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 increasingly prone to devastating disasters, every second is crucial. Twitter serves as a rich source of real-time information during crises, providing invaluable insights into casualties, infrastructure damage, and urgent rescue efforts. It is vital to develop a model capable of classifying disaster-related tweets into distinct categories, thereby facilitating rescue operations and saving lives. Our study aims to bridge this gap by conducting an in-depth comparative analysis of classical machine learning, ensemble and deep learning algorithms, all while confronting the class imbalance issue head-on. We assess these model’s effectiveness using four distinct disaster events: hurricanes, earthquakes, floods, and wildfires. Our findings reveal that deep neural network models achieve F1-scores ranging from 0.41 to 0.70, outperforming classical machine learning classifiers with scores between 0.62 and 0.66 and ensemble learning algorithms with scores from 0.54 to 0.64. This underscores th remarkable efficacy of deep neural network models in classifying disaster-related tweets.

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

Machine Learning, Deep Learning.

# I. INTRODUCTION AND MOTIVATION

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 [1]. 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. RELATED WORK

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 [2]. 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 [3]. 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

The authors N. Assery, Y. Xiaohong, S. Almalki, R. Kaushik and Q. Xiuli introduced an analytical framework for identifying disaster related tweets [4]. 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 [5]. 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.

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’ [6]. 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. 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 [7]. 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 [8]. The data is imbalanced as it can be clearly seen in Table I, hence we need to balance out the data by

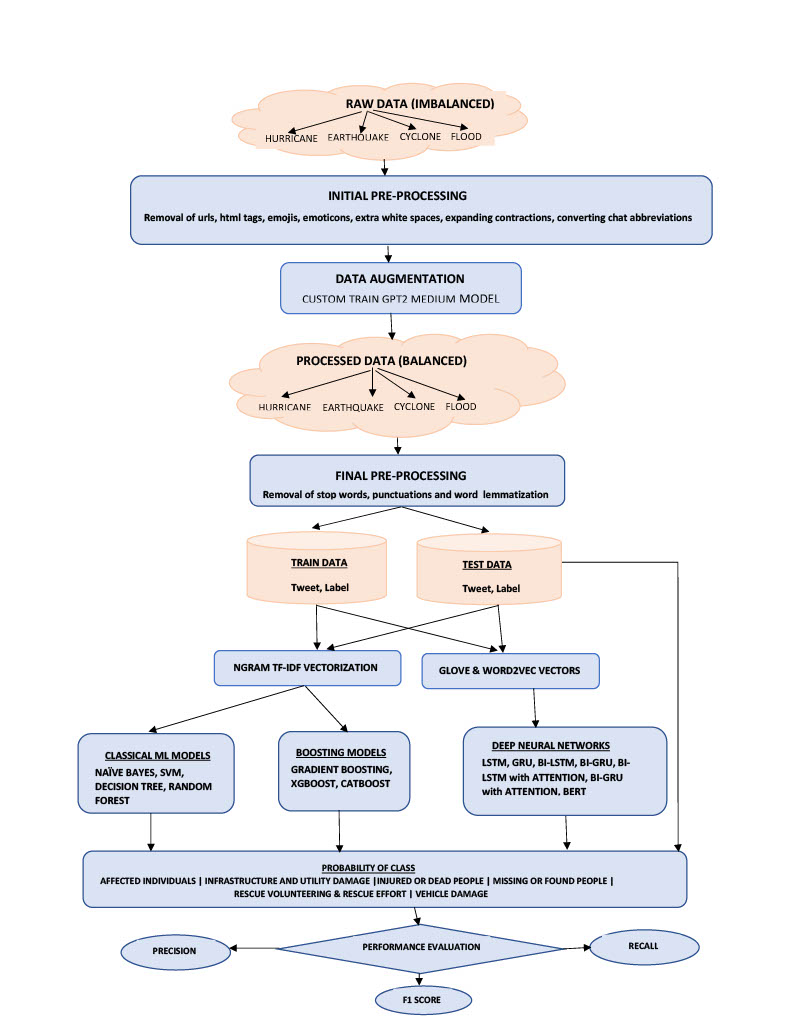


FIGURE1:MODEL FLOWCHART

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[9] and GloVe[10] 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. RESULTS

Leveraging the power of Sklearn and Keras Python libraries, we've implemented a comprehensive range of classical machine learning, ensemble learning, and deep neural network models. To address the challenge of class imbalance, we custom-trained a GPT-2 medium model on a crisis disaster dataset. Our performance evaluation focused on precision, recall, and F1-score metrics.

