Mining social media text for disaster resource management using a feature selection based on forest optimization

Ashutosh Bhoi [a], Rakesh Chandra Balabantaray [a,*], Deepak Sahoo [b], Gaurav Dhiman [c,d,e], Manish Khare [f], Fabio Narducci [g], Amandeep Kaur [d]

[a] Department of Computer Science and Engineering, IIIT Bhubaneswar, Bhubaneswar, India
[b] Department of Faculty of Emerging Technologies, Sri Sri University, Cuttack, India
[c] Department of Computer Science, Government Bikram College of Commerce, Patiala, India
[d] University Centre for Research and Development, Department of Computer Science and Engineering, Chandigarh University, Gharuan, Mohali, India
[e] Department of Computer Science and Engineering, Graphic Era Deemed to be University, Dehradun, India
[f] DA-IICT, Gandhinagar, India
[g] Universita degli Studi di Salerno, Fisciano (SA), Italy

ARTICLE INFO

Keywords:
Disaster management
Social media
Forest optimization
Feature selection
Classification

ABSTRACT

Resource management is an essential task that needs to be performed by the government or any disaster management agency during natural disasters. During these critical circumstances, people mostly depend upon a social media platform to share and collect information about the situation of the affected localities. The huge volume of real-time data can be useful in disaster assessment, response, and relief activities. We have presented a system which analyzes tweets during natural disasters and categorizes them according to the availability or need for general or medical resources along with their location information (if any) mentioned in the tweets. Several statistical classifiers are applied to show their usefulness for a better solution. Optimal feature representation is the heart of any machine learning based classification model. Here, we have applied a forest optimization-based wrapper feature selection algorithm to improve the classification accuracy. FIRE, SMERP, and CrisisLex dataset are used to evaluate our system and its effectiveness is demonstrated for smooth management of the resources. From the experimentation, it is found that forest optimization algorithm (FOA) wrapped multinomial naive bayes classifier gives an accuracy of 91.41 percent and f-measure of 88.33 percent on the FIRE dataset. The execution time of the model is quite less which will be very helpful for this challenging task.

1. Introduction

Social Networks are becoming popular day by day since the last decade. Social media applications such as Twitter, Facebook, WhatsApp, and many more are the major source of real-time data available online. In recent times, due to intensive use of smartphone, people are very much active in social media during natural calamities. A natural disaster is the effect of a natural hazard (e.g., flood, tornado, hurricane, volcanic eruption, earthquake, heat wave or landslide) that leads to financial, environmental or human losses. The resulting loss depends upon the resilience of the affected population, and the disaster management system (Velev & Zlateva, 2012). The natural hazard may lead to disaster if there is any vulnerability in the management of resources and responsibilities for handling all humanitarian aspects in disaster response and recovery system. These losses can be reduced to a minimum level with the help of proper disaster management system (Goolsby, 2010).

Communication is one of the rudimentary tools for disaster management. During disasters, there is an increased communication since people seek to contact family, friends, and relatives in the disaster affected zone. They may also ask for information regarding food, shelter, electricity, medical facilities, transportation, and many more. Traditional methods were used for data collection through phone calls, direct inspection or interviewing the people for gathering information (Hughes & Palen, 2009). However, in the past decade, social media has played a significant role in disseminating information about these disasters by allowing people to share information and ask for help (Castillo, 2016). This increased communication in social media during a disaster generates large volume of data. Hence, the data can be

* Corresponding author.
E-mail addresses: c115001@iiit-bh.ac.in (A. Bhoi), rakesh@iiit-bh.ac.in (R.C. Balabantaray), deepsahoo@gmail.com (D. Sahoo), gdhiman0001@gmail.com (G. Dhiman), manish_khare@daiict.ac.in (M. Khare), fnarducci@unisa.it (F. Narducci), amanphd786@gmail.com (A. Kaur).

effectively used to evaluate damage, investigate impacted population, and proper mapping of needful resources like water, food, medicine, and shelter (Middleton, Middleton, & Modafferi, 2014). These data can also be used as a time reference and resource mapping for different geographical location based on the need of the affected people.

The managerial process normally includes planning, organizing, directing, and controlling. The most important aspect of any disaster management process is the allocation of finite resources. The resources may be food, water, cloth, medicine, healthcare equipments, and general infrastructures like electricity, mobile networks. Human resources like healthcare personnel, electrical and network personnel are also the most important entities during post disaster response and recovery. Therefore, these highly essential resources must be transported to the particular location based on the requirement or urgency. For this challenging task, social media texts collected during the disaster period may be very much helpful. Our proposed system collects such tweets and categorizes them based on the requirement or availability of resources as defined in Table 5. This will be very much helpful for the disaster management agencies to manage resources during the crunch period.

The extraction of useful information from social media text is a tedious task. These unstructured texts are noisy due to presence of so many spelling and grammatical errors, acronyms, genre-specific phenomena, use of hashtags, mentions, and many more. However, with the help of efficient text analysis techniques, these noisy texts can be proved beneficial for resource management and relief operations. The disaster management agencies may prioritize different operations based on the requirements of the affected people along with their locations. The objective of this work is to extract the information from tweets collected within the duration of a disaster and categorize them into predefined classes for efficient disaster resource mobilization. The FIRE, SMERP, and CrisisLex datasets are used to evaluate our system.

In recent years, many embedding based models are suggested by researchers which are performing satisfactorily for the tweet classification task (Kersten, Bongard, & Klan, 2021; Naseem, Razzak, Musial, & Imran, 2020; Zubiaga, 2020). The main limitations of these models are their strong requirement for large scale training data and powerful computational resources for implementation. These shortcomings motivate us to think of an alternate solution which must be simple and quite efficient with reasonable amount of training data. Unlike these models, our suggested wrapper-based model is not so much dependent on high volume of data and computational resources. In all these embedding based text classification systems large volume of training sets are used (mostly in thousands), whereas in our case its very less (in hundreds only as it is difficult to find large number of tweets for each target class). There are some pre-trained embedding models but they are not robust enough to work satisfactorily on all type of datasets. In addition to this, our model takes less time (i.e. around 97 s) in terms of execution whereas all these sophisticated models are taking longer period of time to complete the task.

The objectives of our research are mentioned in Section 2. The related work is discussed in Section 3. Section 4 elaborates in detail about the proposed methodology. The detail about the case study and dataset is mentioned in Section 5. Results and discussion are presented in Section 6. Finally, we conclude our work along with the findings and future scope for improvement in Section 7.

2. Research objectives

During any disaster situation, essential resource mobilization is an important task. Here, we have presented a system which utilizes the social media text like tweets to perform this task in an optimum way. We summarize our contributions as follows:

- Our proposed system is a three-tier complex text categorization system which can be used by disaster management agencies for post disaster relief operations. The first two steps are used to filter the non-disaster and non-relevant tweets from the collected tweets set. The third step of our presented system categorizes the tweets into the class of need or available tweets which may be very helpful to locate the areas and their needful resources.
- In the present era, embedding driven systems require powerful resources and huge volume of data. Our proposed wrapper based classification system competes in performance with them due to better feature optimization.
- This forest optimization algorithm (FOA) based wrapper system used in our work is very efficient due to its simplicity and performance.

3. Related work

In the last decade, online social networking sites such as Facebook, Twitter, Instagram, Google+, and many more are providing various social media tools. During recent emergencies like Nepal earthquake and Chennai flood in 2015, Haiti earthquake in 2010, Sichuan Earthquake of China in 2008, the use of social media has increased significantly. Currently, there is no established method to monitor these data so that they can be effectively used for emergency services. The problem of disaster management using social media has been touched upon by researchers in various dimensions in last few years. A systematic literature review (SLR) is performed using the SLR Tool[1] available online (Kitchenham et al., 2009). It identifies, selects and critically appraises research in order to answer a clearly formulated question. It is quite helpful to filter qualitative articles published in top conferences, journals, and workshops as many researchers are using SLR techniques during their research works.

If we unfold the literature, Caragea, Kim, Mitra, and Yen (2010) have suggested various feature representation techniques for learning text classifiers like support vector machines (SVM) and naive bayes. In Zielinski, Middleton, Tokarchuk, and Wang (2013), the authors have presented four novel perspectives for natural disaster observations using intense Twitter crawling, analysis, geo-parsing, and tweet classification. They have used this prototype as a decision support system along with the tsunami warning system. Gattani et al. (2013) have developed an application which processes Twitter data to extract the named entities, and categorizes tweets into predefined topics. A framework is recommended to enable the modeling of catastrophe and its formulation for the technical study of catastrophic social media outcomes (Houston et al., 2015). Ngai, Tao, and Moon (2015) have discussed social media research that confers to a better interpretation of the sources and consequences of the adoption and utilization of social media. A real-time modeling system is presented to distinguish areas which may have flooded using data collected only through social media (Smith, Liang, James, & Lin, 2017). Social media extends options for engaging citizens in the crisis management by both spreading information among public and accessing it back from them (Simon, Goldberg, & Adini, 2015). Khalifa, Redondo, Vilas, and Rodríguez (2016) have proposed to mine geo-tagged data from location-based social networks in order to analyze urban mass confirming to different parameters such as size, period, configuration, inspiration, cohesion, and propinquity.

