*Predicting Real Disaster from Twitter with Machine Learning*

**Md. Jakaria**  
*Computer Science & Engineering*  
*Pabna University of Science and Technology*Pabna-6600, Bangladesh  
jakaria.pust@gmail.com

**Md. Khaled Ben Islam**  
Assistant Professor  
*Computer Science & Engineering   
Pabna University of Science and Technology*

Pabna-6600, Bangladesh  
mdkhaledben@gmail.com

*Abstract*— The number and impact of disaster seem to be increasing in the last decades and their consequences have to be managed in the best possible way. As blessing a lot of people, nowadays uses social media and express their opinions and thoughts in real time. Because of this, more agencies, including disaster relief organizations, news agencies are interested in automated way to monitoring Twitter. So the goal of this research is to build a model to predict which tweets are about real disaster and which ones are not using the machine learning algorithms. To help the disaster affected people, to reduce the harmful effects of disaster, to help the disaster management agents it is very important to know details about the disaster, types of disaster whether it's about a real disaster or not. This is the purpose of this research work. Several Machine learning algorithms have used and combined prediction result with Voting Class Classifier. The system is working well, with its tweet classification accuracy of 79%. The study has the potential to facilitate disaster managers for better response operations during emergencies or post emergency period.

Keywords—prediction, real disaster, disaster management, twitter, social media.

# Introduction

Communication could be a center part of disaster planning, response and recovery. Effective disaster communication may prevent a disaster or reduce its impact, whereas ineffective disaster communication may cause a disaster or make its effects worse. So it’s very important things to deal with disaster. These days, information and communication technology is being utilized broadly during various periods of disaster for relief activities [1]. Social media, such as Twitter and Facebook, plays a critical role in disaster management. It is ranked as the fourth most popular source for accessing emergency information [2]. With the availability of mobile phones people are getting involved with social media like Twitter. So there exits plethora of data. But most of the data are in the form of unstructured data [3]. So it’s a big challenge to work with those unstructured data. Before using those data in a programmed frame work model they should be in the form of structured data. In our thesis we have predicted which tweets are about real disaster and which ones are not. We have collected data from twitter as our data set.

Researchers have found that social media like Twitter do better than regular news media or government emergency services [5], including the East Japan Earthquake [4], to distribute information in emergencies. In reality, Twitter is used during a large-scale fire emergency and live traffic updates for multiple real-time messages such as the one required for support. They also include storm, fire, traffic jam, riots, heavy rainfall, and earthquakes. In previous Sakaki developed an earthquake reporting system via Twitter messages [6] [7]. As reported by the Japan Meteorological Agency (JMA), their system was able to detect 93 percent of earthquakes. They had used simple linguistic features like word count and target event word context etc. to train an Earthquake detection classifier. Several new measures were also proposed for task-based performance measurement of event detection techniques. To measure run-time performance they used the data stream management system to implement all available event detection techniques [8]. A good prediction and warning can save lives on the other hand a wrong information can mislead a team. For the purpose of disaster management, different Disaster management agents like local level, national level, International level or NGOs have to keep eye on social media for helping the affected people. But it is so difficult for a man to keep an eye always on social media. On the other hand, it’s not always clear whether a person’s words are actually announcing a disaster. So it’s crucial that the prediction of disaster is real or not. To help those disaster affected people, Government agencies, News agencies, NGOs we have done this research work and we have a good accuracy rate in our model.

# Methodology

Machine learning algorithms have been used to predict which Tweets are about real disasters and which ones are not about disasters. The methodology in the present study has a number of steps, which have been described in the subsequent sections.

## Data

In training our model and validating its accuracy, adequate amounts of Twitter data are important. In our work we have collected tweets about different types of disaster. We have used Twitter API with tweepy python library to capture live tweets related to Disaster like flood, hailstorm, bushfire, earthquake etc. Over 8000 thousands tweets have been collected from different locations and for different types of disasters and the collected tweets were in English language. In our dataset there exists total five columns- Id, Location, Text, Keyword and Target (human specified it may 0 or 1 in binary). After that, data set needs to be split into two sets training set and testing set.

