Disaster Tweets Classification

GAURAV KANOJIA[1], ADITYA RASTOGI[2]
[1, 2] *Maharaja Agrasen Institute of Technology, Sector 24 Rohini, New Delhi, India*

Abstract- ***Social media's enormous data output offers a singular opportunity for disaster research. As a top social media network, Twitter produces more than 500 million Tweets daily. Twitter is increasingly used by authorities to track disaster situations and develop quick rescue plans due to its real-time capability. Building a precise predictive model to recognize disaster Tweets, which might not have enough context due to the length restriction, is difficult. Determining the optimal algorithm to drive a recommendation engine that will aid in real-time crisis occurrences for the purpose of delivering relief to the affected, gathering news, etc. will therefore be the goal of this project.***

I. INTRODUCTION

Here Introduce the paper, and put a nomenclature if necessary, in a box with the same font size as the rest of the paper. The paragraphs continue from here and are only separated by headings, subheadings, images and formulae. The section headings are arranged by numbers, bold and 9.5 pt. Here follows further instructions for authors.

Twitter is social media platform which enables people to communicate their thoughts via text, pictures and videos. It is a platform with more than 1 million daily active users. According to recent data, an approximate of 500 million tweets are posted on Twitter every day.The special aspect of Twitter which makes it more appealing as a forum for us to consider data from is the fact that it has an extremely active and responsive user base.Mostly, all breaking news is posted earlier in twitter than any mainstream media. Hence, this microblogging social network experiences a deluge of information flow during natural disasters. Situation based mining of information from the twitter data, can play a significant role in disaster response and recovery. The large volume and velocity of data flow on twitter during disaster makes it tedious for the disaster rescue volunteers to manually analyze and retrieve information from them. An automated system that could retrieve relevant information from this enormous twitter data during a disaster, could be useful for the disaster relief volunteers to accomplish their duty efficiently amidst the chaos. In this paper, we propose an artificial intelligence based real time disaster response system disaster, which assists the volunteers by identifying the relevant tweets from the real time twitter data and classifying them under the domains. Disaster is empirically validated across various machine learning algorithms for classification using the tweets posted. The versatility of Disaster across different disasters and its improved classification accuracy makes it flexible and robust to handle any location-based emergencies.

We will be comparing Bidirectional Logic Regression, Naïve bayes and Voting TF to see which of the algorithms perform better and can be used in a practical situation in which we will have to classify the tweets as a disaster or not. We will begin by presenting our findings from recent research in the field, then move on to outlining the technique used and the dataset selected.

II. LITERATURE REVIEW

Disasters in our life affect humans in disastrous ways. Throughout our history, mankind has sought to predict, manipulate, and provide relief in times of natural disaster. Our work aims to contribute towards awareness about a disaster in real time and help in disaster relief or for news purposes. We have chosen Twitter as a forum for data collection because we believe that it contains accurate and relevant data for disaster relief purposes. Social media data has been analyzed in multiple studies so that they can be put to good use during disasters in real time[1][2]. Present studies use geographical tags from tweets to calculate the influence of social media in determining reaction of people in disasterstricken areas in order to explore relevant reasons.

In [3] Beverly used dataset consisting of data obtained from tweets during Habagat flooding in Manila. They used Naive Bayes and Support Vector Machines as classification models Similarly, Guizhe [4] used a dataset from Kaggle and developed a model which is aware of sentiments contextual model named Senti BERT-BiLSTM-CNN. Ak Ningsih [5] applied logistic regression on the disaster tweets dataset from Kaggle. Kevin Stowe [6] used data in the form of tweets related to Hurricane Sandy and created an algorithm for classifying the same. WinduGata [7] analyzed data from the Indonesia earthquake disaster and applied Naive Bayes with gini index and using smote upsampling imbalance techniques. Study [8] used Logistic Regression for detecting tweets that were automatically generated. Mozafari et al. [9] used transfer learning paradigm on the model to recognize hateful tweets. In order to detect sarcasm, Eke et al. [10] used voting classification so that it would be helpful in sentiment analysis tasks. Algur et al. [11] used classification algorithms like Naive Bayes, Logistic Regression, J48, Random Forest and SVM on tweets that had been converted into vectors with the help of count vectorizer and TF-IDF .Rinaldo Turang [12] aims to develop a web-based application that can classify tweets of netizens into these four categories of topics about health, music, sport, and technology using one of machine learning methods called Logistic Regression.

