Transformers with Disaster tweets

***Abstract*— Nowadays, a huge amount of data is available on social media platforms. The data contains valuable information which can be used for various applications and studies. This study exploits twitter data to classify them as disaster tweets on non-disaster tweets. The tweets are classified using transformer models i.e. DistilBERT, Roberta, XLNet, and ELECTRA. The performance of these models is compared using the metrics; accuracy, area under the curve receiver operating characteristic, and f1-score. Thus, on the testing set, RoBERTa achieved the highest f1-score of 0.808. Lastly, the study can be extended by experimenting with various preprocessing techniques and hyperparameter optimization.**

Keywords—Natural Language Processing,Twitter, Disaster twitter, review

I. Introduction

Nowadays. Twitter is one of the most popular online social networks worldwide. It has around 650 million registered users, moreover, it is the third most popular worldwide network [1]. Therefore, millions of tweets are posted each day which can be used to analyze online behavior through this data. Many applications use the data for various purposes; sentiment analysis, fake news detection, cyberbullying, etc [2][3][4]. Similarly, data from Twitter can also be utilized for recognizing a disaster, which can lead to better disaster management by the organizations and timely decisions for the government as well [5]. However, it is not always unambiguous whether the tweet is related to a disaster or not. For example, in the tweet: "Aftershock was the most terrifying best roller coaster I've ever been on. *DISCLAIMER* I've been on", the word aftershock is used as the name of a roller coaster, but it can also be used in the context of an earthquake. The purpose of this work is to classify whether the tweet is related to a disaster or not. There have been several studies to classify the disaster tweets like in [6], various machine learning algorithms are applied and compared. Similarly, some research has been done on only one particular disaster and area such as a Hurricane in Texas, or Earthquake in Nepal [7],[8]. In addition, classification on language-specific tweets with specific disaster types [9]. Most of the proposed techniques for this task are based on machine learning techniques or deep learning techniques. In this classification problem, the context of the text holds immense significance as described in the preceding example. Nevertheless, the machine learning models and deep learning models do not have long-term dependencies, because of which they cannot contribute to extracting accurate contextual information from the text. To mitigate these issues, transformer models are used as they cater to the long-term dependencies with attention [16]. Therefore, in this study which extends the work of [14], four transformer models are applied for the classification of tweets. The models applied in this study are; Distillbert, RoBERTa, XLNet, and Electra. Furthermore, the models are then evaluated using accuracy, area under the curve receiver operating characteristic, f1score. The highest f1-score 0.808 based on the testing data is achieved by RoBERTa.

The rest of the paper is as follows. Section 2 describes the literature review for the problem. The proposed method is explained in Section 3. Section 4 gives experimental results and analysis. The paper is concluded in Section 5

II. Literature Review

The data on social media has valuable information. The data extracted from tweets can be used in various applications. The authors Samane et al have proposed a methodology to classify the tweets related to disaster using natural language processing techniques [6]. The dataset is taken from kaggle which has 7613 instances among which 4342 are not disaster related tweets and 3271 are disaster-type tweets. First, the text preprocessing is applied using the techniques; tokenization, normalization, stemming, lemmatization, stop word removal, and noise removal. Second, for feature extraction TF-IDF is used followed by classification algorithms i.e. SVM, Multinomial Naive Bayes, Logistic Regression, XG classifier, Random Forest, Decision Tree and K-Nearest Neighbor. The performance of these models are then analyzed using accuracy and F-Score. Moreover, SVM performance was the best in this study with 81% of accuracy and 80 % F-score. However, KNN performed the worst with an accuracy of 65 % and F-score of 33%. Lastly, more word tokenization methods should be included, as tf-idf is also known for sparsity in dataset.

In the paper[7], Sreenivasulu et al have proposed a method using deep neural networks to classify the informative and uninformative tweets related to disaster. The dataset used in the study, collected the tweets from August 26,2017 to September 20,2017. The dataset is specific to Hurricane Harvey which hit Texas, USA. The authors have proposed the combination of Convolutional Neural Network and Artificial Neural Network for classification. Furthermore, the proposed model is compared with CNN, ANN and SVM on the basis of parameters; precision, f1-score, accuracy, and recall. The proposed architecture of CNN includes five layers in total; an input layer, a convolution layer, pooling layer, dense layer and output layer. Moreover, the words are vectorized using Crisis word embeddings. The extracted features from CNN are inputs of ANN which are used for classification. Hence, the proposed model performed better with precision, recall, f1-score and accuracy of 0.76, 0.76, 0.76 and 0.75, respectively. Lastly, the model is limited to only one type of disaster, hence, the experimentation over other datasets would give more validation to the proposed model. Also, the model can be tweaked such as adding more layers for further experiments on the problem.

