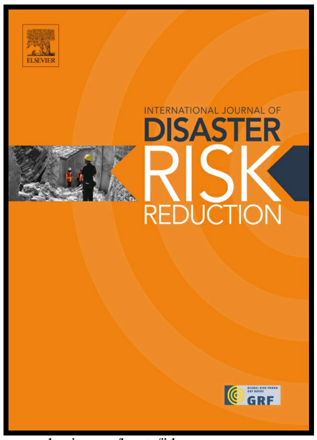
Author's Accepted Manuscript



www.elsevier.com/locate/ijdr

Mining Crisis Information: A Strategic Approach for Detection of People at Risk through Social Media Analysis

J. Rexiline Ragini, P.M. Rubesh Anand, Vidhyacharan Bhaskar

PII: S2212-4209(17)30348-5
DOI: https://doi.org/10.1016/j.ijdrr.2017.12.002
Reference: IJDRR728

To appear in: *International Journal of Disaster Risk Reduction*

Received date: 31 August 2017
Revised date: 6 December 2017
Accepted date: 7 December 2017

Cite this article as: J. Rexiline Ragini, P.M. Rubesh Anand and Vidhyacharan Bhaskar, Mining Crisis Information: A Strategic Approach for Detection of People at Risk through Social Media Analysis, *International Journal of Disaster Risk Reduction,* https://doi.org/10.1016/j.ijdrr.2017.12.002

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Mining Crisis Information: A Strategic Approach for Detection of People at Risk through Social Media Analysis

J. Rexiline Ragini[1], P. M. Rubesh Anand[2], Vidhyacharan Bhaskar[3*]

[1]Department of Computer Applications, Hindustan University, Chennai – 603103, India
[2]Department of Electronics and Communication Engineering, Hindustan University, Chennai – 603103, India
[3]Department of Electrical and Computer Engineering, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132, USA
rexilineragini@gmail.com
rubesh.anand@gmail.com
vcharan@gmail.com
[*]Corresponding author.

Abstract

Situational awareness of the rapidly changing environment in the event of disaster is vital for effective response and recovery management. The major challenges in achieving such awareness are lack of access to various sources of information and tools. Social media plays a vital role in understanding the real situation at the place of disaster as the information is received directly from the affected people. If the collected information is leveraged effectively, the crisis situation can be brought under control and the risks of the disaster affected people or disaster prone areas are reduced. This indeed minimizes the casualty and helps the affected people in serving with their basic needs and medical emergencies. In this paper, we propose a hybrid method for segregating and classifying the texts[1] received from the people who are at risk in the affected region. The proposed hybrid method combines rule based methodology and machine learning algorithms with linguistic features for segregating the texts and classifying them according to the needs. The results of the real-time text classification algorithm help the emergency responders to locate the people at risk and reach them during the hour of their need.

Keywords

Disaster management; Support Vector Machine; Sentiment analysis; Text classification; Social media analysis; Situational awareness

1. INTRODUCTION

The occurrence of disaster is hard to predict but, their effects can be minimized through scientific methods. There is always a chaos and disorganization at the time of the disaster which ruins all the

[1]Text, Message and Tweet are used interchangeably throughout the paper which refers to information received from disaster affected or prone area.

Accepted manuscript

ACCEPTED MANUSCRIPT

good plans that are laid for recovery. During any natural or man-made disaster, accurate information is needed to reach the affected people and create a safe environment for them. In such situations, social media has lots of potential in helping the affected people. It helps the affected people by conveying the necessary information to the rescue personnel [1]. Business organizations have already utilized the power of social media to improve their operations and sales, but the disaster management team and government organizations are yet to utilize this technology to its full potential [2]. The major key elements of disaster management are prevention, preparation, response, and recovery. Social media has the potential to manage all the four dimensions effectively. The increase in the number of mobile phones and Twitter users creates a huge volume of data every day. Especially, social media generates a large amount of data during and after a disaster [3]. This information is considered to be important due to reasons like real-time nature, faster arrival of information, and tends to be accurate as it is received directly from the affected people. Once this information is leveraged properly, the situation can be managed efficiently to help the affected people at the right time.

Social media analysis was first used in crisis response during the 2010 Haiti earthquake where much of the relief work was carried out with the help of tweets and text messages [4]. During the Haiti event, the entire world supported the relief work through the novel use of cyberspace. Though the earthquake damaged most of the communications systems, the internet connectivity system in Haiti was robust since their internet service providers used satellites rather than underwater fiber optic cable. The texts created from Twitter and other messaging services were mainly used by the affected people for conveying information about the current situation. The responders exchanged positive and promising information about the relief operations, which indeed built confidence and courage among the survivors. In 2011, Tohuku-Japan earthquake and tsunami affected people used tweets to mobilize the relief work [5]. During the 2012 Hurricane sandy in the US, an online map of New York and New Jersey was created to locate the places of gas availability [6]. These incidents have triggered Japan, United States of America and other countries to start joint research programs for utilizing social media data in finding workable methods to ease the toll of disasters. On the other hand, crowdsourced data driven initiatives have started to bring all the community people together during their times of need. Previous studies classified the tweets to detect the occurrence of earthquake [7], crowdsourcing information during and after disaster [8], classification of various needs of the people during disaster [9], and various challenges involved in utilizing social media in emergency response [10]. According to recent research, the usage of social media data for the search and rescue operations during any crisis is a challenging task due to the high volume of information collected from various sources and the inability to segregate the information that are received from the affected people [11].

In this paper, a hybrid method is proposed to rescue the people in the disaster affected regions through text analysis. The text classification methodology in the proposed model segregates the texts received directly from the affected people and intimates the appropriate rescue personnel for necessary actions. Linguistic features and rule based methodology are mainly focused as the texts from affected people may not be in proper grammatical format due to the chaotic situation at the place of disaster.

Accepted manuscript

ACCEPTED MANUSCRIPT

The rest of the paper is organized as follows. Section 2 provides a back ground for the current work by reviewing the existing literature on disaster relief using social media. The problem description is explained in Section 3. The proposed model for text classification is explained in Section 4. The experimental results with comparative analysis and discussions are included in the Section 5 and 6. Finally, Section 7 summarizes and provides the conclusions.

2. LITERATURE REVIEW

In the past few years, extensive studies on the use of social media in various fields had been reported, but the usage of social media in disaster relief is still in its inflection point. Chatfield *et al.* [12] studied the pattern in which the affected citizens and government organizations shared the information. The study also highlights the benefits of using social media in emergency response for both affected people and government organizations. Starbird *et al.* [13] and Munro *et al.* [14] had researched on the use of crowdsourcing various information during disaster and mass emergencies. The study highlighted several tools that were developed for disaster relief after Haiti earthquake, including the ‘tweak the tweet’ launched by the authors. Sen *et al.* [15] highlighted the use of crowdsourcing information during crisis and discussed about the various challenges involved in making crowdsourcing as an effective tool for rescue operations. Schimak *et al.* [16] and Zook *et al.* [17] discussed about the various tools that were used during Haiti earthquake along with the power of crowdsourced information and the new ways of interacting between the people who are physically at distant places. The various studies that focused on the information sharing pattern during times of disaster includes Genoa flooding in Italy [18], Tohuku earthquake in Japan [19], Red River Floods that occurred in 2009, and Oklahoma Grassfires in 2009 [20].

Support Vector Machine (SVM) is a supervised machine learning algorithm which is used for classification and regression analysis. SVM has proved to be the best learning algorithm for text classification than any other learning algorithms [21]. Joachims *et al.* had proved that SVM eliminates the need for reducing the higher feature dimension space and has an automatic parameter tuning property which is best suited for text classification [22]. Sentiment analysis is the process of segregating texts into positive or negative texts. Sentiment analysis technique was first implemented by Bo Pang and Lee in the movie review dataset for classifying reviews into positive or negative using the machine learning approaches. The study was concluded by examining the challenges in classifying the text into positive or negative cases [23]. Kamal et.al had used a hybrid method using rule based methodology and SVM which yielded better accuracy on the electronic products over other methods [24]. A novel feature based sentiment analysis model was built by Wang *et al.* [25] using support vector machines for Chinese, this method modifies relationship between word and punctuation. This method produced better classification accuracy than other methods. Sentiment analysis technique was used by Yan *et al.* [26] for predicting Chinese stock markets through local microblogs. SVM and probabilistic neural network were used to predict the stock forum behavior and results revealed that SVM outperformed neural networks. Sentiment analysis has been carried out by Alkalbani *et al.* [27] on cloud customers who utilizes software as a service (SaaS). The results revealed that an accuracy of

Accepted manuscript

ACCEPTED MANUSCRIPT

92% was achieved with 3-fold cross validation. A hybrid method using linguistic features and SVM was proposed by Sharma *et al.* [28] with Chi square and point wise mutual information for selecting the best linguistic features. The study on travel reviews was conducted by Qiang *et al.* [29] for seven travel destinations in and around the US and Europe through the evaluation of machine learning approaches like SVM, naïve bayes, and n-gram model. The study concluded that SVM and n-gram models outperform naïve Bayes approaches. Sentiment analysis was applied in health forums for clustering the documents related to medicine. In this study, various keywords were extracted related to medical terms, including symptoms, treatments, effectiveness, and side effects to form a virtual document for clustering [30].