In the recent past, a crisis and emergency risk communication (CERC) model has been developed to manage crisis and its implications for emergency management organizations (Lachlan, Spence, Lin, Najarian, & Del Greco, 2016). A real-time tracking system is introduced for large social media data, named social big board, for crisis management, which exhibits crisis issues and trends in a plot (Choi & Bae, 2015). Cresci, Cimino, Dell'Orletta, and Tesconi (2015) have proposed a two-fold crisis mapping system for damage detection using SVM classifier and finding the location using the geo-parsing technique. An algorithm

[1] https://www.slr-tool.com/Identity/Account/Login

is proposed to detect the states of an event with the help of web-based resources in order to make people clearly aware of an emergency event and assist the social group or government to effectively exercise the emergency events (Xu et al., 2016). A multi-stage process is proposed to extract meaning for generating corroborated emergency reports from social media data (Andrews, Gibson, Domdouzis, & Akhgar, 2016). Kryvasheyeu et al. (2016) have showed that actual and perceived risk, together with physical catastrophe effects are immediately noticeable through the extremity and configuration of Twitter's message stream. It is also established that the per-capita Twitter activity firmly associates with the per-capita economic damage imposed by the hurricane. Emergency services have not used the social media properly either to tell their state of awareness and retaliation or to convey the citizens what duty they should perform (Spielhofer, Greenlaw, Markham, & Hahne, 2016). A two-phased sentiment analysis framework is suggested for informal Turkish text-based on aspect extraction and matching of aspects to their corresponding sentiment words present in the text (Karagoz, Kama, Ozturk, Toroslu, & Canturk, 2019). In Reynard and Shirgaokar (2019), the authors have suggested strategies for extracting information from tweets, and presents possible ways to establish a probability-based understanding of needs to mentor disaster management agencies. Dynamic acquisition and processing of social media data in the run time can further improve the performance of any disaster management system (Kabir, Gruzdev, & Madria, 2020). In Behl, Rao, Aggarwal, Chadha, and Pannu (2021), the authors have tried to explore the reusability of previously trained disaster models on recent COVID-19 pandemic data for its testing.

Here, the problem which we have attempted to address is a complex text categorization problem. In the literature, several text categorization methods are proposed for structured text like newswire text (Aphinyanaphongs et al., 2014; Lai, Xu, Liu, & Zhao, 2015; Wang, Zhang, Liu, Lv, & Wang, 2014). However, these state-of-the-art systems do not perform satisfactorily on social media text, like tweets and Facebook posts due to the presence of noise. Therefore, a robust disaster resource management system using social media text classification is the need of the hour. Like other classification tasks, feature selection may be an important step in the text classification task to improve the accuracy. In the literature, several feature selection approaches are proposed by the researchers for the text classification task (Aphinyanaphongs et al., 2014; Chandrashekar & Sahin, 2014). A handcrafted feature selection-based approach is suggested by Sriram, Fuhry, Demir, Ferhatosmanoglu, and Demirbas (2010) to improve the tweets categorization accuracy. Jiang, Liou, and Lee (2011) have recommended a fuzzy-similarity based self-constructing feature clustering approach for efficient text classification. Therefore, selection of the most relevant features can be an important step along with the classification algorithm to improve the accuracy of our disaster management system. Many optimization based feature selection algorithms are also used for this categorization task (Boussaïd, Lepagnot, & Siarry, 2013). An ant colony based feature selection technique is suggested by Tabakhi, Moradi, and Akhlaghian (2014) to select the optimal features to improve the classification accuracy. Zhang, Wang, Phillips, and Ji (2014) have used a binary particle swarm optimization based feature selection algorithm for the spam detection task. A genetic programming based new encoding scheme of hierarchical structure is suggested by Arif, Li, and Iqbal (2017) to classify high dimensional real-valued feature vectors for social media text categorization task. Nguyen, Yang, Zhu, Li, and Jin (2018) have proposed a heuristic reinforced learning for scheduling natural disaster emergency response. Researchers have also surveyed on many metaheuristic-based algorithms to find their strengths and weaknesses on different problem domains (Dokeroglu, Sevinc, Kucukyilmaz, & Cosar, 2019; Hussain, Salleh, Cheng, & Shi, 2019).

Recently, FOA is applied in many problem domains for efficient feature selection and dimensionality reduction (Kostrzewa & Brzeski, 2019; Nouri-Moghaddam, Ghazanfari, & Fathian, 2021). In Naz, Zafar, and Khan (2019), the authors have suggested an ensemble based sentiments categorization system employing FOA. Mohanty, Rup, Dash, Majhi, and Swamy (2018) have proposed a FOA based feature selection approach for effective mammogram classification. FOA is hybridized with other evolutionary algorithms for selecting the optimal features in gene expression data (Baliarsingh, Vipsita, & Dash, 2020; Nouri-Moghaddam, Ghazanfari, & Fathian, 2020). It is applied on many UCI repository datasets and found to be robust for machine learning based applications. All these recent successes of FOA motivated us to apply this tree based algorithm for this resource mobilization task during any emergency situation.

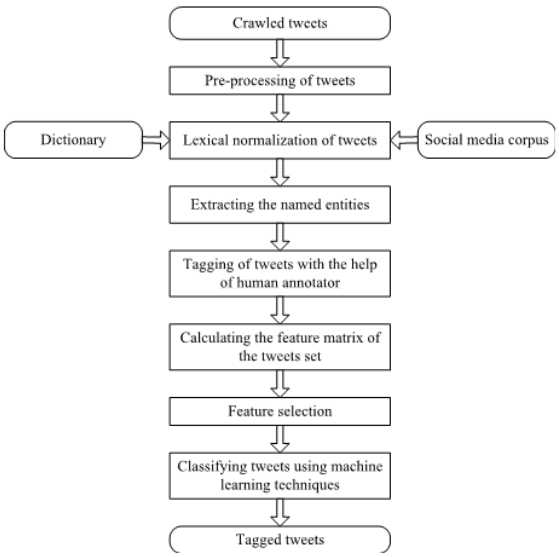

Fig. 1. Solution architecture.

The misclassification is a common problem in the analysis of Twitter data. In the sentiment analysis task, few works have been devoted to minimize the misclassification rate. Parikh and Movassate (2009) have identified various spelling issues for this task and proposed methods to handle these errors for reducing this rate. In Zhang and Desouza (2014), the authors have suggested that the misclassification rate can be lowered down by reducing the sparseness of the feature matrix and by grouping some terms or features that often appear together. Bakliwal et al. (2013) have presented different reasons for misclassification, like focus only on adjectives, absence of sentiment words etc., and has considered a combination of feature and lexicon based approach to reduce the misclassification rate. In addition to these techniques various deep learning based approaches are suggested for text categorization (Kang, Choi, & Lee, 2019). However, they have requirement for strong hardware support and enormous training data. Due to these limitations, traditional machine learning model with good feature selection algorithm is still needed for many tasks. From the recent literature, it is found that multinomial naive bayes works quite well on textual data (Hossain, Sharif, & Hoque, 2020). In several natural language processing (NLP) applications, multinomial naive bayes is shown as an accurate, fast, and reliable classifier. It is quite efficient for analysis of microblogs contents like emotion detection, sentiment polarity detection, business review, and many more (Delizo, Abisado, & De Los Trinos, 2020; Khotimah & Wasono, 2020; Sharupa et al., 2020). It may be due to the multinomial distribution of features in microblogs which encouraged us to apply it on our noisy text to find a robust solution for disaster resource mobilization task.

4. Proposed methodology

The proposed system classifies the tweets related to natural disasters into different pre-specified categories. The categories are based on the requirement or availability of different kind of resources during natural disasters. The architecture of the proposed system is depicted in Fig. 1.

4.1. Crawling and preprocessing of tweets

We have crawled the tweets from the Twitter using their APIs available. The crawled tweets are from certain duration during any natural hazards using a list of disaster related keywords. It may also contain non-disaster tweets along with the disaster tweets. So, first we need to classify the disaster related and unrelated tweets from this tweet set.

After crawling the tweets, we preprocess each of the tweets available for classification. Twitter provides a facility for its users to use their own language other than English. Hence, a language identification module is used to filter only the English tweets. Then, the user IDs, URLs, email IDs, HTML tags (if any present) are removed from the tweets. Finally, various delimiters such as comma, semicolon, colon, period, and extra spaces are removed from the tweets in the preprocessing.

4.2. Lexical normalization of tweets

Text normalization is one of the most important task for any NLP applications. Its importance is still more for any application based on informal text (Han, Cook, & Baldwin, 2013). The common noises present in these microblogs are missplellings (e.g., earthqake → earthquake), phonetic substitutions (e.g., 4 m → from), shortenings (e.g., fwd → forward), acronyms (e.g., NM → not much), slang (e.g., troll), emphasis (e.g., nooooo → no), and punctuation (dont → don't). To make our proposed system a reliable one, we need a robust normalization mechanism. Here, we have applied a neural Seq2Seq model along with the hybrid neural model that utilizes a word based encoder–decoder architecture for in vocabulary tokens and a character level sequence to sequence model to correct complex normalization errors (Lourentzou, Manghnani, & Zhai, 2019).

4.3. Extracting the named entities

In this phase, we have extracted the location and organization entities from the tweets using GATE Twitter named entity recognizer (Bontcheva et al., 2013). These two attributes (location and organization names) are used as handcrafted features during classification. The word capitalization is an important feature for named entity recognition, which is used to extract the named entities. The country name is important in the location list, however, the dataset we have considered, i.e. Nepal earthquake, is only restricted to one country. Keeping this in view, we have eliminated the country names from the location list and did not consider a country name as location in this case.