The paper [8], demonstrates the classification of disaster type data by proposing a method which includes ELMo embeddings followed by a dense classifier. The model is applied on three Twitter datasets i.e. California Earthquake with 2012 instances, Nepal Earthquake with 12140 instances, and Typhoon hagupit with 11487 instances. Firstly, the tweets are sent to the embedded layer to extract vectors of the words. Secondly, the models are trained to classify the labels. The models used for comparative study include; Support vector machine with BoW embeddings, Convolutional Neural Network, Multilayer Perceptron-Convolutional Neural Network with Crisis word embedding and the proposed model which used ELMo with dense classifier. In addition, the class labels are categorised into Donations and Volunteering, Infrastructure and utilities, Sympathy and Support, Affected Individual, Other Useful Information and irrelevant information to the disaster. Thirdly, the models are evaluated with the parameters; precision, recall, f1-score and accuracy. However, the proposed model achieved the highest accuracies i.e. 77.57 % for Nepal Earthquake, 78.79 % for California Earthquake and 90.09 % for Typhoon Hagupit. Nevertheless, the model is trained on specific datasets, with more disasters-type, a better validation of the model can be demonstrated.

Furthermore, the authors Alaa et al , have presented their findings on disaster from Arabic tweets [9]. They have focused on the high-risk floods and proceeded with their study using machine learning classifiers and deep neural network classifiers. The dataset was collected from twitter specific to four events back in 2018. The dataset consisted of 4,037 tweets among which 24 percent were irrelevant tweets. After the text preprocessing steps, the models were implemented for experimentation. The machine learning models applied in this paper were, Naive Bayes, and Support Vector Machine. In addition, the deep learning model included; Convolutional Neural Networks, Long Short-Term Memory, Convolutional LSTM and Bidirectional LSTM. The experiment demonstrates that RNN outperformed the machine learning classifiers. The models were evaluated on the basis of accuracy metric. The highest test accuracy was
77.95 and 77.63 of the models; LSTM, Bi LSTM respectively. However, the models were not evaluated on other important performance metrics such as precision, recall, f1-score. Lastly, the model is trained for specific disaster types, hence, more data can be used for further work in this study.

The authors Abhinav et al [10], have done a comparative analysis of machine learning techniques and deep learning algorithms. The dataset has been taken from [10] in regards to four disaster events such as hurricane, flood, earthquake and wildfire. The dataset consisted of 5,559 tweets in total. After the data preprocessing step, the text is vectorized using various techniques. For the conventional machine learning models, the text vectorized using TF-IDF and for deep learning models, GLoVe and Crisis are used. Next, the machine learning and deep learning models are applied. In this paper, authors have applied seven machine learning algorithms such as SVM, K-nearest neighbor, Random Forest, Logistic regression, Gradient Boosting, Decision Tree and Naive Bayes. Further, five deep learning models are applied which includes; CNN, LSTM, Gated Recurrent Unit (GRU), Bi-GRU, and, GRU- CNN. The parameters used for evaluation of these models include precision, recall and f1score. The results conclude that GRU-CNN performed the best among deep learning models and machine learning models with precision of 0.82, recall of 0.80 and f1-score of 0.81. However, in machine learning models, Gradient Boosting performed with the highest precision, recall, f1-score of 0.66,0.67 and 0.67, respectively. Lastly, the deep learning models were performed with the same parameters such as, batch size, learning rate etc. The tuning of these hyper parameters can lead to better performance of these models.

The authors Jay et al [11], aspire to scrape the news from relevant websites and determine the news related to disaster by using machine learning and natural language processing techniques. The implementation of the system involves two modules. Module 1 involves the processes related to scraping of data and Module 2 deals with the steps leading to the classification disaster-type news. In Module 1, the selected websites included Times of India, NDTV India and Indian express, using the tool scrapy. The dataset used in this study consisted of 11k instances. Further, in module 2, the data is preprocessed in valuable form. Next, text vectorization is implemented for the data to be given as an input to the classifications models. The text vectorization used in the system includes; bag of words and tf-idf vector. Lastly, the classification models are used; Naive Bayes, Logistic regression, Support Vector Machine (SVM) Ensemble Methods including Bagging and Boosting models. The models were evaluated using the performance metrics; precision recall and F1 score. Hence, the logistic regression and Support Vector Machine performed better with the precision, recall and F1 score of 0.89. However, the authors have preferred Logistic Regression over SVM in this study because it showed less number of false positives than SVM. Lastly, the paper does not extend to contextual embeddings and using various deep learning methods, for a more elaborated study of the use-case.