LanyMaceda [13] earthquake-related tweets collected were classified and manually annotated based on the four (4) labels identified namely, drill/training, earthquake feels, extent of damages, and government measures and rehabilitation. Kelvin KiemaKiilu [14] study to develop a reliable tool for detection of hate tweets. They develops an approach for detecting and classifying hateful speech that uses content produced by self-identifying hateful communities from Twitter. K.Ushasree Santoshi [15] discusses about the ways to analyze the tweets and classify them into spam and ham based on the words involved in tweets. deep learning methods to classify and detect spam tweets like SVM, clustering methods and binary detection models that are used Naïve Bayes classifier. Dasari Siva Krishna [16] collected the seven natural disaster events from the crisisNLP. These datasets are different disaster events which contains the people's opinions on that specific event. We preprocess the information which converts the tweet information into machine understandable vectors. These vectors been processed by the different machine learning algorithms. We consider the individual performance of each ML algorithm on different disaster datasets upon chosen the best five algorithms for voting techniques. Anam Yousaf [17] uses Seven Machine Learning models which are implemented for emotion recognition by classifying tweets as happy or unhappy. With an in-depth comparative performance analysis, it was observed that proposed voting classifier(LR-SGD) with TF-IDF produces the most optimal result.

III. METHODOLOGY

3.1 Dataset

The data we were looking for was available on Kaggle which was -https://www.kaggle.com/c/nlp-getting-started. The data is focus on tweets and disasters and probably comes with a set of labels indicating whether a tweet is a real disaster or not. The data set contain different parameters like id, keyword, location, text and target. In the text.csv we had 7316 tweets in total from which target parameter (1 as disaster and 0 as non – disaster) got us that the 4342 were labelled as disaster tweets and 3271 tweets were labelled non disaster tweets.

3.2 Implementation

Firstly, we started visualizing data before further processing. Visualizing data can be a helpful step in the data processing and analysis process. It allows us to gain insights into the structure and distribution of our data, and can help us identify patterns and trends that may not be immediately obvious from looking at raw data. The ways we use to visualize tweet data, depending on the specific characteristics of the data is visualizing lengths of the tweets (average of 18 words), average words lengths of tweets, most common stop words in the text data ("the") and most common punctuations in the text data (" - ").

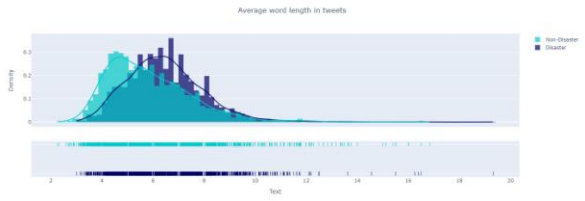

Fig. 1 –Average word lengths in Tweets

Table 1 – Classification Report of Voting Classification

Data	Precision	Recall	F1 - score	Support
0	0.80	0.93	0.86	869
1	0.88	0.68	0.77	654
Accuracy				
Macro Avg	0.84	0.81	0.81	1523
Weighted Avg	0.83	0.82	0.82	1523

Table 2– Classification Report of Logistic Regression

Data	Precision	Recall	F1 - score	Support
0	0.81	0.88	0.84	869
1	0.82	0.73	0.77	654
Accuracy				
Macro Avg	0.82	0.81	0.81	1523
Weighted Avg	0.82	0.82	0.81	1523

Table 3– Classification Report of Naïve Bayes

Data	Precision	Recall	F1 - score	Support
0	0.82	0.85	0.84	869
1	0.88	0.76	0.78	654
Accuracy				
Macro Avg	0.81	0.80	0.81	1523
Weighted Avg	0.81	0.81	0.81	1523

IV. RESULTS

After splitting the data, it was further encode into Bags of words, Term Frequency and Inverse Document Frequency. The feature set that was eventually separately fed into our models gave the following results as mentioned by Table 1, 2 and 3. As stated in Table 1, we get a validation accuracy of 81.6%, validation recall of 82.77%, validation precision of 81.2% and validation F1 of 81.62%.As stated in Table 2, we get a validation accuracy of 81%, validation recall of 80%, validation precision of 79.1% and validation F1 of 81.10%. As stated in Table 3, we get a validation accuracy of 83.1%, validation recall of 82.8%, validation precision of 84% and validation F1 of 83%. We have considered accuracy and recall because we consider these two metrics to be relevant to our goal. Recall was used to minimize the false negatives corresponding to our dataset i.e the number of tweets related to disaster but are classified as non-disaster/normal tweets. We have received good recall and accuracy with our Voting Classification. We get a validation accuracy of 83.1%, validation recall of 82.8%, validation precision of 84% and validation F1 of 83%. Our Voting classification has better performance than our other models in all the metrics, thereby making it the more suitable choice for usage in disaster tweet prediction system.