Some researchers have utilized sentiment analysis technique in crisis domain for detecting the sentiments of the people at the time of disaster. Qadir *et al.* [31] emphasized on the recent developments in using social media during emergency situations. Neppalli *et al.* [32] analyzed the sentiments expressed by people through social media. The results were visualized in a map centered on the hurricane which indicated that sentiment of users changed according to the location and distance from the place of disaster. Verma *et al.* [33] built sentiment analysis model which automatically detected tweets related to situational awareness and categorized them into various categories like, personal or impersonal style, subjectivity, formal and informal linguistic text. Mandel *et al.* [34] used features like bag of words, pruning and lexicons. A sentiment classifier was trained with hurricane Irene dataset for classifying the tweets based on the level of concern. Bayesian network was also used in detecting the sentiment of tweets related to the California gas explosion [35]. Sakaki *et al.* [36] used tweets as an early warning system for earthquake detection. In order to detect the target event, a classifier was built with features such as list of keywords, number of words and relevant context. Along with those features, Kalman filtering and particle filtering were applied to estimate the center of the trajectory. Paul *et al.* had utilized Twitter data to sense earthquake as an early warning system [37]. The Japan earthquake was studied by Vo *et al.* [38] to identify crowd emotions like calm, unpleasantness, sadness, anxiety, fear and relief. Text classification was applied to the Kenya Weatgate Mall attack dataset to understand how the emergency responders use social media data to improve their operations [39]. Torkildson *et al.* [40] had developed a series of emotion classifiers to perform sentiment analysis on the Twitter messages related to 2010 Gulf oil spill. The prototype proposed in the work helped in visualizing the distribution of emotions and their propagation. Lewis *et al.* developed a tool for language translation for most of the minority languages. A cookbook had been developed which can be used during and after a crisis to translate the language [41]. Albuquerque *et al.* researched on the usage of social media along with the authoritative data for identifying the useful information in managing the disaster situation. The results revealed that messages that were received around 10 km surrounding from the affected area were relevant to the incident [42]. Sumalatha *et al.* had developed a web service which helped people to upload the images of the current situation at the place of disaster which in turn sends the information about the nearest relief center [43]. Nazer *et al.* classified the help request related tweets from the disaster data using the content and context of the tweets [44]. Acar and Muraki analyzed the pattern in which people tweeted and the analysis revealed that people who were in the affected area tweeted about their unsafe situation and survival. The other

Accepted manuscript

people who were in the areas far from the affected location post messages about the secondary needs such as transportation and shared information to their relatives and friends that they were safe. The study also revealed that classification of unreliable tweets posed the biggest problem during the times of disaster [45].

Though there are studies that reveal people in the affected area tend to tweet about their unsafe and survival related situations, these studies are ineffective in segregating such texts. This paper concentrates on segregating the texts from the people who are trapped in remote area or struggling for survival during disaster. The text classification is performed through the hybrid method which combines rule based methodology and SVM based machine learning algorithm.

3. EXISTING SCENARIOS AND PROBLEM DEFINITION

3.1 Social media in disaster management

The most challenging task of the rescue personnel is to identify the exact location of the affected people in times of need and emergency. The usage of social media for disaster management shows a promising way of reaching out most of the affected people through geotagging and hash tags (#). The location of the affected people can be easily identified with the geoannotation property of the tweets. Though the number of geoannotated tweets are less, it is likely to grow in the forthcoming years [42]. In some cases, the affected people may be unaware of the contact information for rescue and emergency related queries and help. During such situations, social media like Twitter comes handy for the people to tweet with the disaster name hash tags which indeed reaches out to everyone who involve in the rescue operation. As the needs of the people affected by disaster varies with respect to location, time, stress level and physical health, the disaster management team needs to be updated on the current situation for effective response and recovery.

3.2 Description of the case study and Data sets:

This section provides a brief description of our case study, which is followed by a detailed explanation of the data set that are used in the analysis of Twitter data.

3.2.1 Floods in South Asian Countries:

The disasters that are considered in this research include, India-Pakistan floods that happened in September 2014, a severe cyclonic storm named HUDHUD in October 2014 and another severe cyclonic storm named Nilofar. In September 2014, heavy floods and landslides happened due to torrential rain in the border of India-Pakistan causing severe damages. During the floods in Kashmir, 2600 villages were affected out of which 390 villages were completely submerged in water. In a tweet, "*People of #Pakistan have unfortunately suffered from destructive floods now for the fifth consecutive year*." indicates that the flood occurs almost every year in these areas. During the Cyclonic storm named HUDHUD, the city of Visakhapatnam which is located in South India experienced a huge loss of life and severe damages. Over 2 million families were affected by Hudhud and the avalanche that happened as an effect of the cyclone. The third incident is Nilofar cyclonic storm. The cyclone was

Accepted manuscript

ACCEPTED MANUSCRIPT

formed in the North Indian Ocean and it was about to hit Gujarat state in India. Later, the cyclone got weakened and did not cause much effect. But there was great anxiety among people, which created a lot of hits in social media. The effects and the nature of the disaster are analyzed along with the people's tendency to tweet or text during a disaster. The analysis of the tweets collected from these incidents reveals that the affected people has expressed their day to day problems and the real situation at the place of disaster. The process of categorizing the tweets for helping the rescue personnel is a challenging task as some information in the tweets are hard to predict as shown in Fig. 1.

2014-09-08 14:08 IST	The waters are till the first floor, in some cases higher and this is prime locality in the heart of Srinagar #kashmirfloods.
2014-09-09 02:47 IST	#KashmirFloods How can rescue and relief operations be carried out if hooligans pelt stones at rescuers? J&K politicians must advise locals
2014-09-09 23:00 IST	"I went to the police, I tried to find a private boat, but nothing, there is no help." http://t.co/wtrM3ammEj

Fig. 1. Samples of tweets with mixed feelings of the people.

3.2.2 Data Set Description:

The disaster related data are collected from Twitter database for the above mentioned disasters. The Twitter data collected for the text analysis contained 70,817 tweets. The corpus for India-Pakistan floods, Kashmir floods contained 30,817 tweets, HUDHUD contained 30,000 and Nilofar contained 10,000. A part of these tweets are collected using the Streaming Twitter API. As Twitter allows to collect only the past seven days' data using the Streaming Twitter API, there is no provision to collect the historical data using it. The rest of the data are collected using a third party vendor 'Followthehashtag' [47]. Besides the text message, few other fields that are captured with the tweets include, unique Id for each tweet, the information about the user who created the tweet, hashtags (#), the time stamp and the geo co-ordinates of the location. The frequency of the tweets and the date range in which the tweets are collected is shown in Fig. 2.

Accepted manuscript

ACCEPTED MANUSCRIPT

Fig. 2. Analysis of tweets related to India – Pakistan floods, Nilofar and Hudhud disaster.

3.3 Problem Description:

The most commonly identified basic needs of the affected people are food, water, medical emergency, rescue, shelter, collapsed structure and electricity [9]. The diverse dimensions of people's necessity during any disaster is shown in Fig. 3.

ACCEPTED MANUSCRIPT

Fig. 3. Diversity of people's need during disaster.

The disaster related tweets are categorized using keyword filtering technique [20] which is a common practice in Twitter analysis. The keywords are coined for each category of the identified needs. The keywords are selected by identifying the words that are found more than five times and also relevant to each of the category. The identified keywords are utilized to filter the required data from a large set of text. Table 1 shows the coined keywords pertaining to each category of need.

