73.91	1.1
Logistic Regression	74.5	1.06	73.68	1.11
Vote Classifier	74.19	1.27	75.13	1.15

Count = Count + 1.

Measure mean of accuracy.

Step 5: End.

In our study, we make another comparative study among the standard classifiers and our newly created voting classifier. The result is shown in Table 9.5. Our table reflects that logistic regression is much more accurate than all the other classifiers. Our voting classifier's performance is also up to the mark.

9.6 INFORMATION EXTRACTION

The aim of Information Extraction (IE) is to analyze unrestricted text for excerpting information regarding pre-determined forms of events, entities or relationships. There are few steps involved to excerpt information from the classified tweets, mainly which are helpful for the responder. The steps are shown below.

Information Extraction Algorithms for Tweets

Step 1: Select only those tweets that are classified as informative to the responder.

Step 2: Clean the tweets using the following steps:

1. Remove all the inconsistencies between letters and make sure that all the letters are in lower case.
2. All the non-ASCII characters (emojis) are removed.
3. Website links are removed.
4. 'rt', new line character ("\n") are removed.

Step 3: Each tweet is tokenized.

Step 4: Data are split into 70:30 ratio. 70% for training purpose and 30% for testing purpose. 'Punkt Sentence Tokenizer' has been used for this purpose.

Step 5: Parts of Speech (POS) tagging is applied on tokenized strings. POS tagging is used for assignment of parts of speech to each word (and other token), such as noun, verb, adjective, etc.
Step 6: Entity is detected from the POS tagged tuple.
Step 7: Relation detection between entities using 'Chunking' is done.
Step 8: End

Using the above algorithm, information can be extracted from classified tweets from disaster response. Coding part has been done in Python framework using nltk package. Some test results are given below.

Case 1:

Testing instance 1: 'students stranded at feet water at kollapakam lalaji memorial school need boat to rescue ph s'

Resulting Tree:

Fig 4: System generated POS tagging tree for testing instance 1.

Case 2:

Testing instance 2: 'drivers are stuck in binny mills without food since yesterday shelp'

Resulting Tree:

Fig 5: System generated POS tagging tree for testing instance 2.

Case 3:

Testing instance 3: 'this guy needs help asap s chennaihelp'

Resulting Tree:

Fig 6: System generated POS tagging tree for testing instance 3.

9.7 CONCLUSION

During disaster, tweets grow exponentially. These huge volumes of tweets are very difficult to analyze and therefore extraction of relevant information is very tough. Using machine learning algorithms, we can automatically select a relatively small sub-set of tweets from the large data sets, thereby causing the information extraction much easier and less time-consuming. In our work, we have demonstrated this idea taking a smaller dataset of around 3,000 tweets. The aim is to extract information from the tweets which would help the responder during natural disasters to be prepared with more planning and reduced effect. In this paper, we have established that machine learning algorithms can excerpt relevant information from disrupted social media. Using classification techniques for classifying tweets and POS tagging for information extraction, we are able to mine informative and possibly actionable information from the tweets in the time of natural disasters. There are a lot of research scopes in future.

Additional information is needed to check the authenticity of the tweets, actual area of the incident, so that responder can response properly.

REFERENCES

[1] C. Seebach, R. Beck, I. Pahlke, Situation awareness through social collaboration platforms in distributed work environments.

[2] W.S.N. Reilly, S.L. Guarino, B. Kellihan, Model-based measurement of situation awareness, in: 2007 Winter Simulation Conference, IEEE, 2007, pp. 1353–1360.

[3] K. Starbird, L. Palen, (how) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising, in: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, ACM, 2012, pp. 7–16.

[4] S. Vieweg, A.L. Hughes, K. Starbird, L. Palen, Microblogging during two natural hazards events: what Twitter may contribute to situational awareness, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, 2010, pp. 1079–1088.

[5] T. Heverin, L. Zach, Microblogging for crisis communication: examination of Twitter use in response to a 2009 violent crisis in the Seattle-Tacoma, Washington, area, in: ISCRAM, 2010.

[6] Y. Qu, C. Huang, P. Zhang, J. Zhang, Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake, in: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, ACM, 2011, pp. 25–34.

[7] F. Cheong, C. Cheong, Social media data mining: a social network analysis of tweets during the 2010–2011 Australian floods, PACIS 11 (2011) 46.

[8] S. Kumar, G. Barbier, M.A. Abbasi, H. Liu, Tweettracker: an analysis tool for humanitarian and disaster relief, in: ICWSM, 2011.

[9] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, J. Rodrigue, A demographic analysis of online sentiment during hurricane Irene, in: Proceedings of the Second Workshop on Language in Social Media, Association for Computational Linguistics, 2012, pp. 27–36.

[10] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, P. Meier, Extracting information nuggets from disaster-related messages in social media, in: ISCRAM, 2013.

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

[12] S. Venkatesan, H. Yadav, H.R. Rao, M. Agarwal, Exploring multiplexity in Twitter—the 2013 Boston marathon bombing case.

[13] Z. Ashktorab, C. Brown, M. Nandi, A. Culotta, Tweedr: mining Twitter to inform disaster response, in: ISCRAM, 2014.

[14] Online, https://en.wikipedia.org/wiki/2015_South_Indian_floods. (Accessed 29 September 2016).

[15] G. James, Majority Vote Classifiers: Theory and Applications, Ph.D. thesis, Stanford University, 1998.

[16] E. Bauer, R. Kohavi, An empirical comparison of voting classification algorithms: bagging, boosting, and variants, Machine Learning 36 (1) (1999) 105–139.

[17] G. Tsoumakas, I. Katakis, I. Vlahavas, Effective voting of heterogeneous classifiers, Machine Learning: ECML 2004 (2004) 465–476.

[18] S. Tulyakov, S. Jaeger, V. Govindaraju, D. Doermann, Review of classifier combination methods, in: Machine Learning in Document Analysis and Recognition, 2008, pp. 361–386.

[19] S. Saha, A. Ekbal, Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition, Data & Knowledge Engineering 85 (2013) 15–39.