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***— In a world where disasters strike with alarming frequency, every second counts. Twitter is a goldmine of real-time information during such crises, containing crucial details about casualties, infrastructure damage, and urgent rescue efforts. However, extracting valuable insights from this vast and complex data source is no easy feat. While many research have proposed categorizing disaster-related tweets, little attention has been paid to the crucial issue of class imbalance. And what's more, the potential of ensemble algorithms in this space has been vastly unexplored. In this paper we aim to bridge this gap by conducting a comparative analysis of classic machine learning, ensemble learning and deep learning algorithms, all while tackling the class imbalance problem head-on.

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

Machine Learning, Deep Learning.

# I. PROJECT INTRODUCTION

In times of natural disasters, reaching out to affected populations promptly is a matter of life and death. Rescue and relief agencies face a daunting challenge in accurately locating victims amidst a high volume of rescue calls, while also having to prioritize rescue activities according to the victims' urgent needs. Amidst all this chaos, the use of Twitter and Facebook generates an enormous amount of data, making it nearly impossible for aid agencies to manually sort through each tweet and determine which rescue and relief efforts should be prioritized. It is therefore of utmost importance to design a system that can accurately categorize these tweets into diverse humanitarian aid categories. Yet, given that tweets are limited to 280 characters and frequently contain uncommon acronyms and spelling errors, pre-processing and automatically categorizing them presents a significant challenge

This paper takes the idea of categorizing disaster-related tweets to the next level by proposing an innovative approach that has not been widely explored in previous studies. The proposed model goes beyond a simple comparative analysis of classic machine learning and deep learning algorithms, and also explores the potential of ensemble learning algorithms such as XGBoost, Light GBM, and CatBoost. One of the unique aspects of this work is its focus on addressing the issue of class/data imbalance, which is a crucial factor in accurately categorizing disaster-related tweets. Table1 highlights the skewed distribution of data across different categories, making it clear that traditional and deep learning models may not be effective in dealing with such challenges. By using ensemble learning techniques and carefully considering class imbalance, this paper offers an impactful solution to the challenges of categorizing disaster-related tweets.

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 |

# II. LITERATURE REVIEW

Categorizing disaster-related tweets has become a crucial task in recent times, and a multitude of studies have proposed different approaches to tackle this challenge. In recent years, deep neural network-based models have gained significant traction in this field, outpacing traditional machine learning techniques. The application of cutting-edge deep learning models such as Transformer and BERT have shown tremendous promise in accurately categorizing disaster-related tweets. However, classic machine learning algorithms such as Support Vector Machines, Random Forest, Decision Trees, Naïve Bayes, Logistic Regression, and Gradient Boosting still remain widely popular in this field. These algorithms continue to play a critical role in disaster-related tweet classification, and their effectiveness should not be overlooked.

The authors T. Ramya and J. Anita Christaline had done a comparative study on identification of informative tweets from crisis related data on Twitter [1]. The study focuses on crisis-related datasets, classification of tweets, pre-processing methods, and machine learning algorithms used in the study. The authors have evaluated the performance of the classifiers using various metrics such as AUC, precision, recall, and F1-score. The results indicate that SVM outperformed compared with other classifiers. The study also identifies several limitations such as the quality of the dataset, the accuracy of the pre-processing methods, and the generalizability of the models to other datasets. Overall, the paper provides a comprehensive analysis of the existing literature on the topic and provides useful insights into the challenges and opportunities for future research in this area.

The authors N. Assery, Y. Xiaohong, S. Almalki, R. Kaushik and Q. Xiuli introduced an analytical framework for identifying disaster related tweets [2]. They conduct experiments using two different datasets related to hurricanes and compare the performance of Naive Bayes, Random Forest, and Support Vector Machines (SVM), Decision Trees, KNN and Logistic Regression classifiers. The study also explores the effectiveness of different text vectorization techniques such as term frequency-inverse document frequency (TF-IDF) and count vectorizer. The results show that almost all algorithms except KNN outperforms, and that count vectorizer performs better than TF-IDF vectorization. However, the study is limited to only two datasets and the comparison is not done with other recent state-of-the-art ensemble models. Moreover, the study does not explore the impact of hyperparameter tuning on the performance of the classifiers.

In 2022 K. Asinthara, M. Jayan and L. Jacob conducted a study on analysing the sentiments of disaster related tweets using Machine Learning and Deep Learning techniques [3]. Supervised Learning techniques such as SVM, Naïve Bayes and LSTM were used to develop the classifier. The authors find that their approach outperforms other models in terms of classification accuracy and F1 score. However, the authors note that their model may not be effective in situations where the dataset is imbalanced or the tweets are poorly written.