In the paper [12], the authors have used this data to classify the real and fake disaster tweets. The dataset of 10,878 tweets is used in this study, which was shared by the company figure-eight. First, the data is preprocessed such as removing the duplicates, URL's, stop words, punctuation etc. Furthermore, stemming is performed, followed by tokenization. Second, the GloVe algorithm is used as a word embedding method which will be an input to the proposed model. Next, the authors have used Bidirectional LSTM for text classification. The model proposed by Aryan et al consists of three Bidirectional lstm layers and two dense layers with a dropout of 0.2. The activation function "Relu" is used, with sigmoid function to classify the output. In addition, the performance metrics accuracy is used to evaluate the model. The model obtained a training accuracy of 85.6 percent. Nevertheless, the model should have been tested on various word embedding methods.

The author Ashish Kumar, has compared conventional machine learning algorithms with a deep learning model known as BERT [13]. In the paper, various methods are applied to classify the disaster-types tweets. The dataset is taken from Kaggle competition, which consists of 10,876 tweets. The authors have used three types of word embeddings; bag of words, context-free and contextual embeddings. For a bag of words, the machine learning models are used for prediction; Decision tree, Random Forest and Logistic Regression. Next, with deep learning models such as Bi-LSTM and softmax, context- free embeddings are used i.e. Skip-gram, FastText, GloVe etc. Furthermore, the authors used Bi-LSTM and SoftMax with contextual embedding using the BERT pre-trained model. The performance metrics applied for evaluation include: accuracy, F1-score and Auc. The results demonstrate that contextual embeddings are a better choice for this application. The model using BERT and Bi-LSTM achieved the train and test accuracy of 0.835 and 0.830, respectively. Lastly, other transfer learning models can also be used with Bi-LSTM for more analysis.

In the paper [14], the authors have presented the work which uses Natural Language Processing techniques to classify if the tweet was related to a disaster or not. The dataset has been taken from Kaggle competition which includes 10875 tweets. The vector tokenizer used in this paper is BERT tokenizer. In this study, the authors have used a transfer learning technique known as BERT. Furthermore, only one feed-forward layer is added and the tweets are classified. The model is evaluated using accuracy and f1-score. The model reached the accuracy of 83% and a F1-score of 81%. Hence, the model has performed well as compared to other conventional methods. However, the model has not been compared with other transformer models.

TABLE I. MODEL COMPARATIVE ANALYSIS

Paper	Techniques	Dataset	Performance metrics	Limitations
[6]	SVM	Kaggle	Accuracy, precision, recall, f1score	Only one method for word tokenization
[7]	CNN with ANN	CrisisMMD	Accuracy, precision, recall, f1score	Limited to only one disaster type
[8]	Elmo embedding with dense classifier	CrisisNLP , CrisisLex, AIDR	Accuracy, precision, recall, f1score	Limited to only one disaster type
[9]	LSTM BI-LSTM	Twitter API 3	Accuracy	Limited to specific disaster type. Evaluated only with accuracy
[10]	GRU-CNN	CrisisMMD	Recall, f1-score	No hyperparameter tuning
[11]	Text vectorization followed by classification	News	Precision, recall and f1-score	Does not extend to contextual embeddings and doe not utilise deep learning based models
[12]	Word embeddings along with bidirectional LSTMs	Twitter	Accuracy	Not tested on various word embeddings
[13]	Word embeddings	Kaggle	Accuracy, F1-score and Auc	Model not compared with other transfer learning models
[14]	BERT	Kaggle	Accuracy and F1score	Model not compared with other transfer learning models

III. METHODOLOGY

In this section, the methods leading to the classification of the tweets are discussed.

i. Dataset:

The dataset has been taken from Kaggle's ongoing competition "Real or Not? NLP with Disaster Tweets". The dataset consists of 10,000 tweets among which 7503 tweets are used for training and 3243 are used for testing that. Furthermore, out of 7503 tweets, 20% is used for the validation of the model.

ii. Data Preprocessing

The attributes are comprised of id, text, location, keyword, and target. The target values are annotated by hand, the value will be 1 if the tweet is about a disaster, and 0 otherwise. The columns keyword and location have 80% and 33% missing values, respectively.