4.4. Tagging of tweets with the help of human annotators

There are many statistical machine learning algorithms available for classification. It is observed from the literature that if labeled data is available, then supervised algorithms may give better results than that of the unsupervised algorithms. The main focus of this work is resource mobilization during a disaster with prime emphasis on medical domain. Keeping this in view, we have identified eight class labels or tagset i.e. FMT0 to FMT7 in line with FIRE 2015 tagset which is described in Table 5 of Section 5. The tagging task has been carried out by ten students individually. Hence, for maintaining uniformity in the process we have computed the inter judgmental agreement (Kappa measure)[2] among various pairs of students for few tweets. The average value of Kappa measure computed over various pairs of students is 0.78. This score signifies the proportion of times the annotators (students) would be expected to agree by chance. The details regarding this experiment are mentioned in Table 1.

Table 1
Kappa measures for human annotators.

Sl.No.	Annotator pair	Kappa measure
1	Student-1 & student-2	0.83
2	Student-3 & student-4	0.8
3	Student-5 & student-6	0.75
4	Student-7 & student-8	0.70
5	Student-9 & student-10	0.84
	Average among all	0.78

4.5. Calculating the tf_idf matrix of tweets set

The tweets are stemmed with the porter stemmer after the tagging of tweets is done successfully. There are many common words in the tweets which do not contribute to the classification accuracy. These stopwords need to be removed. In the stopword list, we have not removed words such as not, have, have not, do not, had, had not, does not, and many more, as these words are important features for classification of the resource into required or available classes. After stemming, the tf_idf score of each tweet is calculated to represent the tweet set into a feature matrix. Here, we have considered each tweet as a sample and each distinct word in the tweet set as a feature in the feature matrix.

After generating the tf_idf matrix, the score values are normalized using min–max normalization (Jain, Nandakumar, & Ross, 2005) as follows:

$$Normalized\ score = \frac{actual\ score - min\ score}{max\ score - min\ score} \tag{1}$$

Finally, we have added the location and organization names as two handcrafted features in the normalized feature matrix. These two are the only handcrafted features included in the feature set as they are very important for the classification of this task.

$$tf\_idf[i][loc] = \begin{cases} 1, & \text{if } any\ location\ name\ mention\ in\ the\ ith\ tweet \\ 0, & \text{if } no\ location\ name\ mention\ in\ the\ ith\ tweet \end{cases} \tag{2}$$

$$tf\_idf[i][org] = \begin{cases} 1, & \text{if } any\ organization\ name\ mention\ in\ the\ ith\ tweet \\ 0, & \text{if } no\ organization\ name\ mention\ in\ the\ ith\ tweet \end{cases} \tag{3}$$

4.6. Feature selection

In the tf_idf matrix, there are many words which occur rarely in different tweets. This leads to the creation of a sparse tf_idf matrix. Hence, we have applied a feature selection algorithm to reduce the features (or words) in the feature matrix (or tf_idf matrix) for efficient classification and improving computation time. Here, we propose a forest optimization-based wrapper feature selection algorithm for this task.

[2] https://en.wikipedia.org/wiki/Cohen's_kappa.

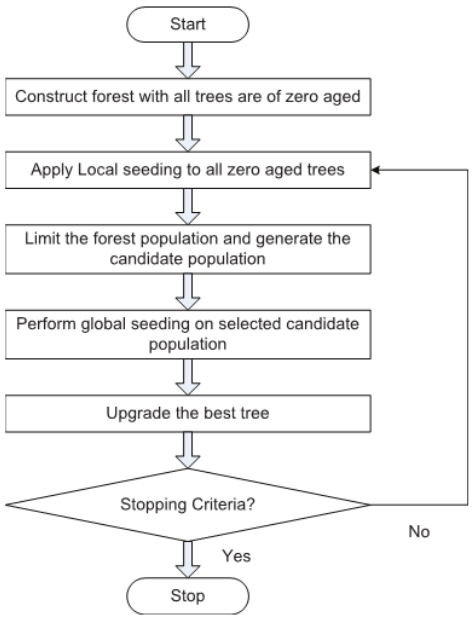

Fig. 2. Steps of forest optimization algorithm.

4.6.1. Forest optimization-based feature selection

Feature selection plays an important role in most of the classification tasks. The significance of feature selection is very high for the text categorization task. Proper feature selection techniques improve the classification accuracy as well as reduce the computation time. Many optimization techniques are used for the feature selection task. Ghaemi and Feizi-Derakhshi (2014) have proposed a tree-based new evolutionary algorithm, named forest optimization algorithm which is tested on many optimization tasks. In Ghaemi and Feizi-Derakhshi (2016), the authors have suggested FOA to remove the irrelevant and redundant features which further improves the classification accuracy. In our proposed disaster management system, a wrapper-based feature selection algorithm called as FOA is used to select the optimal features. This algorithm is quite fast in execution as its total execution time is 97 s in our task. The steps of the FOA are mentioned in Fig. 2.

4.7. Classifying the tweets using machine learning techniques

The feature matrix is generated after feature selection is given as input to different classifiers like SVM (Cortes & Vapnik, 1995), k-nearest neighbors (Cover & Hart, 1967), decision tree (Quinlan, 1986), and multinomial naive bayes (Metsis, Androutsopoulos, & Paliouras, 2006) to classify the tweets based on the pre-specified category. However, before classifying the tweets based on the mention of resource need or available, we need to classify whether the tweet is related to disaster or not. After getting the set of disaster related tweets, our system classifies them as relevant or not based on mention of any kind of resource in the tweet. Finally, the system classifies the tweets based on the pre-categorized classes. At last, the accuracy of each classifier is calculated. The scikit learn package is used here for implementation of all the models.

4.7.1. Support vector machine classifier

SVM is one of the most popular machine learning classifiers based on supervised learning. It is a non-probabilistic approach based classifier which represents each sample as points in the hyperspace and maps them such that different categories are separated by a hyper-plane. It

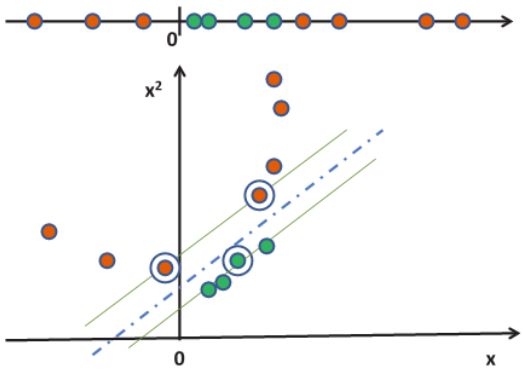

Fig. 3. Transforming nonlinear separable sample points to linear separable points in high dimensional space.

transform the nonlinear separable sample points to linearly separable points in higher dimensional space as mentioned in Fig. 3. Those points which are on the decision boundary are called as support vectors. The gap between the separating hyper-plane and the decision boundaries are of maximum possible margin.

The multi-class classification problem can be solved by training multi-class SVM with the help of a binary classifier over a feature vector $\chi(\vec{x}, y)$ derived from both the input features and the class label. During the time of testing, the classifier selects the class label as:

$$y = \text{argmax}_{y^l} \vec{w}^T \chi(\vec{x}, y^l) \tag{4}$$

The mapping from input space to high dimensional feature space is done with the help of kernel function.

$$K(\vec{x}_i, \vec{y}_j) = (\vec{x}_i^{\,T}, \vec{y}_j) \tag{5}$$

With the help of some transformation $\Phi : \vec{x} \mapsto \chi(\vec{x})$ the classifier can be formulated as:

$$f(\vec{x}) = sign(\sum_i \alpha_i y_i \chi(\vec{x}_i^{\,T}) \chi(\vec{x}_j) + b) \tag{6}$$

where

$$K(\vec{x}_i, \vec{x}_j) = \chi(\vec{x}_i^{T}) \chi(\vec{x}_j) \tag{7}$$

4.7.2. Multinomial naive bayes classifier

Multinomial naive bayes classifier is a typical case of traditional naive bayes classifier which works well with the multinomial distribution of features. In text document, the contributing features are the words or tokens. The words or phrases in a sentence or document have high dependencies or relationships to their neighboring words. They follow the multinomial distribution which encourages us to apply the multinomial naive bayes classifier for our problem. In a number of NLP applications, multinomial naive bayes has proven to be an accurate, fast, and reliable classifier. Here, we have used the $sklearn \cdot naive\_bayes \cdot MultinomialNB$ for the implementation task.

In the multinomial classification model, feature vectors (tweets) represent the tf-idf values of each word present in the tweets. The multinomial distribution works well for both integer and fractional feature counts. The probability of a tweet t being in class c is determined as follows:

$$P(c|t) \alpha P(c) \prod_{1 \le k \le n_t} P(w_k|c) \tag{8}$$

where $P(w_k|c)$ is the class conditional probability of word w_k present in tweet class c. P(c) is the prior probability of a tweet being in class c.

A sample (tweet) or feature vector $F = (f_1, f_2, \ldots, f_n)$ is a histogram with f_i is the tf-idf values of each word in the tweet. The classification

Table 2
Tweet class distribution in FIRE dataset.