TABLE 1: SUMMARY OF THE CATEGORY LIST AND ITS ASSOCIATED KEYWORDS

Category	Keywords
Water	Water, drink, thirsty
Food	Food, starve, hungry
Shelter	Shelter, house, living place, sleep
Medical emergency	Medicine, clinic, hospital, medicine, doctor, nurse
Electricity	Electricity, power, electricity, light, fan
Trapped	trapped, stranded, help, save, rescue, struck, caught, risk, hazard, danger, evacuate, critical

The number of tweets filtered with the identified keywords is 6,842. The distribution of the tweets in each category is shown in Fig. 4. This classification and analysis reveals that there are high proportions of tweets from trapped people who are experiencing hard time for their survival. The most important aspect during any crisis is to locate those trapped people and save them.

Accepted manuscript

ACCEPTED MANUSCRIPT

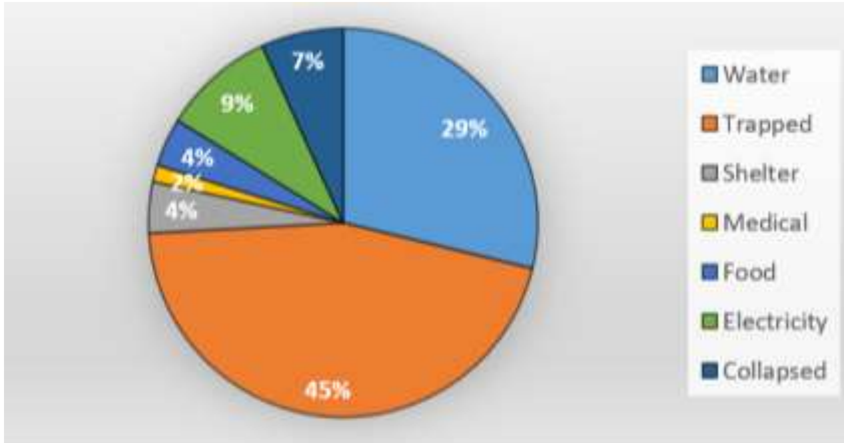


Fig. 4. Distribution of tweets according to various needs of people that are collected from past disaster data.

On further analysis of these tweets, it is identified that not all tweets are from the trapped people who are struggling for survival, but there are tweets from people who narrate the incidents they hear or interpret. These third-party tweets unnecessarily confuse the rescue team and wastes their time and resources. In order to understand whether the text message is from the affected people or not, the texts are classified into two categories namely positive and negative. The positive text is from the affected people and negative means the text that contains the keywords for people trapped but are not actually from the affected people as shown in Fig. 5. The segregation of the tweets only from the affected people is important as the responders need not waste time in relying on other fake information. The rest of all the texts which debates about the disaster should be eliminated.

Accepted manuscript

Examples of positive text:

25 ppl stuck in house no 152 balgarden, near shamshanghat. Pls update and pass msg to rescue team My no 9892910906
house 117/118 jawahar ngr fmly stranded with toddler amp old lady no contact since 2 days masood a shayiq
Please tweet this: My father Mohd Muzaffar Shaw is at Ram Munshi
We are trapped in no 3, Krishna street, please save us.

Examples of negative text:

Ali announces Lakh for Hudhud cyclone relief fund Lets SAVE Vizag
Anger rises in India s Kashmir Valley as people remain trapped a week after floods
Nilofar Not so Far now Saturday is the red alert day god save us
Rescue working round the clock to save thousands caught in Pakistan floods

Fig. 5. Examples of positive and negative tweets that are collected from past disaster data.

The summarized problem definition is that an automated text classification and analysis system in real-time is highly necessary for reducing the causalities. The output of the automated system should also help the responders in building trust, courage, and confidence among the people during the events of the disaster. The primary focus of this paper is to segregate the tweets of the people who are trapped or struggling for survival with highest accuracy during any disaster.

ACCEPTED MANUSCRIPT

4. PROPOSED FRAMEWORK FOR REAL-TIME TEXT ANALYSIS

The proposed framework for text classification of disaster related data utilizes the power of social media in sharing information between the people and the rescue personnel for organizing the rescue operations. The data flow diagram of the proposed model consists of three phases namely, data collection, processing and output phase as shown in Fig. 6. The data collection is the initial phase, which consolidates the data collected from Twitter and other messaging services along with the questionnaire issued to the people for answering about their responses during any crisis or disaster. The data collected through questionnaire are mainly to analyze people's views and words usage when they are in critical situation. In the processing phase, the rules and linguistic features are applied to the input data for text classification. The results obtained from text classification is fed to the machine learning algorithm which utilizes Support Vector Machine (SVM). The various hybrid combinations of rules and machine learning algorithm are executed to finalize the method that yields maximum accuracy. Once the text is accurately classified, the output phase receives the necessary information for further communication. The output phase provides the results in multiple ways. It either sends an alert message to the rescue personnel about the victims or the information is stored for further analysis. The geo annotation property of the tweet allows to plot the location of the classified tweet in a map. The map helps the rescue personnel in reaching the exact place from where the victims tweet.

Accepted manuscript

ACCEPTED MANUSCRIPT

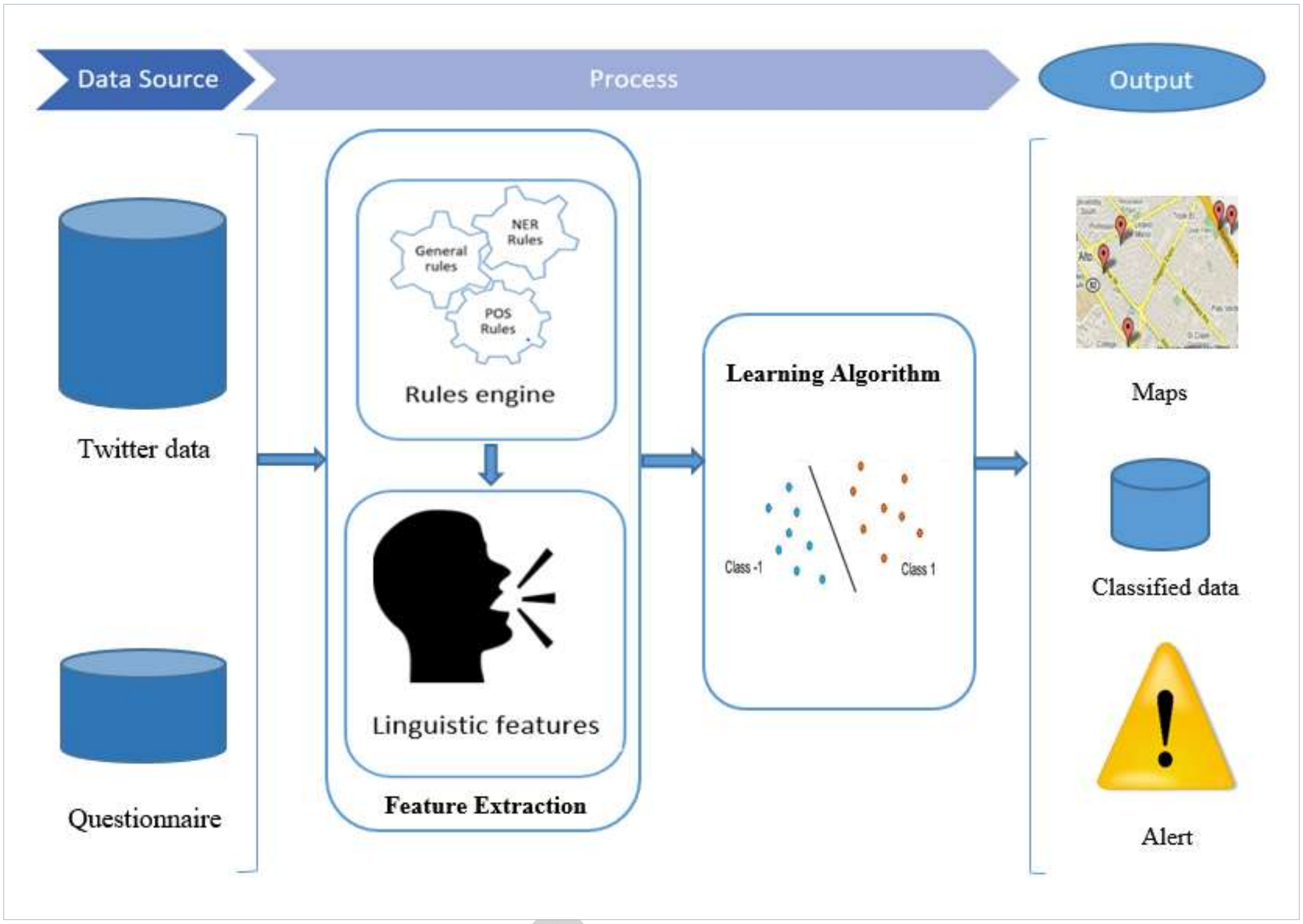


Fig. 6. The proposed data flow diagram for real-time text processing and classification.