The authors Dalela P. K, Senjyu, T., Mahalle, P.N., Perumal, T., Joshi presented a supervised learning-based approach for classifying disaster-related tweets with a case study on Cyclonic Storm ‘FANI’ [4]. They have used a combination of data pre-processing techniques and supervised learning algorithms like Linear SVC, Logistic Regression, Multinomial Naïve Bayes, Random Forest, XGBoost to classify the tweets into 9 different categories. The results show that Linear SVC and Logistic Regression outperform in terms of accuracy. The limitation of this study is that classes are skewed and accuracy score alone cannot be the only error metric to determine the effectiveness of the classifier. Furthermore the study only considers a limited set of machine learning algorithms and does not explore more complex models such as deep neural networks and ensemble learning.

Christidou et al. (2022) proposed a machine learning-based method for automatically recognizing essential information contained in Twitter that is important to emergency responders during a natural disaster and in the aftermath of a natural disaster [5]. They then applied different machine learning algorithms such as Naïve Bayes, Logistic Regression, SVM, Classification and Regression trees (CART), K-Nearest Neighbor (KNN) and Adaboost to classify the tweets as informative or non-informative. Best results were achieved with the Logistic Regression and SVM algorithms. The model falls short in addressing the critical challenge of class imbalance and lacks exploration of advanced techniques such as Boosting and Deep Neural Networks, which could potentially unlock greater insights and improve the overall performance of the system. The strides made in recent disaster-related tweet classification research are impressive, with models achieving high levels of accuracy in classifying tweets related to various disasters. Nonetheless, challenges remain, including skewed class distribution, difficulty classifying certain disaster types, and limited generalization to novel disaster types.

The paper's notable contribution and distinguishing factor can be succinctly stated as:

1. Four disaster-related datasets - hurricane, wildfire, flood, earthquake will be used to train and test traditional machine learning, ensemble learning, and deep learning models.
2. Custom training GPT2 medium language model and leveraging the same for generating synthetic samples thereby mitigating the problem of imbalanced data distribution.
3. Examining the significance of different text vectorization techniques including 2 different word vector models

III. PROPOSED MODEL

This study utilizes a dataset published in [6]. The data is imbalanced as it can be clearly seen in Table I, hence we need to balance out the data by generating synthetic samples. To achieve this, we will be custom training GPT2 medium language model on different datasets listed in Table I and thereafter generate the synthetic samples using the custom trained GPT2 medium model. GPT-2 is essentially a sophisticated Language Model at heart that is based on Transformer Architecture and is trained on 40GBs of Web Text. But before custom training the GPT2 model we need to pre-process the data and remove urls, html tags, emojis, emoticons, white spaces, expanding contractions and converting chat abbreviations thereby removing noise from the data. Once GPT2 model has been custom trained, we will generate synthetic samples and balance out the data. Now the resulting balanced data will be processed further wherein words will be lemmatized and punctuations and stop words will be removed from the text. Post processing the data will be split into train and test set in the ratio of 80:20.

For transforming the data all possible n-gram term frequency-inverse document frequency features ranging from 1-3 will be used for training classical ML algorithms and ensemble algorithms. In case of Deep Neural Networks 2 different word vectors Crisis[7] and GloVe[8] will be used. The utilization of pre-trained word vectors has proven to be advantageous in text classification models. By utilizing pre-trained word vectors, we leverage transfer learning.

Post data transformation, 7 different classical and ensemble ML algorithms and 7 different Deep NN algorithms will be used for performing the comparative study. In case of classical ML algorithms Naïve Bayes, Decision Tree, Random Forest and SVM will be trained whereas in ensemble ML algorithm Gradient Boosting, XGBoost and CatBoost will be implemented. With regards to Deep NN following models will be trained upon LSTM, GRU, Bi-LSTM, Bi-GRU, Bi-LSTM with Attention, Bi-GRU with Attention and BERT.

Once the model has been trained the performance of the model will be evaluated using Precision, Recall and F1 score. The workflow of the proposed model can be seen in Figure 1.

# IV. SYSTEM DEFINITION

The study utilizes traditional, ensemble, and deep learning methods to categorize tweets related to disasters into six distinct categories. For traditional machine learning different flavours of the term frequency-inverse document frequency feature vectorization technique will be utilized. In contrast, two distinct word vector model will be deployed for Sequence-to-Sequence Classification. To address the issue of imbalanced class distribution, GPT2 medium model will be custom trained on each of the different datasets. The resulting trained models will then generate synthetic samples, providing a more balanced representation of the data. Moreover, for estimating model’s performance we will be using different evaluation metrics like – precision, recall and F1 score. By evaluating the precision and recall of each model, we can determine how well each model is able to correctly identify positive instances while minimizing false positives. F1 score will provide a balanced measure of precision and recall, taking into account both metrics to provide an overall measure of the model's effectiveness. Thus precision, recall, and F1 score together will provide quantitative and objective means of evaluating and comparing the performance of different ML models and will also helps us in making informed decisions about which model to select for a particular task.