Classes	Non-disaster	Disaster							
		Relevant							Non-relevant
		FMT1	FMT2	FMT3	FMT4	FMT5	FMT6	FMT7	FMT0
#Tweets	47268	524	480	475	335	274	440	402	2170

model for Multinomial Naive Bayes classifier is

$$\bar{y} = \underset{k\epsilon\{1,\ldots,n_t\}}{\operatorname{argmax}} \; p(C_k) \prod_{i=1}^{n} p(f_i|C_k) \tag{9}$$

where n_t is the number of distinct words present in the tweet.

$$p(C) = \frac{N_c}{N} \tag{10}$$

where N_c is the number of tweets present in the class c.

N is the total numbers of tweets in the sample.

The likelihood of observing a histogram of a sample F is given by

$$p(F|C_k) = \frac{(\sum_i^n f_i)!}{\prod_i^n (f_i)!} \prod_i^n p_{ki}^{f_i} \tag{11}$$

5. Description of case study and dataset

This section discusses in detail about the case study and dataset used for evaluation of our proposed system. Proper distribution of urgently needed resources as per the requirement is the most important thing that needs to be performed by disaster management agencies during an emergency situation. Since social media communication during natural disasters increases rapidly due to the wide usage of smartphones, these data can be mined effectively to be used for resource management. There are several kinds of noise associated with these social media data which must be removed first to extract the meaningful information from it. The objective of this research work is to develop a system prototype which will classify the tweets according to availability or requirement of different urgently needed resources, locations, casualties or any organization (e.g., Government, private, NGOs) mentioned in the tweet.

Dataset

We have used one real time dataset to evaluate our proposed system due to unavailability of any gold standard dataset for our task. A large set of tweets related to Nepal earthquake is collected from Forum for Information Retrieval Evaluation (FIRE). A wide variety of tweets are present in the FIRE 2016 and 2017 dataset, which are used to evaluate our system performance. As some tweets are common in both the datasets, we have filtered and removed the redundant tweets from them. After completion of the preliminary operations (from step 4.1 to step 4.4), 5100 tweets are collected for tagging using human annotator. We have further collected tweets related to sports, entertainment, politics, tourism, and finance randomly with the help of Twitter streaming APIs. We have filtered 47268 tweets based on the relevance to our evaluation process from the above collected tweets of different domain. All these tweets are considered as non-disaster tweets. Our proposed system is a multistage, multi-class categorization system. The first two stages in the classification pipeline are binary classification. In the first stage, it is disaster vs non-disaster and in the second stage, it is relevant vs irrelevant. However, the final stage is a multi-class categorization task as there are seven numbers of target classes involved in this process. The tagging task carried out carefully several times by the human annotators. The rules along with the tweet class labels mentioned in Table 5 are the guiding principle for the tagging task. We have identified eight class labels or tagset i.e. FMT0 to FMT7 in line with FIRE 2015 tagset. Along with the above eight classes we have added one extra class label as non-disaster. Non-disaster label denotes such tweets which are not related to any disaster. It will be helpful to filter such non-disaster tweets from tweet streaming contents. If any tweet falls in any class between FMT1 to FMT7 are considered as relevant tweet. However, if any tweet which is related to any disaster but does not contain any information as mentioned in classes FMT1 to FMT7 are considered as non-relevant (FMT0) tweet for our task. The details of class distribution of tweets are given below in Table 2. Along with the FIRE dataset, we have used the SMERP (Social Media for Emergency Relief and Preparedness) and CrisisLexT26 (Olteanu, Vieweg, & Castillo, 2015) datasets to evaluate the disaster vs non-disaster task of our proposed system. SMERP 2017 dataset contains tweets related to Italy earthquake occurred in 2015 and CrisisLexT26 dataset contains tweets gathered during twenty six different disasters. The details about CrisisLex and SMERP dataset are mentioned in Tables 3 and 4 respectively.

6. Results and discussion

This section describes experiments that evaluate our system prototype and shows its utility. Here, we have shown the classification accuracy of the entire system with different state-of-the art classifiers used in the literature. The feature selection algorithm reduces to a set of 489 new features from the original set of 4011 features. Here, we have segmented the total samples into different ratios of training and testing varies from 60:40 to 90:10. The results that are given in subsequent tables are the average accuracy of all four cases (i.e. 60:40, 70:30, 80:20, and 90:10) in ten different runs. To provide better generalization, two widely used cross validation, i.e. 5-fold and 10-fold are also incorporated to evaluate the system. The performance of all the classifiers is evaluated based on the most widely used measures, such as accuracy, precision, recall and F1-measure.

Precision is the fraction of true positives to the total data that are classified as true. Mathematically, it is represented as:

$$Precision(P) = \frac{T_p}{T_p + F_p} \tag{12}$$

where T_p is the total number of true positives and F_p is the total number of false positives as classified by the classifier.

Recall is the fraction of true positives to the total number of true data. Mathematically, it is represented as:

$$Recall(R) = \frac{T_p}{T_p + F_n} \tag{13}$$

where F_n is the total number of false negatives as classified by the classifier. The high value of precision signifies the less number of false positives returned by the system. Similarly, the high value of recall signifies the less number of false negatives returned by the system. Normally, when it is tried to increase the precision value the recall value decreases. So to properly evaluate a system another parameter is used named as F1-measure, which is the harmonic mean of precision and recall. Mathematically, it is represented as:

$$F1 - measure = \frac{2 * P * R}{P + R} \tag{14}$$

In most of the disaster-related tweets, there may not be any mention of resources. Therefore, we need to filter those irrelevant tweets in which there is no mention of any kind of resources (neither need nor availability). Binary classifiers are used to perform the above classification task. Once relevant tweets are filtered, then multi-class classifiers can be applied on those tweets to perform our main classification task as defined in the problem.

Table 3
Details of CrisisLexT26 dataset.

No. of crisis	Total Tweets	Labeled tweets	Class labels	Information types	Information sources
26	250K	28K	Informative Not informative	caution, advice infrastructure damage etc.	Government NGOs

Table 4
Details of SMERP dataset.

Name of crisis	Total Tweets	Labeled tweets	Class labels	Topics
Italy earthquake 2016	72K	NA	Relevant Not relevant	Activities of government, NGOs, resource required, resource availability, infrastructure damage, restoration etc.

Table 5
Rules for classification of tweets in FIRE dataset.

Sl. No.	Class label	Mention	Description	Sample tweet
1	FMT1	What resources were available	An appropriate tweet must state the possession of any kind of resource like drinking water, clothes, blankets, shelter, food, water filter, electricity, infrastructure like tents, and human resources like volunteers and so on. Tweets reporting the possession of shipping vehicles for assisting resource distribution process would also be appropriate. Nevertheless, generalized tweets without mention of any resource or tweets asking for fund would not be appropriate.	Food Distribution in sindupalchowk district, sufficient for 7 days for 500 earthquake victims under assistance... http://t.co/eZxCcJRLSl
2	FMT2	What resources were required	An appropriate tweet must state the demand/need of any kind of resource as mentioned in FMT1. Nevertheless, generalized tweets without mention of any specific resource or tweets asking for donation of money would not be appropriate.	After earthquake. Need shelter soon. http://t.co/972GpUnvvV
3	FMT3	What medical resources were available	An appropriate tweet must state the possession of any kind of medical resource like medical equipment, blood, medicines, supplementary food items (e.g., milk for infants), water filter, electricity, ambulance, human resources like doctors/medical staff and resources to build or support medical infrastructure like tents, etc. Generalized tweets without mention of any medical resources would not be appropriate.	We are collecting Medicines, Tents, Water Bottles, Biscuit Packets, for Nepal. We are sending all materials with ... http://t.co/jJTZm0j6FN
4	FMT4	What medical resources were required	An appropriate tweet must state the demand of any kind of medical resource as mentioned in FMT3. Generalized tweets without mention of any medical resources would not be appropriate.	@PMOIndia Nepal waiting for more help. need blood, Medicine, water and food packet.
5	FMT5	What were the need/possession of resources at specific locations	An appropriate tweet must state both the requirement or possession of some resource, (e.g., shelter, food, water, medical resources, tents, power supply, human resources like volunteers doctors/medical staff) as well as a particular geographical location. Tweets containing only the requirement/possession of some resource, without stating a geographical location would not be appropriate.	Nepal earthquake: No food, no clothes, no shelter, say locals from Gorkha district at the epicenter of quake. We... http://t.co/NbhBYilcDl
6	FMT6	What were the activities of various NGOs/government organizations	An appropriate tweet must carry information about relief-based activities of different Government organizations and NGOs in rescue and relief operation. Tweets that carry information about the human resources like volunteers visiting different geographical locations would also be appropriate. However, tweets that do not held the name of any Government organization/NGO would not be appropriate.	BSNL heading towards Nepal to fix their telecommunication network after the earthquake. Scientist were right. Nepal yet to face more disastes.
7	FMT7	What casualties, infrastructure damage and restoration were being reported	An appropriate tweet must state death, injuries, damage or restoration of electricity, mobile or Internet connectivity, some specific infrastructure resources such as structures (e.g., dams, houses, and mobile tower), communication infrastructure (e.g., roads, runways, and railway) etc. Generalized tweets without mention of infrastructure resources would not be appropriate.	RTrajdev_neha: #MSGHelpEarthquakeVictims More than 100 people r knwn to hv died in a \npowerful earthquake dat struck Nepal, wrecking many \n\u2016
8	FMT0	None of the above mention	An appropriate tweet must not mention any of the general or medical resource. There must not be any location information along with any kind of resource or organization or anything about damage or loss.	Nepal wl again stand up n\u2016

In our implementation the fitness function which is used is given as below:

$$fitness = avgAcc = \frac{\sum_{i=1}^{k} testAcc_i}{k} \quad (15)$$

where $avgAcc$ = Average test accuracy of the wrapper classification model
$testAcc_i$ = Test accuracy of the wrapper classification model on the ith fold
k = number of folds used in cross validation

The stopping criterion which is used in all our experimentation is 'number of iterations'. We have set the number of iteration as 300. In local seeding of the forest optimization algorithm it will select those columns from tf-idf matrix where there is mention of any resource words like food, water, shelter, medicine, cloth, doctor, human resources, infrastructure, ambulance, etc. However, for global seeding it will include those negation words like no, not, do not, have not, does not etc. if any present in tweets. Overall, it will improve the feature quality which significantly improve the performance of the classification system. In our implementation, all preprocessing operations are done with the help of python NLTK library. All machine learning algorithms are implemented using scikit-learn 0.22.0 package. The FOA algorithm is coded in python.