The proposed model works in real-time by monitoring the data from social media related to disaster. Once the data is collected related to any disaster or crisis, the model performs the automated text classification and analysis to initiate the rescue operation. The data processing phase with the various classification techniques that are used in the research is dealt in the following sub-sections.

4.1 Data Preparation:

Keyword filtering technique is used in the proposed model for identifying and analyzing Twitter messages containing relevant information [20], [48]. The tweets are initially filtered by keywords that referred to danger or stranded. Tweets containing the keywords such as trapped, stranded, help, save, rescue, struck, caught are retained. Further, few more additional words like, risk, hazard, danger, evacuate, critical which are found common from the web are included. The keyword list prepared from the above mentioned terms is referred to as 'Trapped Keywords'. The filtered number of tweets with the trapped keyword list is 3,553. On manual analysis of these tweets, it is found that 3,521 tweets are not from the trapped people, remaining 32 are from the trapped people

ACCEPTED MANUSCRIPT

who are struggling for survival. Though the tweets from the directly affected people is very less, it is important to segregate these tweets for saving the trapped people. In order to compensate the discrepancy in the number of positive texts, a questionnaire mimicking the disaster environment is framed, and circulated among the university students for their responses. The collected responses are segregated manually to create an additional dataset for text analysis. Since the number of negative text is relatively very high, it is manually analyzed to remove the retweets and the repeated tweets. The resulting corpus consisted of 1593 tweets out of which the total number of negative text is 1199 and the total number of positive text along with questionnaire is 394.

4.2 Feature Extraction

Feature engineering is one of the main components in the text classification system. Identification of right features improves the classification accuracy of the machine learning algorithms. The precision of the text classification mainly depends on the distinctive features that are considered for disaster related data.

4.2.1 Rule Based Feature Extraction

Rule based methodology is one of the common feature reduction technique. The rules are used for segregating the tweets from the disaster affected people. The rules help to annotate the text as positive or negative. The proposed rule based classifier has three divisions namely, general rules, POS tagging based rules and NER tagging based rules to cover all the combinations of positive text such that the data is not biased.

4.2.1.1 General Rules:

General rules are essential as affected people do not always use structured sentences due to the panic situation that prevails in the disaster affected area. In general rules, the tweets with smileys (related to danger, sad or cry) and acronym (SOS (Save Our Soul), ASAP (As Soon As Possible), ppl) are categorized as positive text. The general rules are built with the perception that people often use the acronyms to type the messages faster, and also use smileys to explain the situation better in times of distress.

4.2.1.2 POS Tagging based rules:

Parts of speech (POS) Tagging is the process of assigning the parts of speech to each word in a sentence as a noun, pronoun, verb, adverb or an adjective. The tweets are preprocessed and POS tagging is carried out with the Penn Treebank POS tag set to generate a set of personal pronouns as in Algorithm 1. Once the personal pronouns are identified, the POS tagging rule is built by replacing all the personal pronoun (PRP) with @ symbol forming a set of antecedents. When this antecedent is followed by any of the keywords such as save, rescue, and help, the tweet is classified as positive. The next rule is framed by replacing all the Trapped keywords in the tweets with a % symbol. When the %

Accepted manuscript

ACCEPTED MANUSCRIPT

symbol is followed by a preposition or subordinating conjunction (IN), the tweets are classified as positive rest of all the tweets are classified as negative. Table 2 summarizes the terms and definitions used in the algorithms.

TABLE 2: TERMS AND DEFINITIONS

Term or Notation	Definition
T_{ID}	Tweet ID
T+	Positive Tweet
T-	Negative Tweet
K	List of Keywords
N	Total number of tweets
TK	Tweets filtered with Keywords
M	Total number of filtered tweets
NER	Named Entity Recognition
POS	Parts of Speech
TP	POS tagged tweets
TN	NER tagged tweets
P	Total number of positive tweets
Q	Total number of negative tweets
KR	Keywords related to rescue or help
PRP	Personal Pronoun
IN	Preposition or subordinate conjunction
L	Place or location of the victim

Algorithm 1: Text Filtration and Tagging

```
Input: T, K;
Output: TK, TP, TN;

Let T[1...n], K[1...m], TK[1...s], TP[1...s] and TN[1...s] be new arrays
s ← 1;                                  // Initialization of increment counter //
    for i =1 to n do
        for j = 1 to m do
            if (T[i] = K[j]) then
                TK[s] ← T[i];
                s ← s+1;                // Incrementing the counter //
            end if
        end for
    end for
                // Apply POS and NER Tagging //
```

Accepted manuscript

```
    for i =1 to s do
        TP[i] ← POS(TK[i]);
        TN[i] ← NER(TK[i]);
    end for
```

4.2.1.3 NER Tagging based rules:

Named Entity Recognition (NER) tagging is the process of locating and classifying named entities in text into the categories of names of persons, places, locations, time, date and values. The process of NER tagging is performed through Stanford Named Entity tagger for categorizing the names and locations in the texts. Once the NER tagging is applied, the names and the locations in the entire texts are replaced by $ and @ symbol. In a text, if $ is in first occurrence and followed by any Trapped keyword, then the text is marked as positive. In case, if @ is followed by a preposition and then by a Trapped keyword, the text is marked as positive. Algorithm 2 details the application of various rules to classify the text as positive or negative.

Algorithm 2: Classification of Tweets into Positive or Negative

```
Input: TK, TP, TN, KR, PRP;
Output: T+, T-;
                    // Filtration of POS Tagged Tweets //
Let T+[1...p] and T-[1...q] be new arrays
for i =1 to s do
    if (TP[i](BiGram[KR, PRP])) = True) then        // BiGram linguistic feature is applied //
        T+[i] ← TP[i];
    else if (TP[i](BiGram[KR, IN])) = True) then
        T+[i] ← TP[i];
    else
        T-[i] ← TP[i];
    end if
end for
                    // Filtration of NER Tagged Tweets //
for i =1 to s do
    if (TN[i](Rules[L, KR])) = True) then        // Rules applied for location and keywords //
        T+[i] ← TN[i];
    else
        T-[i] ← TN[i];
    end if
end for
                    // Filtration of Tweets based on general rules //
for i =1 to s do
    if (TK[i](General Rules)) = True) then
        T+[i] ← TK[i];
    else
        T-[i] ← TK[i];
    end if
```

Accepted manuscript

ACCEPTED MANUSCRIPT

end for

4.2.2 Linguistic Features

There are two types of features namely, statistical features and linguistic features that are utilized for text classification systems. According to various studies, it is evident that linguistic features provide better accuracy than statistical features [49]. During any crisis, the affected people are in panic and the messages they convey as text may not be in proper format. Hence, the application of linguistic features provides better results in text classification than statistical features. Though there are various linguistic features like, negation, bag of words, character n-gram, word n-gram, adjective, adverb, punctuation, emotion, aspect based and sentiment lexicons, this study utilizes bag of words and word n-gram as the rest of the features has less impact on the classification task of disaster data.

4.2.2.1 Bag of words:

In text classification, bag of words is the most commonly used method. In this method, the occurrence of each word is considered as a feature to train the machine learning algorithms. In bag of words model, a document is represented as a bag that is a multiset of its words without considering the grammar or word order instead only preserving the multiplicity. A word cloud is an image which contains the list of words used in a particular context, with the size of the word indicating the frequency of the word in the particular corpus. The different words considered in the bag of words model is represented as word cloud is shown in Fig. 7.

Fig. 7. Word cloud representing bag of words in disaster data.

4.2.2.2 N-grams:

N-gram is the extension of the bag of words model. It is a set of co-occurring words with a given window size. These are a set of adjoining sequences of *N* items in the given text. The N-gram model stores spatial information in the text. The length of *N* can be varied from 1 to *n*. In this study, *n*

Accepted manuscript

ACCEPTED MANUSCRIPT

is limited to 3. When N = 1, it is called as unigram. N = 2 is known as bigram, and N = 3 is known as trigram. Unigram is the order less text representation, where only the words are considered. In bigram, the words are considered along with the adjacent words. Trigram considers three words which means two words before and after any word are considered. The results of bag of words and N-gram features are provided as input to the machine learning algorithm for classification.

