Comparative Analysis of Natural Language Processing Techniques for Disaster-Related Tweet Classification

Subham Gupta1, Vivek Verma2, Nikil Rajasekaran Sargunar Bhoopathy3

1. Masters of Applied Computing, University of Windsor, 110093436, gupta2d@uwindsor.ca
2. Masters of Applied Computing, University of Windsor, 110094354, verma17@uwindsor.ca
3. Masters of Applied Computing, University of Windsor, 110093474, bhoopatn@uwindsor.ca

***Abstract***—Tweets are a rich data source in terms of velocity, volume and variety, but to effectively extract information from them is a challenge often talked about. Disaster Related Tweet classification has the potential to redirect emergency services resources to the most critically hit areas in times of natural disasters in an optimized manner. To attain this state, we intend to employ natural processing language techniques in conjunction with Machine Learning and Deep Learning techniques and simultaneously, care about the sustainability of the algorithm and the compute as well.

***Keywords-*** NLP, disaster, classification, twitter,

Machine Learning, Deep Learning.

# I. INTRODUCTION

Twitter is a platform for microblogging. Tweets are short character messages that users post and reply to. Users post about a wide range of topics on Twitter. An estimated 336 million people use Twitter every day and almost 500 million tweets are posted each day, according to a recent survey. Thus, it has a large amount of uncategorized data, which can be used for valuable information discovery.

This research provides a comparative analysis of different Machine Learning and Deep Learning techniques which can categorize tweets about real disaster into 6 categories –Affected Individuals, Infrastructure and Utility Damage, Injured or Dead people, Missing or Found people, Rescue Volunteering or Donation Effort and Vehicle Damage.

# II. PROBLEM STATEMENT

Social Media has become increasingly popular among netizens these days. Social Media text can be categorized and used to predict disasters. When there is a disaster, tweets about it during an emergency contain a variety of information about people who have been hurt or killed, people who are missing or found, infrastructure damage, and utility damage that can assist government agencies and humanitarian organisations in prioritising their aid and rescue efforts. It is crucial to build a model that can categorise these tweets into distinct types to better manage rescue and relief efforts thereby saving lives. For the purpose of categorising disaster-related tweets into six different classes, we want to compare various traditional machine learning and deep learning techniques in this study. Four separate disaster events—hurricane, earthquake, flood, and wildfire—will be used to assess the model’s effectiveness.

# III. MOTIVATION

# During natural disasters, it is crucial for rescue and relief agencies to reach out to the affected population promptly. For government agencies, this task is extremely challenging due to several factors, including inaccurate location information for the victim, the overwhelming number of rescue calls, and prioritizing the rescue activities as per the needs of the victims [1]-[4]. The use of Twitter and Facebook during natural disaster generate a lot of data [5]. The sheer volume of these Tweets makes it challenging for aid agencies to manually sort through each tweet and determine which rescue and relief efforts should be prioritised. Therefore, it is of utmost importance to design a system that can categorize tweets into diverse humanitarian aid categories. Because tweets can only be 280 characters long and frequently contain uncommon acronyms and spelling errors, it is quite difficult to pre-process and automatically categorise them.

# IV. SOLUTION

The idea of categorizing disaster-related tweets has been proposed in several recent works. In comparison to traditional machine learning methods, deep neural network-based models are becoming more and more popular for the task. Transformer and BERT based models are quite extensively used deep learning models whereas Support Vector Machines, Random Forest, Decision Trees, Naïve Bayes, Logistic Regression and Gradient Boosting are most popular classic machine learning algorithm [6]-[11]. The proposed model is an expansion to the current work wherein we not only want to do a comparative analysis of classic machine learning and deep learning algorithm but we also want to explore ensemble learning algorithms like XGBoost, Light GBM and CatBoost because not much of work has been done so far with regards to classification of disaster related tweets using ensemble learning. In the process of classifying the tweets into 6 different class we are also taking into consideration class/data imbalance. Hence it is crucial to compare the effectiveness of traditional and deep learning models when dealing with class imbalance, as the distribution of data is typically skewed toward certain categories. We can clearly see the same in Table1 which shows the count of tweets across different classes for different datasets. Since we also aim at increasing the accuracy of the model, hence we would also like to investigate the use of different combinations of n-gram TF-IDF feature for classic ML algorithms and Glove, Crisis embeddings for representing words in case of Deep Learning Algorithms. Moreover for estimating different model’s performance we will be using different disaster events dataset like – hurricane, earthquake, flood and wildfire[12].