Table 6
Measures on disaster vs non-disaster tweet.

Dataset	Classifier		Avg. result of 4 different ratios				5-fold cross validation				10-fold cross validation			
			A	P	R	F1	A	P	R	F1	A	P	R	F1
FIRE	KNN	K = 3	81.67	83.07	80.74	81.89	81.68	82.96	80.17	81.54	81.14	82.19	79.48	80.81
		k = 5	84.43	84.19	82.37	83.27	84.09	84.07	81.96	83.00	83.92	84.03	81.43	82.71
		k = 7	86.39	85.58	83.49	84.52	86.39	85.49	83.42	84.44	83.27	84.89	82.06	83.45
	SVM	Linear kernel	87.35	85.91	81.82	83.82	87.07	85.86	81.74	83.74	86.98	85.31	81.14	83.17
		RBF kernel	91.26	90.32	85.06	87.61	91.13	90.31	85.03	87.59	91.04	89.94	84.75	87.27
	Decision Tree (C4.5)		89.33	87.74	84.47	86.07	89.19	87.79	84.68	86.20	88.94	87.67	84.07	85.83
	Multinomial Naive Bayes		91.41	90.92	85.88	88.33	91.24	90.92	86.11	88.44	91.21	90.65	85.79	88.15
SMERP 2017	KNN	K = 3	81.24	82.83	80.52	81.66	81.21	82.74	79.92	81.31	80.82	81.94	79.26	80.58
		k = 5	84.13	84.03	82.18	83.09	83.85	83.85	81.74	82.78	83.69	83.78	81.18	82.46
		k = 7	86.09	85.34	83.21	84.26	86.17	85.23	83.18	84.19	83.03	84.71	81.83	83.25
	SVM	Linear kernel	86.97	85.73	81.57	83.60	86.87	85.59	81.49	83.49	86.71	85.04	80.98	82.96
		RBF kernel	90.89	90.04	84.84	87.36	90.93	90.08	84.81	87.37	90.84	89.74	84.52	87.05
	Decision Tree C4.5		89.02	87.52	84.19	85.82	88.83	87.58	84.43	85.98	88.63	87.49	83.90	85.66
	Multinomial Naive Bayes		91.15	90.69	85.67	88.11	91.02	90.92	86.11	88.45	90.96	90.42	85.79	88.04
CrisisLex	KNN	K = 3	81.31	82.87	80.58	81.71	81.25	82.78	79.95	81.34	80.87	81.92	79.23	80.55
		k = 5	84.27	84.15	82.24	83.18	83.97	83.93	81.78	82.84	83.85	83.82	81.21	82.49
		k = 7	86.14	85.41	83.27	84.33	86.26	85.29	83.21	84.24	83.10	84.74	81.19	82.93
	SVM	Linear kernel	86.99	85.79	81.61	83.65	86.92	85.65	81.53	83.54	86.78	85.08	81.20	83.09
		RBF kernel	91.02	90.13	84.89	87.43	90.98	90.16	84.87	87.44	90.93	89.79	81.21	85.28
	Decision Tree C4.5		89.15	87.60	84.24	85.89	88.90	87.64	84.46	86.02	88.77	87.55	81.23	84.27
	Multinomial Naive Bayes		91.29	90.77	85.71	88.17	91.18	90.99	86.16	88.51	91.06	90.49	81.24	85.62

The classification results of disaster or non-disaster tweet are shown in Table 6. We have evaluated this task of our proposed system on FIRE, SMERP, and CrisisLexT26 datasets. For preparation of the non-disaster tweet dataset we collected tweets related to sports, entertainment, politics, tourism, and finance randomly for evaluation of our system. From the above collected non-disaster tweets, we filtered 47268 tweets based on relevance to our evaluation process. We have considered all Nepal earthquake tweets of FIRE dataset and Italy earthquake tweets of SMERP dataset as disaster tweets. On both datasets the proposed system with multinomial naive bayes classifier gives the best results among all. The results are obtained with a reduced set of features, after employing the forest optimization based feature selection algorithm. In this binary classification, the difference between FOA based SVM system and multinomial naive bayes system is quite less. Among all the datasets, our proposed system has shown better results with FIRE dataset, which may be due to maximum number of tweets.

The classification measures of the relevant vs. irrelevant tweets are given in Table 7. The tweets which belong to any of the classes from FMT1 to FMT7 are considered as relevant and the rest are irrelevant. It is useful to filter those relevant tweets which are required to be tagged by the help of human annotators. From all the experiments, it is found that the best results are achieved using FOA based multinomial naive bayes classification system. Similar to the results of Table 6, here also it is observed that the difference between FOA based SVM system and multinomial naive bayes system is not that much higher which may be due to the two class classification problem. Fig. 4 presents the bar chart for the performance measures (5-fold cross-validation) obtained by different classifiers for relevant vs. irrelevant tweet categorization.

The classification results of the overall system as mentioned in Table 8 are obtained with a reduced set of features after employing the feature selection algorithm. The accuracy of k-nearest neighbors (KNN) and the decision tree are nearly same with the value of $k = 5$.

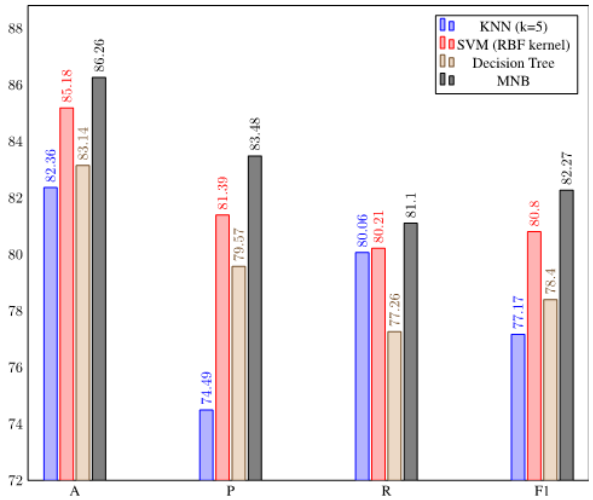

Fig. 4. 5-fold cross validation measures on relevant vs irrelevant tweet.

Our system prototype gives almost the same results with a FOA-based multinomial naive bayes and FOA-based SVM according to the results shown in Table 8. The precision and recall of the system are also quite good for this complex unstructured text classification task. Fig. 5 shows the bar chart for the performance indicators (5-fold cross-validation) of the overall resource mobilization system. The confidence level we got using two-tailed paired t-test is more than 99 percent in this case which may be considered as a good one for our task.

Table 7
Measures on relevant vs irrelevant tweet.

Classifier		Avg. result of 4 different ratios				5-fold cross validation				10-fold cross validation			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
KNN	k = 3	79.38	70.87	79.26	74.83	79.29	70.78	79.17	74.74	79.21	70.69	79.06	74.64
	k = 5	82.49	74.58	80.16	77.27	82.36	74.49	80.06	77.17	82.30	74.41	80.02	77.11
	k = 7	81.23	72.31	80.71	76.28	81.07	72.20	80.59	76.16	81.02	72.15	80.51	76.10
SVM	Linear kernel	80.95	73.42	77.52	75.41	80.87	73.31	77.43	75.31	80.81	73.24	77.35	75.24
	RBF kernel	85.26	81.55	80.28	80.91	85.18	81.39	80.21	80.80	85.11	81.34	80.16	80.75
Decision Tree (C4.5)		83.30	79.71	77.37	78.52	83.14	79.57	77.26	78.40	83.06	79.50	77.20	78.33
Multinomial Naive Bayes		86.44	83.64	81.15	82.38	86.26	83.48	81.10	82.27	86.21	83.44	81.03	82.22

Table 8
Measures on the overall system.