4.3 Machine learning algorithm for text classification

A machine learning algorithm is required to map the extracted features to the text that need to be classified. Support vector machine (SVM) is an artificial intelligence algorithm which works based on the decision boundaries. A decision plane separates the input into various classes based on class membership. Though there are various supervised machine learning algorithms available, SVM is used as it is proven to perform better than any other algorithms in text classification problems [21]. SVM is a supervised classification technique that performs regression and classification tasks by constructing nonlinear decision boundaries. For a given category C = {S_+, S_-}, where S_+ is a set of positive samples and S_- is the set of negative samples. The SVM calculates and constructs a hyper plane or set of hyper planes in the dimensional space that divides the data into sets with a maximum margin. During preprocessing, the entire text is converted into vector X_i consisting of a set of features that represents the corresponding disaster related data. The SVM algorithm calculates and plots a hyperplane through supervised learning that divides the positive and negative texts with a maximum margin. The classification problem is defined as in which side of the hyperplane the test data lies. A kernel is the core of the learning algorithm which works based on the similarity function. There are three different types of kernels in SVM namely, linear, Radial Bias Function (RBF) and polynomial kernels. The linear kernel works with linear functions and its hyper-plane is a straight line. RBF and polynomial kernels constructs the hyper-plane as Gaussian and polynomial curves.

5 EXPERIMENTAL RESULTS AND ANALYSIS

The proposed hybrid method combines the Rules Based Classification (RBC), Linguistic features (LING) and the machine learning algorithm (ML) classifiers in multiple ways. Three machine learning algorithms namely, Naïve Bayes, decision tree and Support Vector Machine (SVM) are used to analyze the performance of the model. Out of all the three, SVM performs better with the hybrid combination. The proposed hybrid classification applies the classifiers in sequence against the collected disaster data. Whenever a classifier is not performing the classification correctly, the text is passed onto another classifier until it is classified correctly. The set of possible hybrid classification configurations for text processing is shown in Fig. 8.

In the hybrid configuration, features extracted from the rules and machine learning algorithm are used in sequence as the underlying mechanics of both of them are different. The results of rule based method is fed as an input to machine learning algorithm for classification of texts. In the rule

Accepted manuscript

ACCEPTED MANUSCRIPT

based methodology, a set of documents is regarded as a mining field from which the patterns (antecedents) are extracted. These patterns are framed with a set of predefined templates by analyzing a huge number of documents in the same category. The rule based classifier makes use of these pattern sets to identify the associated categories from the entire corpus. In contrast, SVM classifier and other machine learning algorithms consider the document collection as a set of features. They assign weights to the features and the entire document is represented as a feature set. When a training set is given as input to SVM, the machine learning algorithm focuses on optimizing the associated weights and other parameters like, C and Gamma to build an optimized model. The classifier makes use of this model to classify the set of unseen data. The combination of rule based methodology and SVM classifier yields better results as the texts are processed and analyzed by different approaches simultaneously.

Fig. 8. Hybrid configuration for text processing and classification.

The proposed hybrid method is implemented and tested using *Scikit-learn* in Python programming language. The data used for training the classifier are collected from Twitter and questionnaire. Text preprocessing is performed initially as the Twitter data collected during disaster lacks the apt usage of words. During preprocessing, foreign words, stop words, URL, # symbols are removed. Parts of speech tagging (POS) is applied to the data using the Pen tree bank tagger. Named entity recognition (NER) is performed using the Stanford NER tagger on the data to identify the names of person and location. Once the rules and the linguistic features are applied to the data, the data are converted into feature vectors. These feature vectors are supplied as input to the machine learning algorithms.

5.1 Test data set description:

The prototype of the proposed hybrid model is tested with the data set that was created by Olteanu *et al.* [46]. The data set consists of tweets collected for 26 crisis events. CrisisLexT6 and ChileEarthquakeT1 data sets are considered as test data set. The CrisisLexT6 collection has tweets about 2012 Hurricane sandy, 2013 Alberta floods and 2013 Queensland floods. In the ChileEarthquakeT1 data set, 2015 Chile earthquake tweets are considered. These tweets are labeled by the relatedness to the events as On_Topic and Off_Topic. The considered Hurricane sandy dataset consists of 10008 tweets out of which 6138 tweets are On_Topic, Alberta floods dataset consists of 10031 tweets out of which 5189 are On-Topic, Queensland flood dataset consists of 10033 tweets out of which 5414 are On_Topic and the Chile earthquake dataset consists of 2187 On_Topic tweets. The

Accepted manuscript

ACCEPTED MANUSCRIPT

keyword filtering technique along with rules are applied on these On_Topic tweets which results in the test dataset of 654 tweets.

5.2 Comparative Analysis of the Machine learning algorithms:

The test dataset is given as input to all the three machine learning algorithms. The multiple combinations of machine learning algorithms with the rules are evaluated. Initially, each machine learning algorithm is applied independently to the disaster data to estimate the performance metrics. Table 3 shows the results of various combinations of rules and machine learning algorithms that are applied for the text classification. The results reveal that the combination of rules with SVM performs better.

TABLE 3: SUMMARY OF THE RESULTS OF DIFFERENT COMBINATIONS OF RULES AND MACHINE LEARNING ALGORITHMS

Combinations / Parameters	Naïve Bayes			Decision Tree			SVM		
	P	R	F_1	P	R	F_1	P	R	F_1
ML	0.81	0.67	0.68	0.78	0.76	0.77	0.79	0.77	0.78
POS ⇨ ML	0.86	0.84	0.84	0.87	0.86	0.86	0.87	0.86	0.87
General ⇨ ML	0.89	0.85	0.85	0.83	0.82	0.81	0.83	0.82	0.81
Combined ⇨ ML	0.88	0.89	0.90	0.91	0.89	0.91	0.93	0.93	**0.93**

The linguistic features such as bag of words and the n-gram are applied on disaster data to estimate the performance of the text classification algorithm. Further, the text is converted into the feature vector through the Count Vectorizer method of the *Scikit learn*. The obtained feature vector is given as input to the machine learning algorithms in different combinations. Table 4 summarizes the results of different combinations of linguistics and machine learning algorithms that are applied for positive and negative text classification.

TABLE 4: SUMMARY OF THE RESULTS OF DIFFERENT COMBINATIONS OF LINGUISTICS AND MACHINE LEARNING ALGORITHMS

Combinations / Parameters	Naïve Bayes			Decision Tree			SVM		
	P	R	F_1	P	R	F_1	P	R	F_1
Bag of words ⇨ ML	0.81	0.67	0.68	0.79	0.77	0.78	0.79	0.77	0.78
Bigram ⇨ ML	0.73	0.42	0.39	0.71	0.60	0.63	0.84	0.83	0.84
n-Gram (n=3) ⇨ ML	0.50	0.22	0.75	0.14	0.16	0.15	0.91	0.91	0.90