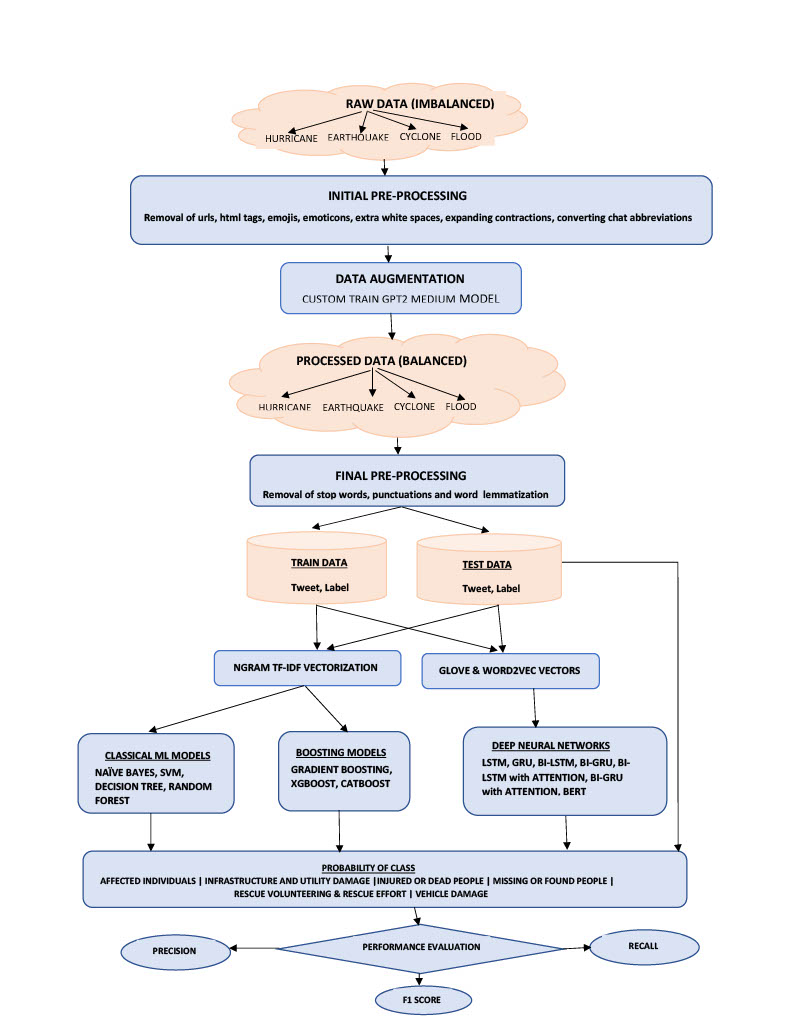


FIGURE1:MODEL FLOWCHART

# V. REFERENCES

[1] Ramya, T., Christaline, J.A. (2022). A Comparative Study on the Identification of Informative Tweets Using Deep Neural Networks During Crisis. In: Mallick, P.K., Bhoi, A.K., González-Briones, A., Pattnaik, P.K. (eds) Electronic Systems and Intelligent Computing. Lecture Notes in Electrical Engineering, vol 860. Springer, Singapore

[2] N. Assery, Y. Xiaohong, S. Almalki, R. Kaushik and Q. Xiuli, "Comparing Learning-Based Methods for Identifying Disaster-Related Tweets," 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 2019, pp. 1829-1836, doi: 10.1109/ICMLA.2019.00295.

[3] K. Asinthara, M. Jayan and L. Jacob, "Classification of Disaster Tweets using Machine Learning and Deep Learning Techniques," 2022 International Conference on Trends in Quantum Computing and Emerging Business Technologies (TQCEBT), Pune, India, 2022, pp. 1-5, doi: 10.1109/TQCEBT54229.2022.10041629.

[4] Dalela, P.K. *et al.* (2021). Classification of Disaster-Related Tweets Using Supervised Learning: A Case Study on Cyclonic Storm FANI. In: Senjyu, T., Mahalle, P.N., Perumal, T., Joshi, A. (eds) Information and Communication Technology for Intelligent Systems. ICTIS 2020. Smart Innovation, Systems and Technologies, vol 195. Springer, Singapore.

[5] Christidou, A.N., Drakaki, M., Linardos, V. (2022). A Machine Learning Based Method for Automatic Identification of Disaster Related Information Using Twitter Data. In: Kahraman, C., Tolga, A.C., Cevik Onar, S., Cebi, S., Oztaysi, B., Sari, I.U. (eds) Intelligent and Fuzzy Systems. INFUS 2022. Lecture Notes in Networks and Systems, vol 505. Springer, Cham.

[6] 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.

[7] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

[8] Muhammad Imran, Prasenjit Mitra, Carlos Castillo: Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In Proceedings of the 10th Language Resources and Evaluation Conference (LREC), pp. 1638-1643. May 2016, Portorož, Slovenia.