## Data Preprocessing

According to IBM, [9] it is estimated that around 80% of all information is unstructured. Tweets contain various sorts of noise and redundancies for example, emoji’s, user mentions, Internet links and so forth. A proper data pre-processing is needed in order to use these tweets for any meaningful purpose.

* Missing data: In our data set there exits some missing values. We need to ignore the row related to those missing values.
* Remove unwanted words: As our text column contains unwanted words like #, =>, numbers, signs, mentions etc. these letters will not be useful in our problem. So we will get only pure text without any markings or numbers. This work has done by specifying our pattern using re library.
* Transform words to lowercase: Transform words to lowercase because upper and lower case have different ASCII codes.
* Remove stop words: Stop words are usually the most common words in a language and they will be irrelevant in the model.
* Stemming words: Stemming is the process of reducing words to their word stem, base or root form.
* Making Corpus: making corpus is important for saving the cleaned data. So that the further works can be done with those data.

## Feature Extraction

In our dataset Location field was not suitable for selecting feature as there were too many unique values in it. That’s why Keyword were used as a feature selection. In the system we had used TFIDF (Term Frequency Inverse Document Frequency) feature selection methods for finding the most important texts from our data set.

## Model

Hence, we are developing a system based on machine learning to categorize disaster and non-disaster tweets. We need to do feature selection and extraction first. After extracting features, we can build model for the prediction. We have used 4 different type of machine learning algorithms like K-Nearest Neighbors, Logistic Regression, Support Vector Machine, Naive Bayes for our text classifier model. In our system, we have combined all the algorithms prediction results by Voting Classifier majority voting. The classifiers are implemented in Python using the machine learning software library called scikit-learn. Also we need to import libraries such as pandas, numpy, nltk. We built the data frame using pandas, so that we can manage our dataset. Then we converted all upper case word to lower case in the dataset. To delete a certain regular expression from our data, we imported PorterStemmer to stem our data. Then removed stop-words from our data set so that the analysis of our data becomes simple. After all these preprocessing we used Tf-IdfVectorizer to vectorize our data. The classifier trained using the training data. Fit method used for training it. After that our model was ready to make predictions.

# Result

With the help of confusion matrix, we had calculated F1 score, Precision, Recall, Receiver operating characteristic (ROC) curve for the measurement of classifiers performances. Confusion matrix gives us insight not only into the mistakes made by a classifier but more importantly, the types of mistakes made. The high priority and low priority levels are class 0 and class 1. Precision is the fraction of actual class data properly categorized, to the total data listed as real. In our case how many tweets are actually high priority out of total tweets, and vice versa. It is represented in Eq. 1

(1)

Where TP is the number of true class samples classified as true, and FP is the number of false class samples classified as true. Recall is the proportion of true data listed as true, to the total number of true data. It is represented in Eq. 2

(2)

Where FN is the number of true samples classified as false. F1 Score is the cumulative mean of precision and recall, bundling precision and recall in a single index. It is represented in Eq. 3

(3)

ROC curve is the plot for different thresholds between the classifier's true positive rate and the false positive rate. The output of the classifier is more reliable, if the curve follows the left border and then the top border of the ROC space. The resulting classification is described in TABLE I for the classifiers in terms of the precision, recall and F1-score.