Accepted manuscript

ACCEPTED MANUSCRIPT

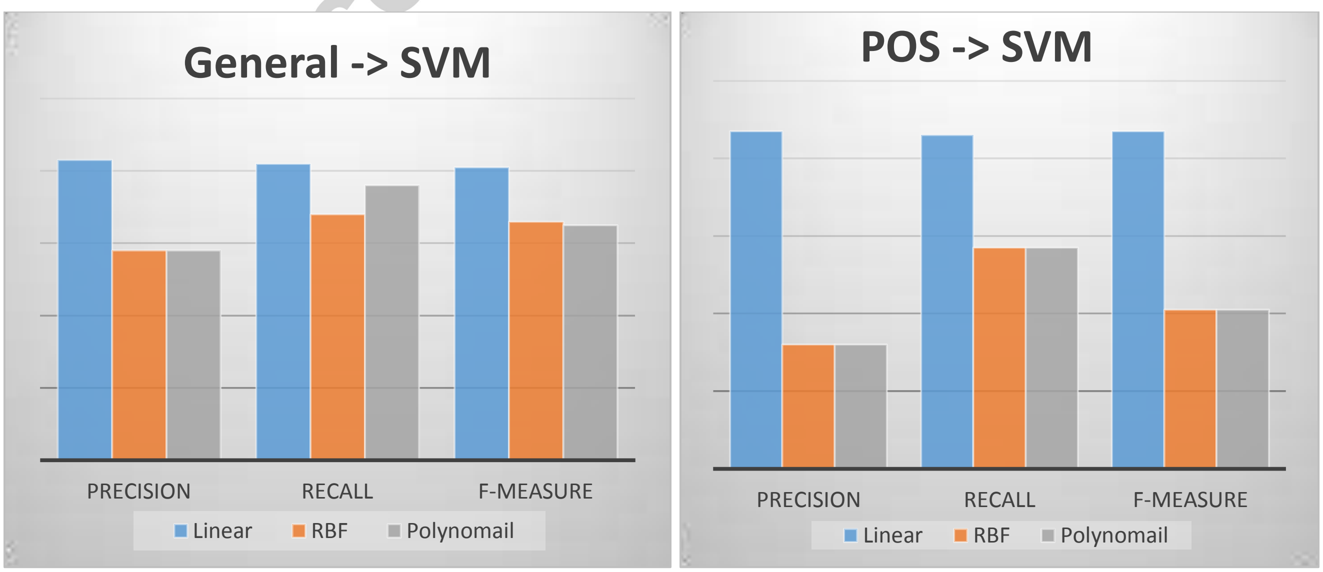
The hybrid method combines all the rules and linguistic feature vectors with the machine learning algorithm. Initially, the rules are applied and the resulting data are processed with linguistic features. The filtered texts are given as input to machine learning algorithm. Table 5 summarizes the results of the hybrid method with different combinations of rules and linguistic features with machine learning algorithms. The results reveal that the hybrid method achieves better accuracy.

TABLE 5: SUMMARY OF THE RESULTS OF DIFFERENT COMBINATIONS OF RULES AND LINGUISTICS WITH MACHINE LEARNING ALGORITHMS

Hybrid method	Naïve Bayes			Decision Tree			SVM		
	Bag-of-words	Bi-gram	n-gram (n=3)	Bag-of-words	Bi-gram	n-gram (n=3)	Bag-of-words	Bi-gram	n-gram (n=3)
POS ⇨ LING ⇨ ML	0.84	0.86	0.81	0.86	0.85	0.86	0.87	0.81	0.87
General ⇨ LING ⇨ ML	0.85	0.76	0.61	0.77	0.82	0.84	0.81	0.84	0.84
Combined ⇨ LING ⇨ ML	0.92	0.72	0.52	0.90	0.83	0.85	**0.93**	0.89	0.87

The results reveal that SVM performs better than Decision tree and Naïve Bayes algorithms. Further, all the three kernels of the SVM, namely linear, Radial Bias Function (RBF) and polynomial kernels are evaluated. The comparative results of linear, RBF and polynomial kernels in all combinations reveal that the linear SVM kernel performs better with the rules as shown in Fig. 9.

Accepted manuscript



ACCEPTED MANUSCRIPT

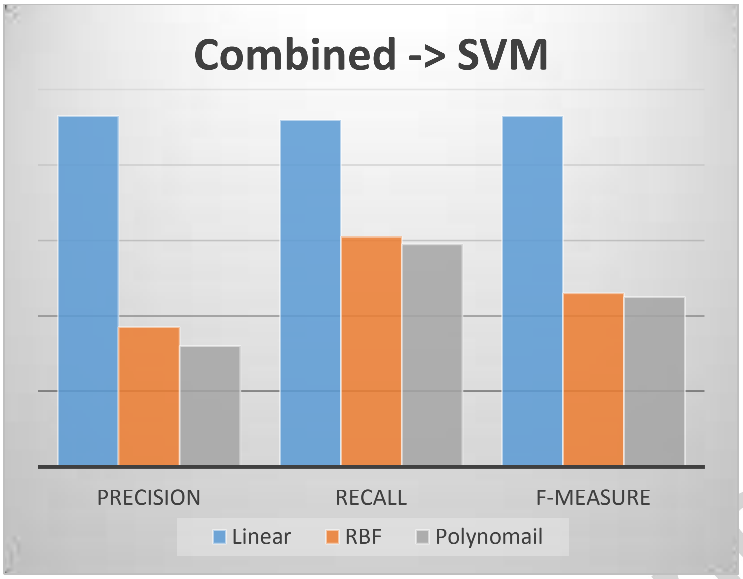


Fig. 9. Comparative performance of linear, RBF and polynomial SVM for data set.

Receiver Operating Characteristics (ROC) curve is an evaluation technique which is used to assess the ability of the algorithm in the classification of the data. Accuracy of the text classification is measured through area under the curve. When the area under the curve is equal to 1, it means that the classification is accurate and perfect. The proposed hybrid model for text classification is tested under different combinations of rules and linguistic features. The performance of the various combinations of hybrid methods are plotted using ROC curve as shown in Fig. 10. In the plotted ROC graphs, it is evident that the hybrid method with rules has higher accuracy compared to independent SVM and SVM with linguistic features. The ROC curve for combined rules and linguistics with SVM lies closer to the left side (nearly 1) than the other curves. This indicates that the performance of the hybrid combination of rules, linguistics and SVM exhibits better classification accuracy than other possible combinations.

Accepted manuscript

ACCEPTED MANUSCRIPT

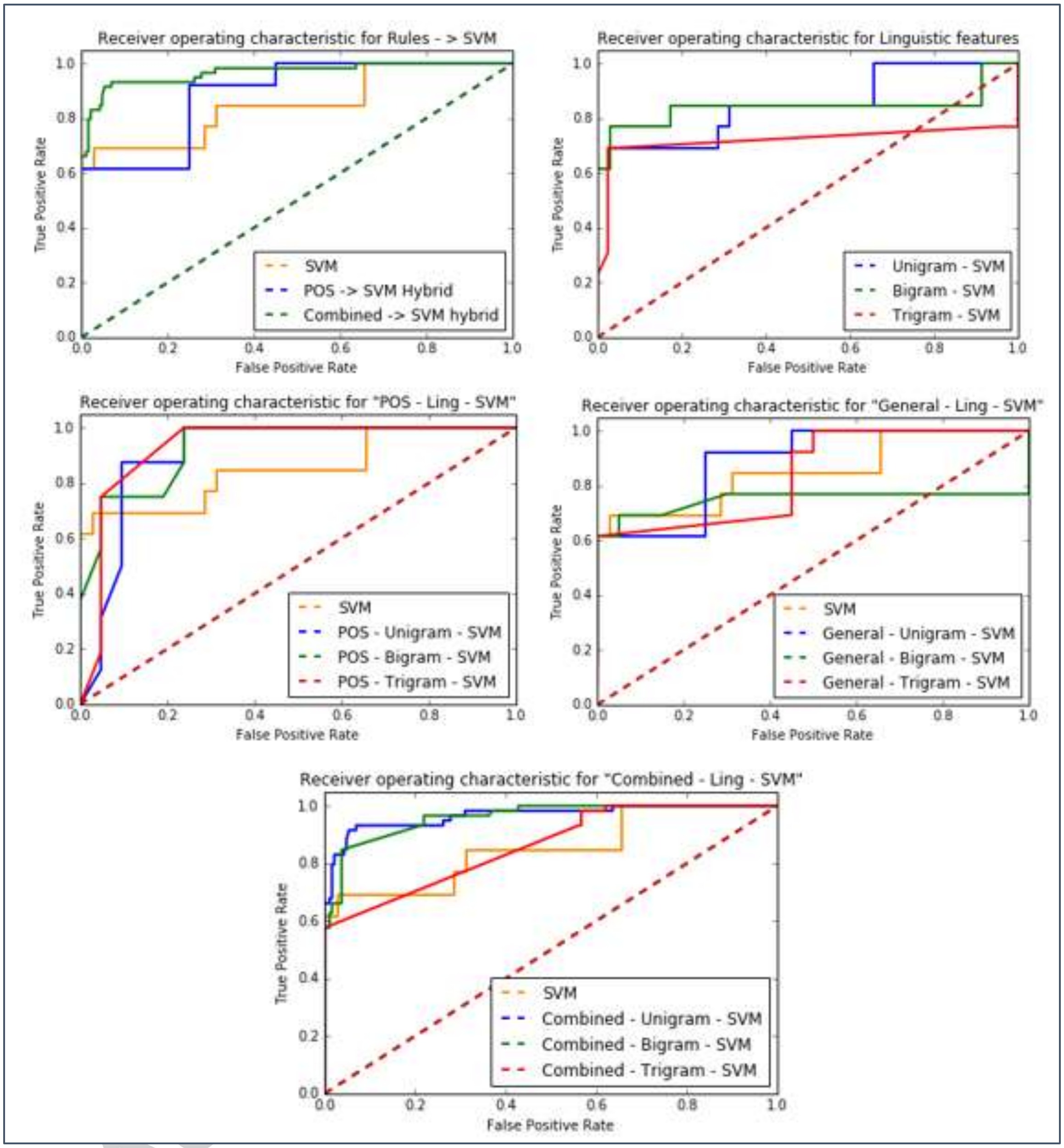


Fig. 10. Receiver Operating Characteristics (ROC) curve for different combinations of the proposed hybrid method.