Classifier		Avg. result of 4 different ratios				5-fold cross validation				10-fold cross validation			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
KNN	k = 3	69.58	67.29	59.29	63.04	69.47	67.22	59.22	62.97	69.44	67.11	59.16	62.88
	k = 5	71.09	72.58	62.49	67.16	71.01	72.53	62.43	67.10	70.97	72.46	62.39	67.05
	k = 7	71.04	71.30	59.78	65.03	70.97	71.23	59.70	64.96	70.93	71.15	59.65	64.89
SVM	Linear kernel	70.51	70.15	58.40	63.74	70.42	70.06	58.32	63.65	70.36	69.95	58.30	63.60
	RBF kernel	73.97	73.71	65.09	69.13	73.88	73.64	65.03	69.07	73.84	73.57	65.00	69.02
Decision Tree (C4.5)		71.88	71.87	63.68	67.53	71.81	71.81	63.62	67.47	71.75	71.75	63.58	67.42
Multinomial Naive Bayes		75.72	72.17	66.51	69.22	75.70	72.11	66.43	69.15	75.66	72.04	66.41	69.11

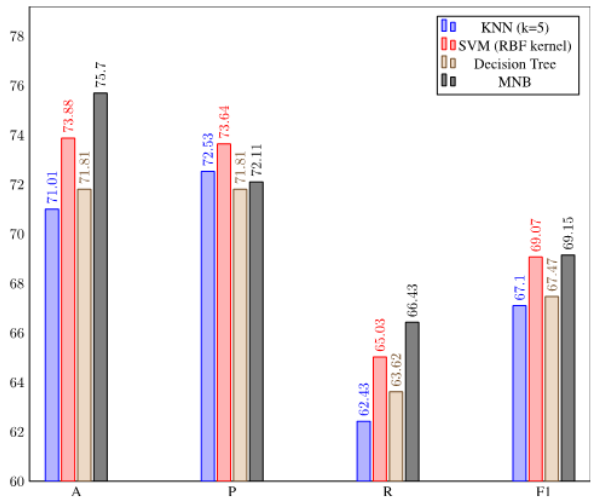

Fig. 5. 5-fold cross validation measures on the overall system.

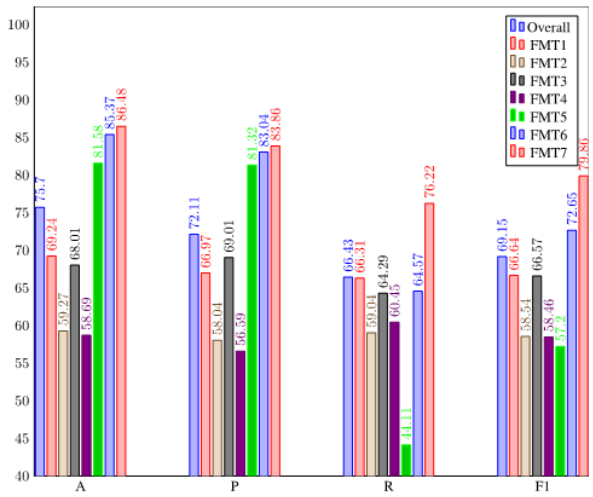

Fig. 6. 5-fold cross validation measures on each topic using FOA based Multinomial Naive Bayes classifier.

Here, in Table 9 classification measures of all the individual topics on FOA wrapper based multinomial naive bayes system are mentioned. As mentioned in Table 8, the overall accuracy, precision, recall, and F1-score are given in the first row of Table 9. It is observed that the results are better in case of availability of resources as compared to need of resources which may be due to the presence of negative words like not, no, do not, and many more in the tweets. It is giving more than 80 percent of accuracy and precision in all the three classes of FMT5, FMT6, and FMT7. Fig. 6 presents the bar chart for the performance measures (5-fold cross-validation) as obtained by different classifiers for categorization of 7 individual topic.

Table 10 shows the performance of our proposed system with various classifiers for the classification of availability or requirement of general resources like food, shelter, water, etc. Here also multinomial naive bayes with FOA gives better accuracy and F1-score than the rest. The classification accuracies of general resources are nearly equal to the overall accuracies, which may be due to the mention of positive and negative words (e.g., have food and haven't food, need water and needn't water) in the tweets. Our proposed system achieves a standard F1-score of 70.16 percent in FOA based Multinomial Naive Bayes classifier. The common issues in misclassification like spelling errors, the sparseness of the term presence matrix, the presence of negative words, and so forth have already been taken care of to the possible extent in the lexical normalization, feature selection, and preprocessing step respectively. It is observed that still few tweets are getting misclassified mostly due to the presence of negative terms such as not, no, do not, etc. (for example "not required" implies available but in some cases it is misclassified into need class). This misclassification task still remains as a great challenge mostly in the sentiment analysis of Twitter data. We are trying to handle these negative words in future by analyzing the tweets semantically in the preprocessing step. Fig. 7 manifests the bar

Table 9
Measures on each topic using FOA based Multinomial Naive Bayes classification system.

Class	Avg. result of 4 different ratios				5-fold cross validation				10-fold cross validation			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
Overall	75.72	72.17	66.51	69.22	75.70	72.11	66.43	69.15	75.66	72.04	66.41	69.11
FMT1	69.31	67.04	66.38	66.71	69.24	66.97	66.31	66.64	69.17	66.86	66.22	66.54
FMT2	59.33	58.12	59.13	58.62	59.27	58.04	59.04	58.54	59.19	57.91	58.91	58.41
FMT3	68.05	69.09	64.36	66.64	68.01	69.01	64.29	66.57	67.97	68.87	64.17	66.44
FMT4	58.76	56.65	60.52	58.52	58.69	56.59	60.45	58.46	58.65	56.44	60.33	58.32
FMT5	81.67	81.4	44.14	57.24	81.58	81.32	44.11	57.20	81.51	81.19	44.02	57.09
FMT6	85.45	83.11	64.63	72.71	85.37	83.04	64.57	72.65	85.28	82.88	64.49	72.54
FMT7	86.53	83.94	76.27	79.92	86.48	83.86	76.22	79.86	86.42	83.72	76.13	79.74

Table 10
Measures on general resources available (FMT1) vs required (FMT2).

Classifier		Avg. result of 4 different ratios				5-fold cross validation				10-fold cross validation			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
KNN	k = 3	70.08	68.15	57.30	62.26	70.02	68.09	57.27	62.21	69.68	67.87	57.03	61.98
	k = 5	74.83	72.26	61.91	66.69	74.79	72.21	61.89	66.65	74.47	72.08	61.68	66.48
	k = 7	74.36	72.04	61.05	66.09	74.33	71.99	61.01	66.05	74.02	71.74	60.72	65.77
SVM	Linear kernel	72.69	71.41	60.48	65.49	72.65	71.34	60.41	65.42	72.38	71.18	60.14	65.20
	RBF kernel	76.92	76.67	64.26	69.92	76.87	76.58	64.63	70.10	76.59	76.46	64.43	69.93
Decision Tree(C4.5)		74.71	73.35	64.52	68.65	74.64	73.29	64.47	68.60	74.37	73.19	64.32	68.47
Multinomial Naive Bayes		77.08	77.13	64.34	70.16	77.05	76.95	64.32	70.07	76.97	76.91	64.26	70.02

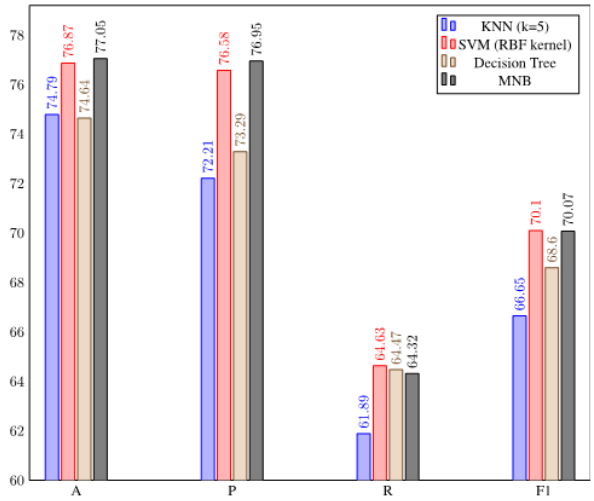

Fig. 7. 5-fold cross validation measures on general resources available (FMT1) vs required (FMT2).

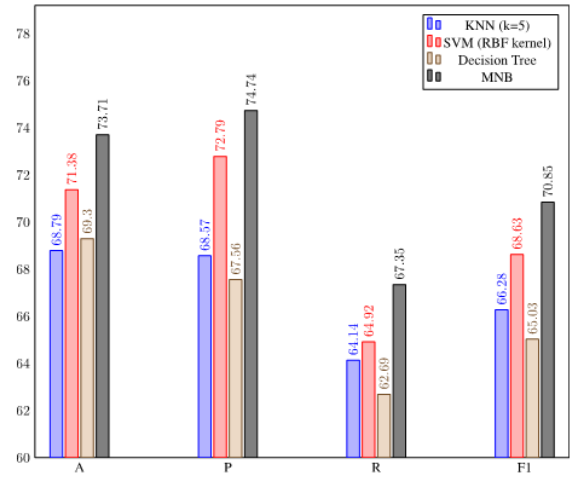

Fig. 8. 5-fold cross validation measures on medical resources available (FMT3) vs required (FMT4).

chart for the performance measures (5-fold cross-validation) obtained by several classifiers for general resource available vs. required tweet classification.

We have used various classifiers such as multinomial naive bayes, KNN, SVM, and decision tree to classify the implication of availability or requirement of medical resources in a tweet. The F1-score and all other performance measures achieved by the above mentioned classifiers are listed in Table 11. Here, the overall classification accuracy is better in comparison to the medical resource need or available. The difference in F1-score of KNN and decision tree based classifier is marginal. Here also FOA-based multinomial naive bayes classifier gives the highest F1-score of 70.93 percent. Fig. 8 shows the bar chart for the performance measures (5-fold cross-validation) obtained by different classifiers for medical resource available vs. required tweet classification.