Furthermore, the id column has no contribution to the prediction so it is removed. Also, the keyword and location attributes are dropped because of the missing ratio of the data. The text and target columns are free from missing values. Hence. the dataset comprises text and target columns where the target column is the class label and the text consists of tweets. The text is further preprocessed such as removal of links, hash, etc. For text preprocessing a library known as tweet-preprocessor is used, which supports cleaning of the data such as removing URLs, hashtags, emojis, etc. Next, for the tokenization of the data, transformer word embeddings are used, which are corresponding to the models used in the experimentation.

iii. Models

As concluded in the aforementioned related work, that the work done by the researchers in the respective domain is mostly based on machine learning algorithms and deep learning algorithms. However, in a couple of papers, Bidirectional Encoder Representation from Transformer (BERT) is used for word embeddings, and in the paper [14], it is used for the corresponding problem. Since transformer models mitigate the problem of long-term dependencies, and with help of the pre-trained models, with less data, good accuracy can be achieved [16]. Hence, in this study, extending the work of [14], other transformer models such as DistilBERT, RoBERTa, XLNet, and ELECTRA are used for the prediction of real disaster tweets.

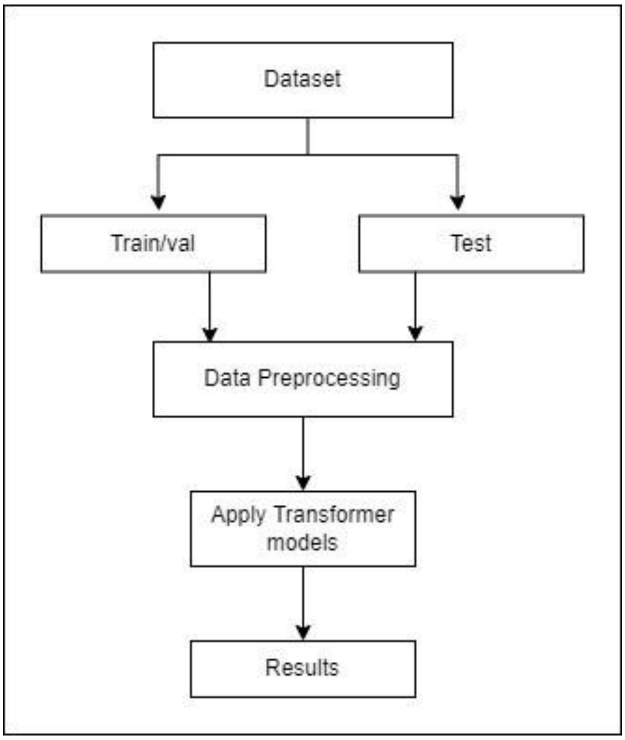
DistilBERT is a generalized pre-trained for of BERT, it is 40% smaller, 60% faster and it has the capability to understand the language as it retains 97% of it[17].

RoBERTa was proposed to cater the limitations of BERT.The authors of the paper [18], observed that BERT was significantly undertrained, hence, the improved version of BERT known as RoBERTa was proposed. The proposed model was trained longer, with more data, it removed the objective of the next sentence objective, it was trained on long sequences, and the changing of masking pattern was also applied.

XLNet is based on an autoregressive model (AR) which means that it takes the context words for prediction. Contrary to XLNet, BERT uses autoencoders for the language model and assumes the predicted masks to be independent of other tokens. Furthermore, masks are used in pretraining which consequently results in pre-train-finetune discrepancy at finetuning time. To cater to the mentioned issues, XLNet used the AR model and introduced Permutation Language Modelling so that the context can be learned forward and backward [19]. *ELECTRA* uses an alternative approach for pre-training the model, contrary to BERT, which uses masked language modeling. ELECTRA uses two transformer models, the generator that replaces the tokens in the sequence, and a discriminator, which identifies the corresponding tokens replaced by the generator. Nevertheless, the computation is 1/4 times less than RoBERTa.

The preceding models are then fined tuned over the natural disaster tweet dataset. Each model is trained over 10 epochs, using the Simple transformer library, which is based on Hugging face transformers. Next, the train batch size is 32 and the maximum sequence length is 128.

Figure I



IV. RESULTS

In this work, four transformer models are implemented for the classification of disaster tweets. Furthermore, these models were evaluated based on accuracy, area under the receiver operating characteristic (AUROC),f1-score, along with the observation of false-negative(FN), false-positive (FP), true-negative (TN), true-positive(TP).