6 DISCUSSIONS

A systematic approach for identifying the people who are trapped and struggling for survival during a disaster is implemented through hybrid method. This method is trained and tested based on the disaster data that is collected from Twitter. The text classification system utilizes the hybrid method that combines rule based methodology and machine learning algorithm to segregate the text from the affected population.

ACCEPTED MANUSCRIPT

The existing literature reveals that the data gleaned through social media are of significant importance to the emergency responders [46]. The emergency responders need a defined operational procedure in the event of any crisis, which gives them better situational awareness. During any crisis, the affected people are often the first responders who help the people nearby and also protect themselves. They also assist the rescue personnel directly and indirectly. Social media has rich source of information to organize disaster response and recovery activities [50]. During and after any disaster, more people use social media to share their experience, describe their feelings, communicate with other people and express about the lack of basic needs [15]. The invention of various technologies like, big data and parallel programing have made these huge volume of data manageable. Though these data have become manageable, the various challenges in using them during the disaster response and recovery include, the inability to categorize the data into useful types and inability to trust the data [32]. Due to the nature of social media, all the contributors of these data need not be from the affected people. In many occasions, other people who are concerned about the people who are affected involve in an extensive discussion about the disaster situation in social media. Such discussions create unwanted text messages that has to be filtered and eliminated which otherwise leads to confusions in the relief operations.

In segregating the texts from the affected people, the proposed method classifies such data as positive and the rest as negative. In the proposed hybrid combinations, the application of the combined rules with the bag of words and SVM gives better classification accuracy. This is due to the reason that the combined rules consider all the rules that are framed to capture maximum number of positive text. A confusion matrix is a table representation used to describe the performance of the classification model. The confusion matrix for the hybrid combination of combined rules with bag of words and SVM algorithm exhibits good classification accuracy of disaster data as shown in Fig. 11. However, the system produces 15% false negative rate due to the usage of native language words and slangs. The misclassification rate can be reduced by increasing the training data size for the machine learning classifier.

Accepted manuscript

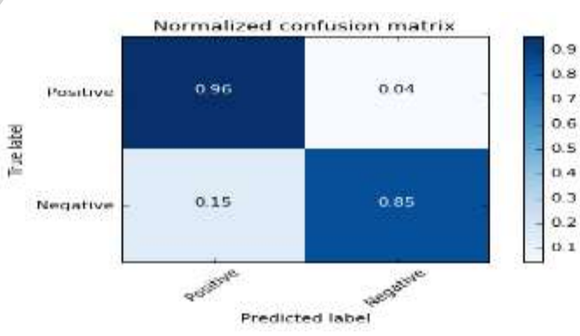


Fig. 11. Confusion Matrix for hybrid combination of combined rules and SVM.

The strongest contribution of the proposed hybrid method is the segregation of the text from the affected people who are trapped and struggling for survival during disaster. While there are many studies that focus on analyzing the sentiment of the people during a disaster [32], [34], [35], [40], our

ACCEPTED MANUSCRIPT

research is different from them since it applies sentiment analysis technique to segregate the text messages only from the affected people who are trapped. In particular, rules are proposed to classify the tweets into positive and negative helps the rescue personnel to locate and save the trapped people. The main challenge in this research is the lack of sufficient data in the positive category. The discrepancy in the number of positive and negative category texts impacts the classification accuracy of the machine learning algorithm. Though the positive category texts are proportionately less, segregating them is important since they help to save lives of the affected people and also soothes the people who are struggling without food and water for a long time. Although there are challenges in the data processing, the proposed research can be considered as a pilot study. The future work can focus on creating a dataset by integrating the data from social media for various disasters which will have a considerable amount of positive data. Furthermore, the rules proposed can be extended depending on the collected positive data which may improve the classification accuracy. Also, an extensive study on each category of the basic needs of the people will help in effective classification which in turn ease the work of the emergency responders. Despite of the challenges involved in the text classification, our proposed hybrid model serves as a basis for segregating the real-time tweets from the affected people who are at risk. This indeed helps the rescue personnel to locate and identify the trapped people which reduces the count of casualties at the place of disaster and simultaneously builds confidence among the affected people.

7 CONCLUSIONS

In this paper, a hybrid method for locating people at risk during and after disaster through real-time text classification and analysis is proposed. The hybrid method combines rule based methodology, linguistic features and machine learning algorithm to classify the texts. Various combinations of the hybrid method are analyzed in order to identify the best combination that produces the highest accuracy in text classification. The experimental results reveal that the combination of the rules, linguistic features and linear SVM improves the performance of the text classification system. The proposed real-time hybrid model receives the feed from social media and segregates the text from the disaster affected people. The results of the proposed model can be used to reduce the casualties by monitoring the social media during any disaster. The output from the proposed model is helpful for the emergency responders in rescuing the people at risk.

Accepted manuscript

REFERENCES

[1] B. Stollberg, T. De Groeve, The use of social media within the Global Disaster Alert and Coordination System (GDACS), WWW'12 - Proc. 21st Annu. Conf. World Wide Web Companion. (2012) 703–706. doi:10.1145/2187980.2188185.

[2] K.M. Carley, M. Malik, P.M. Landwehr, J. Pfeffer, M. Kowalchuck, Crowd sourcing disaster management: The complex nature of Twitter usage in Padang Indonesia, Saf. Sci. 90 (2016) 48–61. DOI: 10.1016/j.ssci.2016.04.002.

[3] Y. Xiao, Q. Huang, K. Wu, Understanding social media data for disaster management, Nat. Hazards. 79 (2015) 1663–1679. doi:10.1007/s11069-015-1918-0.