The classification accuracy of location, organization or damage related tweets are illustrated in Table 12. The accuracies shown here are better than the overall accuracies of the system which may be due to the lack of negative terms. Here, multinomial naive bayes classifier with FOA-based feature selection technique outperforms the other classifiers. In location related tweets, the named entity recognition accuracy is reasonable due to the presence of regional location names present in the tweet. However, the accuracy of our system for location is relatively low w.r.t. other classes like organization and damage. Fig. 9 presents the bar chart for the performance indicators (5-fold cross-validation) as obtained by numerous classifiers for location vs. organization vs. damage mentioned tweet categorization.

To check the robustness of the proposed FOA-MNB model, the standard deviation of f-measure is estimated w.r.t to all runs in Table 13. Instead of accuracy, f-measure is chosen as it is more reliable on unbalanced datasets. From the results it is found that the standard deviations are below 1 percent on both average of four different ratios and 5-fold cross validation. However, in 10-fold cross validation the standard deviation is found to be just above 1 percent which may be due to lowering of testing samples in this case. This less value of

Table 11
Measures on medical resources available (FMT3) vs required (FMT4).

Classifier		Avg. result of 4 different ratios				5-fold cross validation				10-fold cross validation			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
KNN	k = 3	65.62	65.50	58.08	61.57	65.59	65.44	58.07	61.54	65.41	65.29	58.02	61.44
	k = 5	68.81	68.62	64.20	66.34	68.79	68.57	64.14	66.28	68.63	68.41	63.99	66.13
	k = 7	67.96	68.85	61.78	65.12	67.91	68.78	61.75	65.08	67.74	68.57	61.68	64.94
SVM	Linear kernel	67.65	68.14	60.73	64.22	67.58	68.06	60.64	64.14	67.39	67.88	60.55	64.01
	RBF kernel	71.43	72.86	64.98	68.69	71.38	72.79	64.92	68.63	71.22	72.60	64.79	68.47
Decision Tree (C4.5)		69.34	67.65	62.71	65.08	69.30	67.56	62.69	65.03	69.07	67.41	62.61	64.92
Multinomial Naive Bayes		73.76	74.80	67.44	70.93	73.71	74.74	67.35	70.85	73.50	74.58	67.26	70.73

Table 12
Measures on location (FMT5) vs organization (FMT6) vs damage (FMT7).

Classifier		Avg. result of 4 different ratios				5-fold cross validation				10-fold cross validation			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
KNN	k = 3	69.62	67.49	55.16	60.71	69.59	67.45	55.11	60.66	69.31	67.17	54.86	60.39
	k = 5	75.54	73.55	64.02	68.45	75.50	73.52	63.97	68.41	75.24	73.24	63.76	68.17
	k = 7	75.29	72.18	61.46	66.39	75.18	72.13	61.37	66.32	74.85	71.78	61.13	66.03
SVM	Linear kernel	73.71	70.41	60.58	65.13	73.69	70.34	60.46	65.03	73.36	70.05	60.15	64.72
	RBF kernel	78.02	77.68	64.62	70.55	77.99	77.65	64.57	70.51	77.72	77.38	64.37	70.28
Decision Tree (C4.5)		74.61	72.91	67.28	69.98	74.52	72.86	67.19	69.91	74.28	72.53	66.93	69.62
Multinomial Naive Bayes		79.88	75.35	68.54	71.78	79.73	75.31	68.43	71.71	79.44	75.03	68.27	71.49

Table 13
Robustness of the proposed model with estimation of standard deviation for f-measure (F1).

Task	Standard deviation of f-measure (F1) on FIRE dataset		
	Avg. result of 4 different ratios	5-fold cross validation	10 fold cross validation
Disaster vs non-disaster	0.89	0.90	1.03
Relevant vs non-relevant	0.97	0.95	1.06
Overall system	0.95	0.97	1.11
FMT1 vs FMT2	0.87	0.88	1.02
FMT3 vs FMT4	0.88	0.90	1.04
FMT5 vs FMT6 vs FMT7	0.96	0.95	1.07

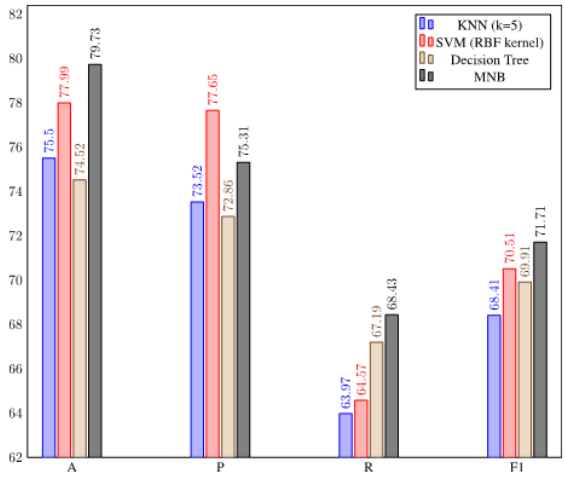

Fig. 9. 5-fold cross validation measures on location (FMT5) vs organization (FMT6) vs damage (FMT7).

standard error in all three types of train-test split makes the model robust for all the tasks performed to evaluate the proposed system. However, 5-fold cross-validation is strongly recommended for its low standard error along with high accuracy and f1-score.

To validate our results statistically we have applied the Friedman test which is one of the most popular non-parametric statistical test (Demšar, 2006; Derrac, García, Molina, & Herrera, 2011). From the F1-measure results (avg. results of all 4 different ratios) of Table 6 the Friedman test is performed as per the formula given below:

$$FM = (12/(n * k * (k+1)) * (\sum_{i=1}^{k} R_i^2) - 3 * n * (k+1)) \quad (16)$$

n = number of datasets
k = number of classifiers
R_i = total ranks for ith dataset

As we are comparing the results of 7 classifiers on 3 datasets, the FM value is calculated as 17.46.

The p-value is found to be 0.00773. We have chosen significance level (α) as 0.05. The null hypothesis is rejected as the p-value is less than the significance level. Hence, from the alternate hypothesis it is observed that the results are significant.

Each of the machine learning based model have some strength and limitations. Regarding the strength our system is quite simple and very effective for this challenging disaster resource mobilization task. Our proposed model handles various noises present in social media text and performs very good as observed from all performance measures. Unlike embedding based models our model is not so much data and resource dependent. If we highlight the limitation, then there is still further scope for improvement in the misclassification rate. In our future work we will try to address it using robust semantic analysis techniques. If there is complete failure of Internet then the system will face challenge as it will not get the required amount of information from people in that affected area. We may think of collecting information from neighboring areas. Hopefully, we will come out with a concrete solution for this situation in our future work.

7. Conclusion and future work

In this work, we have proposed a system prototype for the classification of tweets aiming at assisting the resource management task

during natural disasters. We first pre-process the tweets using natural language processing techniques, then normalize the lexical variant out-of-vocabularies using contextual and string similarity information. Further, normalized tweets are transformed into a tf_idf matrix to run on different state-of-the-art classifiers. At last, the proposed system is experimented on Nepal earthquake tweets and it is found that the proposed FOA wrapper-based multinomial naive bayes classification system can be beneficial during natural emergencies due to its reasonably higher accuracy. This improvement in performance is mainly due to better feature selection and dimension reduction, which further improves the execution time. We have taken a less number of tagged tweets due to unavailability of tagged data and we expect that in the future accuracy of the classifier can be further improved by increasing the number of tagged tweets. It still has plenty of scope for improvement if context information can be exploited properly and also by reducing the preprocessing and normalization error. In our future work, along with the syntactic features we will be using some robust semantic analysis techniques to reduce the misclassification rate and further improve the performance of our system.

Acknowledgments

The authors wish to thank the Forum for Information Retrieval Evaluation (FIRE) and Social Media for Emergency Relief and Preparedness (SMERP) for providing the Nepal and Italy earthquake tweet datasets. The authors also wish to thank Prof. Carlos A. Iglesias for his valuable comments to improve the paper.

All authors approved the version of the manuscript to be published.

References

Andrews, S., Gibson, H., Domdouzis, K., & Akhgar, B. (2016). Creating corroborated crisis reports from social media data through formal concept analysis. *Journal of Intelligent Information Systems*, *47*, 287–312.

Aphinyanaphongs, Y., Fu, L. D., Li, Z., Peskin, E. R., Efstathiadis, E., Aliferis, C. F., et al. (2014). A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. *Journal of the Association for Information Science and Technology*, *65*, 1964–1987.

Arif, M. H., Li, J., & Iqbal, M. (2017). Solving social media text classification problems using code fragment-based XCSR. In *2017 IEEE 29th international conference on tools with artificial intelligence* (pp. 485–492). IEEE.

Bakliwal, A., Foster, J., van der Puil, J., O'Brien, R., Tounsi, L., & Hughes, M. (2013). Sentiment analysis of political tweets: Towards an accurate classifier. In *Proceedings of the workshop on language analysis in social media* (pp. 49–58).

Baliarsingh, S. K., Vipsita, S., & Dash, B. (2020). A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm. *Neural Computing and Applications*, *32*, 8599–8616.

Behl, S., Rao, A., Aggarwal, S., Chadha, S., & Pannu, H. (2021). Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises. *International Journal of Disaster Risk Reduction*, *55*, Article 102101.

Bontcheva, K., Derczynski, L., Funk, A., Greenwood, M. A., Maynard, D., & Aswani, N. (2013). TwitIE: An open-source information extraction pipeline for microblog text. In *RANLP* (pp. 83–90).