TABLE I

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifiers | Class | Precision | Recall | F1 score |
| KNN | 0 | .70 | .94 | .80 |
| 1 | .82 | .42 | .56 |
| LR | 0 | .79 | .85 | .82 |
| 1 | .76 | .68 | .72 |
| SVC | 0 | .80 | .84 | .82 |
| 1 | .75 | .69 | .72 |
| BNB | 0 | .79 | .85 | .82 |
| 1 | .76 | .67 | .71 |
| GNB | 0 | .75 | .88 | .81 |
| 1 | .77 | .57 | .66 |
| MNB | 0 | .79 | .85 | .82 |
| 1 | .76 | .67 | .71 |
| Voting | 0 | .78 | .88 | .83 |
| 1 | .79 | .65 | .71 |

From the TABLE I we found that Logistic Regrassion and Support Vector Machine performed really well. They had 0.72 F1 score. Naive byes classifiers and Knn also performed well for our prediction model. After that we had combined all the classifiers result with voting classifier by majority voting. In “Fig.1”, the Confusion matrix and Roc curve of Voting classifier are represented. The confusion matrix indicates how many samples are categorized according to which class. Of the 1523 samples used for processing, the voting classifier classified 109 class 0 (high priority) samples as class 1 (low priority) which is represented in “Fig. 1”. Roc curve is also showed the models performance. With the analysis of different types of algorithms we got the Accuracy result from them. AS measuring their performance of Recall, Precision, and F1 Score we had gotten the Logistic Regression Classifier Algorithm had a better Accuracy resultant of 78%. But we had combined all algorithm results and got total 79% Accuracy. Which is represented in “Fig.2”.

1. Roc curve and Confusion matrix of voting classifier
2. Comparison of different ML Algorithms

# Discussion and Conclusion

# The current research attempts to use social media effectively in predicting real disaster tweets that will assist with the use of an automatic tweet parsing system during a disaster or emergency situation. In 79 percent cases, users tweeting about general disaster-related information are properly classified. The system received tweets from the tweet status of the peoples. It created some missing data after evaluating Basic Data Extraction we handled it in the manner of the Deleting Columns Technique. Then we pre-processed and appended our twitter data into a new corpus. We rendered Specific Words dictionary and found features for our model. Then the data set was divided into two groups as a training and test collection. Then we had put it into various machine learning algorithms and found that the Logistic Regression Algorithm works better with 78 percent accuracy. Other algorithms such as KNN had 73%, SVC had 78%, MultinomialNB had 78% accuracy level. Ultimately, Voting classifier combined the accuracy score with hard voting and established an accuracy result of 79 percent.

# In this study, we have proposed a system to analyze social media contents in real time while collecting and efficiently managing social media contents. This study also will be useful for making society aware of the dangerous effects of Disaster. It will work as a warning factor for Disaster Management Agencies. In this research, we have only considered the textual content of the tweets to categorize them, ignoring the Internet links (if any) provided in the tweets. The downside here is that these Internet links may lead to websites, which may include additional information or photographs about the disasters. The current research opens some new directions for discovery by other researchers. The classification accuracy of the system was 79%, which can be further enhanced by considering more parameters. The introduction of other languages would improve the system further, as more users express their opinions in their native languages.

##### References

[1] Kabra, G., & Ramesh, A. (2015). “Analyzing ict issues in humanitarian supply chain management: A sap-lap linkages framework. Global Journal of Flexible Systems Management”.

[2] Lindsay, B. R. (2011). “Social media and disasters: Current uses, future options and policy considerations. Congressional research service reports”.

[3] Adnan, Kiran-Akbar, Rehan-2019 “An analytical study of information extraction from unstructured and multidimensional big data”

[4] Graham Neubig and Yuichiroh Matsubayashi-2011, in theier work “Safety Information Mining What can NLP do in a disaster Graham”

[5] Nurwidyantoro, Arif, Winarko, Edi -2013 “Event detection in social media: A survey”

[6] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). “Earthquake shakes twitter users: real-time event detection by social sensors “

[7] Sakaki, T., Okazaki, M., & Matsuo, Y. (2013). “Tweet analysis for real-time event detection and earthquake reporting system development.”

[8] Weiler, A., Grossniklaus, M., & Scholl, M. H. (2016). An evaluation of the run time and task-based performance of event detection techniques for twitter.

[9] The biggest data challenges that you might not even know you have-IBM 15th January (<https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/> ) (last accessed on 15 January 2020)