We conducted extensive experiments using 1-gram, 2-gram, and 3-gram TF-IDF features on disaster-related datasets, including Hurricanes, Earthquakes, Floods, and Wildfires. Figure 2 and 3 present the results for each classifier across balanced and imbalanced datasets, showcasing various N-gram TF-IDF feature combinations.

In the realm of classical ML algorithms, the decision tree excelled in handling imbalanced datasets, while logistic regression and naive bayes shined with balanced datasets. Ensemble learning's Gradient Boosting (GB) classifier and deep NN's Bi-LSTM algorithm emerged as top performers across all disaster-related events for both balanced and imbalanced datasets. Notably, the Bi-LSTM classifier achieved F1-scores of 0.70, 0.41, 0.66, and 0.69 for Earthquake, Flood, Hurricane, and Wildfire events, respectively.

Our experiments also explored various text vectorization techniques, revealing that unigrams

Figure 2: Performance Evaluation on Imbalanced Data

Figure 3: Performance Evaluation on Balanced Data



FIGURE4 : Bar Plot of Average F1-Score across different algorithms

(1-grams) generally outperformed other N-gram combinations, followed by bigrams (2-grams). In the case of deep neural networks, the attention mechanism proved to be the most effective.

As illustrated in the bar plots in Figure 4, deep neural network-based models consistently outperformed classical and ensemble machine learning techniques across all disaster events. These results demonstrate the superior learning capabilities of deep neural networks compared to their classical and ensemble counterparts. Furthermore, the bar plots in Figure 5 highlight that addressing class imbalance significantly improves the F1-score for classical and ensemble ML algorithms. However, this improvement is less pronounced in Deep Neural Networks, which require larger networks and vast amounts of data to reach their full potential.

V. LIMITATIONS

The constraints of this study include the exclusive use of English language tweets for the classifying tweets, despite the fact that users often post tweets in their local languages during disasters. As a result, a multi-lingual system could be developed for future research. Another limitation is the use of default hyperparameters for batch size, learning rate, and optimizer for all deep neural network-based models in our experiments. In subsequent studies, these hyper-parameters could be further optimized to enhance performance.

Additionally, to address the class imbalance issue, we employed a custom-trained GPT-2 medium model, although using a GPT-2 large language model for custom training is another option, albeit with increased computational power and memory requirements. In future research, when vectorizing tweets for classical ML models, employing a hash vectorizer may help reduce the length of vectorized tweets. Moreover, considering contextual BERT embeddings as an alternative to word embeddings could potentially yield better results.

VI. CONCLUSION AND FUTURE WORK

In this study, we classify disaster-related tweets into six distinct categories to aid humanitarian efforts. Our comprehensive comparison covers a broad range of techniques, from classical and ensemble machine learning to deep learning approaches. We employ various N-gram TF-IDF feature vector combinations for classical and ensemble ML algorithms, while utilizing word embedding vectors for deep learning. Our findings demonstrate the prowess of decision trees in managing imbalanced datasets and the effectiveness of logistic regression and naive Bayes with balanced datasets in the classical ML domain. The Gradient Boosting (GB) classifier from ensemble learning and the Bi-LSTM algorithm from deep learning emerged as top-performing techniques across all disaster-related events, regardless of dataset balance. Further experimentation with text vectorization techniques revealed the superior performance of unigrams (1-grams) in comparison to other N-gram combinations, followed by bigrams (2-grams). In deep neural networks, the attention mechanism proved most effective. This highlights the importance of 1-gram features in classical and ensemble ML classifiers for tweet classification. We encourage future researchers to select appropriate text vectorization techniques and employ hyperparameter tuning to optimize performance.

VII. REFERENCES

[1] Wahid, Junaid Abdul, et al. "Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response." Expert Systems with Applications 195 (2022): 116562.

[2] 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

[3] 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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.