Table II

Models	**Accuracy**	**AUROC**
Distillbert	0.82	0.87
ELECTRA	0.77	0.84
XLNet	0.81	0.88
RoBERTa	0.81	0.87

Table III

Models	**TP**	**TN**	**FP**	**FN**
Distillbert	521	734	135	133
ELECTRA	479	698	171	175
XLNet	529	707	162	125
RoBERTa	522	727	144	132

The preceding values of accuracy, AUROC, TP, TN, FP, and FN are based on the validation data. Furthermore, the results of test data was not disclosed by Kaggle. The f1-score was calculated by Kaggle after submitting the predicted outcomes. In table IV, it can be observed that RoBERTa achieved the highest F1 score.

Table IV

Models	**F1-score**
DistilBERT	0.805
ELECTRA	0.76
XLNet	0.804
RoBERTa	0.808

V. CONCLUSION

This study explores the power of state of the art transformer models. Furthermore, with minimum text preprocessing and hyper parameter tuning, the transformer models performed better than the conventional approaches taken towards the problem. Hence, the transformer models are a formidable choice to approach this problem. The highest f1-score of 0.808 was achieved by roberta with the accuracy of change. Moreover, in future work, these models can be experimented with more text preprocessing techniques, along with hyper parameter optimization.

REFERENCES

[1] A. Onan, “Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks,” *Concurr. Comput.*, vol. 33, no. 23, 2021.

[2] D. Antonakaki, P. Fragopoulou, and S. Ioannidis, “A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks,” *Expert Syst. Appl.*, vol. 164, no. 114006, p. 114006, 2021.

[3] T. Hamdi, H. Slimi, I. Bounhas, and Y. Slimani, “A hybrid approach for fake news detection in twitter based on user features and graph embedding,” in *Distributed Computing and Internet Technology*, Cham: Springer International Publishing, 2020, pp. 266–280.

[4] V. Balakrishnan, S. Khan, and H. R. Arabnia, “Improving cyberbullying detection using Twitter users’ psychological features and machine learning,” *Comput. Secur.*, vol. 90, no. 101710, p. 101710, 2020.

[5] S. G. Arapostathis, “Utilising Twitter for disaster management of fire events: steps towards efficient automation,” *Arab. J. Geosci.*, vol. 14, no. 8, 2021.

[6] Monfared, S. S., & Sedef, B, “Natural Language Processing for Prediction of Disaster Tweets using Machine Learning Methods”,2021.

[7] Madichetty, S., & Sridevi, M. (2019, January), “ Detecting informative tweets during disaster using deep neural networks”, In 2019 11th International Conference on Communication Systems & Networks (COMSNETS) (pp. 709-713). IEEE. 2019.

[8] Madichetty, S., & Sridevi, M. , “Improved classification of crisisrelated data on Twitter using contextual representations”, Procedia Computer Science, 167, 962-968, 2020.

[9] Alharbi, A., & Lee, M., “Crisis detection from Arabic tweets”,In Proceedings of the 3rd workshop on arabic corpus linguistics pp. 7279, 2019.

[10] Kumar, A., Singh, J. P., & Saumya, S. , “A comparative analysis of machine learning techniques for disaster-related tweet classification”, In 2019 IEEE R10 Humanitarian Technology Conference R10-HTC ,47129, pp. 222-227, IEEE, 2019.

[11] Domala, J., Dogra, M., Masrani, V., Fernandes, D., D'souza, K., Fernandes, D., & Carvalho, T., “Automated Identification of Disaster News for Crisis Management using Machine Learning and Natural Language Processing”, In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 503-508). IEEE, 2020.

[12] Karnati, A., Boyapally, S. R., & Supreethi, K. P. , “ Natural Language Processing with Disaster tweets using Bidirectional LSTM”, 2021.

[13] Chanda, A. K. ., “Efficacy of BERT embeddings on predicting disaster from Twitter data”, arXiv preprint arXiv:2108.10698, 2021

[14] BOUSBIB, R., & XAVIER, C., “Real or Not? NLP with disaster tweets”, 2021.

[15] F. Alam, F. Ofli, and M. Imran, “CrisisMMD: Multimodal Twitter Datasets from Natural Disasters,” arXiv [cs.SI], 2018.

[16] A. Vaswani et al., “Attention is all you need,” arXiv [cs.CL], 2017.

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,” *arXiv [cs.CL]*, 2019

[18] Liu *et al.*, “RoBERTa: A robustly optimized BERT pretraining approach,” *arXiv [cs.CL]*, 2019.

[19] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, “XLNet: Generalized Autoregressive Pretraining for Language Understanding,” *arXiv [cs.CL]*, 2019.

[20] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, “ELECTRA: Pre-training text encoders as discriminators rather than generators,” *arXiv [cs.CL]*, 2020.