TABLE 1

SUMMARY OF DIFFERENT DATASET

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Classes | Affected Individuals  (AI) | Infrastructure and Utility Damage  (IUD) | Injured or Dead People  (IDP) | Missing or Found People  (MFP) | Rescue Volunteering or Donation Effort  (RVDE) | Vehicle Damage  (VD) |
| Earthquake | 55 | 114 | 204 | 11 | 360 | 2 |
| Flood | 14 | 35 | 18 | 6 | 124 | 0 |
| Hurricane | 328 | 907 | 159 | 15 | 2625 | 50 |
| Wildfire | 75 | 154 | 105 | 8 | 184 | 2 |

# V. REFERENCES

[1]G. Nalluru, R. Pandey and H. Purohit, "Relevancy Classification of Multimodal Social Media Streams for Emergency Services," 2019 IEEE International Conference on Smart Computing (SMARTCOMP), Washington, DC, USA, 2019, pp. 121-125, doi: 10.1109/SMARTCOMP.2019.00040.

[2]Singh, Jyoti Prakash, et al. "Event classification and location prediction from tweets during disasters." *Annals of Operations Research* 283 (2019): 737-757.

[3]Al Hasan Haldar, Nur, et al. "Location prediction in large-scale social networks: an in-depth benchmarking study." *The VLDB Journal* 28 (2019): 623-648.

[4]Sheykhmousa, Mohammadreza, et al. "Post-disaster recovery assessment with machine learning-derived land cover and land use information." *Remote sensing* 11.10 (2019): 1174.

[5] Munawar, Hafiz Suliman, et al. "Big data and its applications in smart real estate and the disaster management life cycle: A systematic analysis." *Big Data and Cognitive Computing* 4.2 (2020): 4.

[6] Balaji, T. K., Chandra Sekhara Rao Annavarapu, and Annushree Bablani. "Machine learning algorithms for social media analysis: A survey." *Computer Science Review* 40 (2021): 100395.

[7] Yu, Manzhu, et al. "Deep learning for real-time social media text classification for situation awareness–using Hurricanes Sandy, Harvey, and Irma as case studies." *International Journal of Digital Earth* 12.11 (2019): 1230-1247.

[8] Kabir, Md Yasin, and Sanjay Madria. "A deep learning approach for tweet classification and rescue scheduling for effective disaster management." *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2019.

[9] Goswami, Shriya, and Debaditya Raychaudhuri. "Identification of Disaster-Related Tweets Using Natural Language Processing: International Conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications (ICAISC-2020)." *IOT, Smart Cities & Applications (ICAISC-2020)(May 26, 2020)* (2020).

[10] Bikku, Thulasi, et al. "Analysis of Disaster Tweets Using Natural Language Processing." *Intelligent Computing and Applications: Proceedings of ICDIC 2020*. Singapore: Springer Nature Singapore, 2022. 491-501.

[11] Maulana, Iqbal, and Warih Maharani. "Disaster tweet classification based on geospatial data using the bert-mlp method." *2021 9th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2021.

[12]Firoj Alam, Ferda Ofli and Muhammad Imran, Crisis MMD: Multimodal Twitter Datasets from Natural Disasters, In Proceedings of the 12th International AAAI Conference on Web and Social Media(ICWSM),2018, Stanford, California, USA.