	Models	Accuracy Score
5	Voting Tf-Idf1	0.822718
3	Logistic Regression Tf-Idf1	0.820749
0	Logistic Regression BoW1	0.816152
4	Naive Bayes Tf-Idf1	0.813526
1	Logistic Regression BoW2	0.740643
2	Logistic Regression BoW3	0.700591

Fig. 2 - Comparison of accuracy of the chosen models.

CONCLUSION AND FUTURE SCOPE

We believe that predicting tweets as disaster or not can help in real time rescue operations during a disaster, for fundraising to provide relief to affected people during a disaster and can also help in reporting of such events. As we can see our Voting classification model has higher recall and accuracy than logistic regression and Naïve bayes model, thereby justifying the preference of Voting classification over other for use in a disaster tweets prediction algorithm in real-time.

The future scope of this paper can be used to overcome some limitations. There can be many ensembles of algorithms that can be trained on the dataset and their predictions can be compared in order to create a better disaster predictor. Fornow we have chosen to train our Voting Classification model on an English dataset but same can be done for datasets in which texts are in a different language, but we will also need model pre-trained in the required language

REFERENCES

[1] Zahera H M, Elgendy I A, Jalota R and Sherif M A 2019 Fine-tuned Model for Multi-Label Tweets Classification TREC 1-7.

[2] Frank M R, Mitchell L, Dodds P S and Danforth C M 2013 Happiness and the patterns of life: A study of geolocated tweets Scientific reports 3 (1) 1-9.

[3] Beverly Estephany Parilla-Ferrer, Proceso L. Fernandez Jr., PhD, and Jaime T. Ballena IV, PhD "Automatic Classification of Disaster-Related Tweets" International conference on Innovative Engineering Technologies (ICIET'2014)

[4] Song, G.; Huang, D. A Sentiment-Aware Contextual Model for Real-Time Disaster Prediction Using Twitter Data. Future Internet 2021,13,163.

[5] A K Ningsih and A I Hadiana 2021 IOP Conf. Ser: Mater. Sci. Eng. 1115 012032[6] Stowe, Kevin & Paul, Michael & Palmer, Martha &Palen, Leysia& Anderson, Kenneth. (2016). Identifying and Categorizing Disaster-Related Tweets.

[6] W. Gata, F. Amsury, N. K. Wardhani, I. Sugiyarto, D. N. Sulistyowati and I. Saputra, "Informative Tweet Classification of the Earthquake Disaster Situation In Indonesia," 2019 5th International Conference on Computing Engineering and Design (ICCED), 2019, pp. 1-6.

[7] Harrag, F.; Debbah, M.; Darwish, K.; Abdelali, A. Bert transformer model for detecting Arabic GPT2 auto-generated tweets.

[8] Mozafari, M.; Farahbakhsh, R.; Crespi, N. A BERT-based transfer learning approach for hate speech detection in online social media. In International Conference on Complex Networks and Their Applications; Springer: Cham, Switzerland, 2019; pp. 928–940.

[9] Eke, C.I.; Norman, A.A.; Shuib, L. Context-Based Feature Technique for Sarcasm Identification in Benchmark Datasets Using Deep Learning and BERT Model. IEEE Access 2021, 9, 48501–48518. [CrossRef].

[10] Algur, S.P.; Venugopal, S. Classification of Disaster Specific Tweets-A Hybrid Approach. In Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 17–19 March 2021; pp. 774–777.

[11] S. T. Indra; Liza Wikarsa; Rinaldo Turang Using logistic regression method to classify tweets into the selected topics 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS).

[12] LanyMaceda; Jennifer Llovido; Arlene Satuito-Categorization of Earthquake-Related Tweets Using Machine Learning Approaches 2018 International Symposium on Computer, Consumer and Control (IS3C).

[13] Kelvin KiemaKiilu, George Okeyo, Richard Rimiru, Kennedy Ogada Using Naïve Bayes Algorithm in detection of Hate Tweets. International Journal of Scientific and Research Publications, Volume 8, Issue 3, March 2018 ISSN 2250-3153.

[14] K.Ushasree Santoshi; S.Sree Bhavya; Y.Bhavya Sri; B. Venkateswarlu Twitter Spam Detection Using Naïve Bayes Classifier 2021 6th International Conference on Inventive Computation Technologies (ICICT).

[15] Dasari Siva Krishna ; Srinivas Gorla ;Prasad Reddy P V G D Disaster tweet classification: A majority voting approach using machine learning algorithms December 2022.\

[16] AnamYousaf; Muhammad Umer; Saima Sadiq; Saleem Ullah; SeyedaliMirjalili; Vaibhav Rupapara Emotion Recognition by Textual Tweets Classification Using Voting Classifier (LR-SGD) IEEE Access 2020 709–713.