ACCEPTED MANUSCRIPT

[4] B.G. Smith, Socially distributing public relations: Twitter, Haiti, and interactivity in social media, Public Relat. Rev. 36 (2010) 329–335. DOI: 10.1016/j.pubrev.2010.08.005.
[5] S. Doan, B.-K.H. Vo, N. Collier, An analysis of Twitter messages in the 2011 Tohoku Earthquake, Electron. Healthc. (2012) 58–66. doi:10.1007/978-3-642-29262-0_8.
[6] Hurricane Sandy Gas Map. http://www.mappler.net/home/project-highlights/ hurricane-sandy-gas-map/ (accessed 25.03.17).
[7] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors, in: Proc. 19th Int. Conf. World Wide Web, 2010: pp. 851–860. doi:10.1145/1772690.1772777.
[8] S.B. Liu, Crisis Crowdsourcing Framework: Designing Strategic Configurations of Crowdsourcing for the Emergency Management Domain, Comput. Support. Coop. Work CSCW An Int. J. 23 (2014) 389–443. doi:10.1007/s10606-014-9204-3.
[9] C. Caragea, N. Mcneese, A. Jaiswal, G. Traylor, H. Kim, P. Mitra, D. Wu, A.H. Tapia, L. Giles, B.J. Jansen, J. Yen, Classifying Text Messages for the Haiti Earthquake, Proc. 8th Int. ISCRAM Conf. (2011) 1–10. doi:10.1.1.370.6804.
[10] A.M. MacEachren, A. Jaiswal, A.C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, J. Blanford, SensePlace2: GeoTwitter analytics support for situational awareness, in: VAST 2011 - IEEE Conf. Vis. Anal. Sci. Technol. 2011, Proc., 2011: pp. 181–190. doi:10.1109/VAST.2011.6102456.
[11] V.K. Neppalli, C. Caragea, A. Squicciarini, A. Tapia, S. Stehle, Sentiment analysis during Hurricane Sandy in emergency response, Int. J. Disaster Risk Reduct. 21 (2017) 213–222. DOI: 10.1016/j.ijdrr.2016.12.011
[12] A.T. Chatfield, H.J. Scholl, U. Brajawidagda, #Sandy Tweets: Citizens' Co-production of time-critical information during an unfolding catastrophe, in: Proc. Annu. Hawaii Int. Conf. Syst. Sci., 2014: pp. 1947–1957. doi:10.1109/HICSS.2014.247.
[13] K. Starbird, Digital Volunteerism During Disaster: Crowdsourcing Information Processing, Search. (2011) 1–4. http://crowdresearch.org/chi2011-workshop/papers/starbird.pdf.
[14] Munro, Robert. Crowdsourcing and the crisis-affected community. *Information retrieval* 16.2 (2013): 210-266.
[15] A. Sen, K. Rudra, S. Ghosh, Extracting situational awareness from microblogs during disaster events, 2015 7th Int. Conf. Commun. Syst. Networks. (2015) 1–6. doi:10.1109/COMSNETS.2015.7098720.
[16] G. Schimak, D. Havlik, J. Pielorz, Crowdsourcing in crisis and disaster management – challenges and considerations, IFIP Adv. Inf. Commun. Technol. 448 (2015) 56–70. DOI: 10.1007/978-3-319-15994-2.
[17] M. Zook, M. Graham, T. Shelton, S. Gorman, Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake, World Med. Heal. Policy. 2 (2010) 6–32. doi:10.2202/1948-4682.1069.
[18] C. Rizza, A.G. Pereira, Building a resilient community through social network: Ethical considerations about the 2011 Genoa floods, ISCRAM 2014 Conf. Proc. - 11th Int. Conf. Inf. Syst. Cris. Response Manag. (2014) 289–293.
[19] J.Y. Jung, M. Moro, Multi-level functionality of social media in the aftermath of the Great East Japan Earthquake, Disasters. 38 (2014). doi:10.1111/disa.12071.
[20] S. Vieweg, A.L. Hughes, K. Starbird, L. Palen, Microblogging during two natural hazards events: what Twitter may contribute to situational awareness, CHI 2010 Cris. Informatics. (2010) 1079–1088. doi:10.1145/1753326.1753486.
[21] C.W. and M.S. A Basu, Support vector machines for text categorization, in: Proc. 36th Hawaii Int. Conf. Syst. Sci., 2003: pp. 1–7.
[22] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 1998: pp. 137–142. doi:10.1007/s13928716.
[23] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up?: sentiment classification using machine learning techniques, Proc. Conf. Empir. Methods Nat. Lang. Process. (2002) 79–86. doi:10.3115/1118693.1118704.

Accepted manuscript

ACCEPTED MANUSCRIPT

[24] A. Kamal and M. Abulaish, "Statistical Features Identification for Sentiment Analysis Using Machine Learning Techniques," *2013 International Symposium on Computational and Business Intelligence*, New Delhi, 2013, pp. 178-181. doi: 10.1109/ISCBI.2013.43

[25] H. Wang, L. Liu, W. Song, J. Lu, Feature-based sentiment analysis approach for product reviews, J. Softw. 9 (2014) 274–279. doi:10.4304/jsw.9.2.274-279.

[26] D. Yan, G. Zhou, X. Zhao, Y. Tian, F. Yang, Predicting stock using microblog moods, China Commun. 13 (2016) 244–257. doi:10.1109/CC.2016.7563727.

[27] A.M. Alkalbani, A.M. Ghamry, F.K. Hussain, O.K. Hussain, Sentiment Analysis and Classification for Software as a Service Reviews, in: 2016 IEEE 30th Int. Conf. Adv. Inf. Netw. Appl., 2016. doi:10.1109/AINA.2016.148

[28] A. Sharma, S. Dey, Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis, Int. J. Comput. Appl. (2012) 15–20.

[29] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Syst. Appl. 36 (2009) 6527–6535. doi:10.1016/j.eswa.2008.07.035.

[30] Fu-Chen Yang, Anthony J.T. Lee, Sz-Chen Kuo, Mining Health Social Media with Sentiment Analysis, J. Med. Syst. (2016). doi:10.1007/s10916-016-0604-4.

[31] Qadir, J., Ali, A., ur Rasool, R., Zwitter, A., Sathiaseelan, A., Crowcroft, Crisis analytics: big data-driven crisis response, J. Int. Hum. Act. 1.1 (2016): 12.

[32] C. Caragea, A. Squicciarini, S. Stehle, K. Neppalli, A. Tapia, Mapping moods: Geo-mapped sentiment analysis during hurricane sandy, ISCRAM 2014 Conf. Proc. - 11th Int. Conf. Inf. Syst. Cris. Response Manag. (2014) 642–651.

[33] S. Verma, S. Vieweg, W.J. Corvey, L. Palen, J.H. Martin, M. Palmer, A. Schram, K.M. Anderson, Natural Language Processing to the Rescue? Extracting Situational Awareness Tweets During Mass Emergency, Proc. Fifth Int. AAAI Conf. Weblogs Soc. Media. (2011) 385–392.

[34] B. Mandel, A. Culotta, J. Boulahanis, D. Stark, B. Lewis, J. Rodrigue, A demographic analysis of online sentiment during Hurricane Irene, Proc. 2012 Work. Lang. Soc. Media. (2012) 27–36.

[35] A. Nagy, C. Valley, J. Stamberger, Crowd Sentiment Detection during Disasters and Crises, Proc. 9th Int. ISCRAM Conf. (2012) 1–9.

[36] T. Sakaki, M. Okazaki, Y. Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, IEEE Trans. Knowl. Data Eng. 25 (2013) 919–931. doi:10.1109/TKDE.2012.29.

[37] P. S. Earle, D. C. Bowden, M. Guy, Twitter earthquake detection: Earthquake monitoring in a social world, Ann. Geophys. 54 (2011) 708–715. doi:10.4401/ag-5364.

[38] BKH Vo and NIGEL Collier. Twitter emotion analysis in earthquake situations. International Journal of Computational Linguistics and Applications, 4(1):159–173, 2013.

[39] T. Simon, A. Goldberg, L. Aharonson-Daniel, D. Leykin, B. Adini, Twitter in the cross fire - The use of social media in the Westgate mall terror attack in Kenya, PLoS One. (2014). doi:10.1371/journal.pone.0104136.

[40] M.K. Torkildson, K. Starbird, C.R. Aragon, Analysis and visualization of sentiment and emotion on crisis tweets, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). (2014). doi: 10.1007/978-3-319-10831-5_9.

[41] Lewis, William D., Robert Munro, and Stephan Vogel. Crisis mt: Developing a cookbook for mt in crisis situations. Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2011.

[42] De Albuquerque, J.P., Herfort, B., Brenning, A., Zipf, A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. International Journal of Geographical Information Science, 29.4 (2015): 667-689.

[43] M. R. Sumalatha, P. Batsa, A. Sinha and P. Shrinath, “Social media for disaster relief — Geo distributed social service system,” *2015 Seventh International Conference on Advanced Computing (ICoAC)*, Chennai, 2015, pp.1-6. doi: 10.1109/ICoAC.2015.7562783

Accepted manuscript

ACCEPTED MANUSCRIPT

[44] T. H. Nazer, F. Morstatter, H. Dani and H. Liu, Finding requests in social media for disaster relief, 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, 2016, pp. 1410-1413. doi: 10.1109/ASONAM.2016.7752432

[45] Acar, A. and Muraki, Y., 2011. Twitter for crisis communication: lessons learned from Japan's tsunami disaster. International Journal of Web Based Communities, 7 (3), 392–402. doi:10.1504/ IJWBC.2011.041206

[46] Olteanu, Alexandra, Sarah Vieweg, and Carlos Castillo. What to expect when the unexpected happens: Social media communications across crises. Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, 2015.

[47] Twitter Historical data recovery tool. http://www.followthehashtag.com/features/Twitter-historical-data-recover/ (accessed on 18/03/17)

[48] Kongthon, A., Haruechaiyasak, C., Pailai, J., & Kongyoung, S. (2012). The role of Twitter during a natural disaster: Case study of 2011 Thai Flood. In Technology Management for Emerging Technologies (PICMET), 2012 Proceedings of PICMET'12: (pp. 2227-2232). IEEE

[49] M. Gamon, Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis, Proc. 20th Int. Conf. Comput. Linguist. (2004) 841–847. doi:10.3115/1220355.1220476.

[50] Imran, M., Castillo, C., Diaz, F., Vieweg, Processing social media messages in mass emergency: A survey. ACM Computing Surveys (CSUR) 47.4 (2015): 67.

Accepted manuscript