Boussaïd, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization metaheuristics. *Information Sciences*, *237*, 82–117.

Caragea, C., Kim, H., Mitra, P., & Yen, J. (2010). Classifying text messages for emergency response. In *Proceedings of NIPS workshop on machine learning for social computing*. Whistler, BC, Canada.

Castillo, C. (2016). *Big crisis data: Social media in disasters and time-critical situations*. Cambridge University Press.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*, 16–28.

Choi, S., & Bae, B. (2015). The real-time monitoring system of social big data for disaster management. *Computer Science and Its Applications*, *330*, 809–815.

Cortes, C., & Vapnik, V. (1995). Support vector machine. *Machine Learning*, *20*, 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*, 21–27.

Cresci, S., Cimino, A., Dell'Orletta, F., & Tesconi, M. (2015). Crisis mapping during natural disasters via text analysis of social media messages. In *WISE (2)* (pp. 250–258).

Delizo, J. P. D., Abisado, M. B., & De Los Trinos, M. I. P. (2020). Philippine Twitter sentiments during Covid-19 pandemic using multinomial Naïve-Bayes. *International Journal*, *9*.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, *1*, 3–18.

Dokeroglu, T., Sevinc, E., Kucukyilmaz, T., & Cosar, A. (2019). A survey on new generation metaheuristic algorithms. *Computers & Industrial Engineering*, *137*, Article 106040.

Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., et al. (2013). Entity extraction, linking, classification, and tagging for social media: A Wikipedia-based approach. *Proceedings of the VLDB Endowment*, *6*, 1126–1137.

Ghaemi, M., & Feizi-Derakhshi, M. -R. (2014). Forest optimization algorithm. *Expert Systems with Applications*, *41*, 6676–6687.

Ghaemi, M., & Feizi-Derakhshi, M. -R. (2016). Feature selection using forest optimization algorithm. *Pattern Recognition*, *60*, 121–129.

Goolsby, R. (2010). Social media as crisis platform: The future of community maps/crisis maps. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *1*, 7.

Han, B., Cook, P., & Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*, 5.

Hossain, E., Sharif, O., & Hoque, M. M. (2020). Sentiment polarity detection on Bengali book reviews using multinomial Naive Bayes. arXiv preprint arXiv:2007.02758.

Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., et al. (2015). Social media and disasters: A functional framework for social media use in disaster planning, response, and research. *Disasters*, *39*, 1–22.

Hughes, A. L., & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, *6*, 248–260.

Hussain, K., Salleh, M. N. M., Cheng, S., & Shi, Y. (2019). Metaheuristic research: A comprehensive survey. *Artificial Intelligence Review*, *52*, 2191–2233.

Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition*, *38*, 2270–2285.

Jiang, J. -Y., Liou, R. -J., & Lee, S. -J. (2011). A fuzzy self-constructing feature clustering algorithm for text classification. *IEEE Transactions on Knowledge and Data Engineering*, *23*, 335–349.

Kabir, M. Y., Gruzdev, S., & Madria, S. (2020). STIMULATE: A system for real-time information acquisition and learning for disaster management. In *2020 21st IEEE international conference on mobile data management* (pp. 186–193). IEEE.

Kang, J., Choi, H., & Lee, H. (2019). Deep recurrent convolutional networks for inferring user interests from social media. *Journal of Intelligent Information Systems*, *52*, 191–209.

Karagoz, P., Kama, B., Ozturk, M., Toroslu, I. H., & Canturk, D. (2019). A framework for aspect based sentiment analysis on turkish informal texts. *Journal of Intelligent Information Systems*, *53*, 431–451.

Kersten, J., Bongard, J. H., & Klan, F. (2021). Combining supervised and unsupervised learning to detect and semantically aggregate crisis-related Twitter content. In *International conference on information systems for crisis response and management*.

Khalifa, M. B., Redondo, R. P. D., Vilas, A. F., & Rodríguez, S. S. (2016). Identifying urban crowds using geo-located social media data: A Twitter experiment in New York City. *Journal of Intelligent Information Systems*, *2*, 287–308.

Khotimah, N., & Wasono, R. (2020). Sentiment analysis of E-commerc brand review using multinomial text Naïve Bayes. In *International conference on education*: *Vol. 2*.

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–a systematic literature review. *Information and Software Technology*, *51*, 7–15.

Kostrzewa, D., & Brzeski, R. (2019). The data dimensionality reduction and features weighting in the classification process using forest optimization algorithm. In *Asian conference on intelligent information and database systems* (pp. 97–108). Springer.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., et al. (2016). Rapid assessment of disaster damage using social media activity. *Science Advances*, *2*, Article e1500779.

Lachlan, K. A., Spence, P. R., Lin, X., Najarian, K., & Del Greco, M. (2016). Social media and crisis management: CERC, search strategies, and Twitter content. *Computers in Human Behavior*, *54*, 647–652.

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 2267–2273). AAAI Press.

Lourentzou, I., Manghnani, K., & Zhai, C. (2019). Adapting sequence to sequence models for text normalization in social media. In *Proceedings of the international AAAI conference on web and social media*: *Vol. 13*, (pp. 335–345).

Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with Naive Bayes-which Naive Bayes? In *CEAS*: *Vol. 17*, (pp. 28–69).

Middleton, S. E., Middleton, L., & Modafferi, S. (2014). Real-time crisis mapping of natural disasters using social media. *IEEE Intelligent Systems*, *29*, 9–17.

Mohanty, F., Rup, S., Dash, B., Majhi, B., & Swamy, M. (2018). Mammogram classification using contourlet features with forest optimization-based feature selection approach. *Multimedia Tools and Applications*, 1–30.

Naseem, U., Razzak, I., Musial, K., & Imran, M. (2020). Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, *113*, 58–69.

Naz, M., Zafar, K., & Khan, A. (2019). Ensemble based classification of sentiments using forest optimization algorithm. *Data*, *4*, 76.

Ngai, E. W. T., Tao, S. S. C., & Moon, K. K. L. (2015). Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management*, *35*, 33–44.

Nguyen, L., Yang, Z., Zhu, J., Li, J., & Jin, F. (2018). Coordinating disaster emergency response with heuristic reinforcement learning. arXiv preprint arXiv:1811.05010.

Nouri-Moghaddam, B., Ghazanfari, M., & Fathian, M. (2020). A novel filter-wrapper hybrid gene selection approach for microarray data based on multi-objective forest optimization algorithm. *Decision Science Letters*, *9*, 271–290.

Nouri-Moghaddam, B., Ghazanfari, M., & Fathian, M. (2021). A novel multi-objective forest optimization algorithm for wrapper feature selection. *Expert Systems with Applications*, *175*, Article 114737.

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing* (pp. 994–1009). ACM.

Parikh, R., & Movassate, M. (2009). *Sentiment analysis of user-generated twitter updates using various classification techniques*: *CS224N Final Report, 118*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Reynard, D., & Shirgaokar, M. (2019). Harnessing the power of machine learning: Can Twitter data be useful in guiding resource allocation decisions during a natural disaster? *Transportation Research, Part D (Transport and Environment)*, *77*, 449–463.

Sharupa, N. A., Rahman, M., Alvi, N., Raihan, M., Islam, A., & Raihan, T. (2020). Emotion detection of Twitter post using multinomial Naive Bayes. In *2020 11th International conference on computing, communication and networking technologies* (pp. 1–6). IEEE.

Simon, T., Goldberg, A., & Adini, B. (2015). Socializing in emergencies—A review of the use of social media in emergency situations. *International Journal of Information Management*, *35*, 609–619.

Smith, L., Liang, Q., James, P., & Lin, W. (2017). Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. *Journal of Flood Risk Management*, *10*, 370–380.

Spielhofer, T., Greenlaw, R., Markham, D., & Hahne, A. (2016). Data mining Twitter during the UK floods: Investigating the potential use of social media in emergency management. In *Information and communication technologies for disaster management, 2016 3rd international conference on* (pp. 1–6). IEEE.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., & Demirbas, M. (2010). Short text classification in twitter to improve information filtering. In *Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval* (pp. 841–842). ACM.

Tabakhi, S., Moradi, P., & Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence*, *32*, 112–123.

Velev, D., & Zlateva, P. (2012). Use of social media in natural disaster management. *International Proceedings of Economics Development and Research*, *39*, 41–45.

Wang, D., Zhang, H., Liu, R., Lv, W., & Wang, D. (2014). t-Test feature selection approach based on term frequency for text categorization. *Pattern Recognition Letters*, *45*, 1–10.

Xu, Z., Liu, Y., Yen, N., Mei, L., Luo, X., Wei, X., et al. (2016). Crowdsourcing based description of urban emergency events using social media big data. *IEEE Transactions on Cloud Computing*, 1–11.

Zhang, Y., & Desouza, P. (2014). Enhance the power of sentiment analysis. *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *8*, 421–426.

Zhang, Y., Wang, S., Phillips, P., & Ji, G. (2014). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. *Knowledge-Based Systems*, *64*, 22–31.

Zielinski, A., Middleton, S. E., Tokarchuk, L. N., & Wang, X. (2013). Social media text mining and network analysis for decision support in natural crisis management. In *ISCRAM* (pp. 840–845).

Zubiaga, A. (2020). Exploiting class labels to boost performance ing-based text classification. In *Proceedings of the 29th ACM interna ence on information & knowledge management* (pp. 3357–3360).