2024 3rd International Conference for Innovation in Technology (INOCON)
Karnataka, India. Mar 1-3, 2024

Classification of Tweets using a Machine Learning and Natural Language Processing Algorithm for Disaster Prediction

Kanwarpartap Singh Gill [1]
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
kanwarpartap.gill@chitkara.edu.in

Vatsala Anand [2]
Chitkara University Institute of Engineering and Technology, Chitkara University,
Punjab, India
vatsala.anand@chitkara.edu.in

Deepak Upadhyay[3]
Computer Science & Engineering, Graphic Era Hill University,
Dehradun, Uttarakhand, India, 248002
deepaku.cse@gmail.com

Sarishma Dangi[4]
Computer Science & Engineering, Graphic Era Deemed to be University,
Dehradun, Uttarakhand, India, 248002
sarishmasingh@gmail.com

2024 3rd International Conference for Innovation in Technology (INOCON) | 979-8-3503-8193-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/INOCON60754.2024.10512145

***Abstract*— This research investigates the use of machine learning (ML) and natural language processing (NLP) algorithms for the categorization of tweets to anticipate disasters. This study aims to use the extensive and up-to-date social media data, namely from Twitter, to construct a reliable model for distinguishing tweets that pertain to disasters from those that do not. The technique being offered encompasses many key steps, including the gathering of data, pre-processing of the collected data, extraction of relevant features, and the subsequent deployment of several machine learning models. The primary objective is to develop a highly effective and precise system that can classify tweets in real-time, hence enhancing early warning systems and catastrophe management. The efficacy of the model will be assessed using evaluation criteria such as precision, recall, and accuracy. This will position the model as a helpful tool for boosting catastrophe prediction skills. The primary objective of this research is to forecast if a particular tweet pertains to an actual catastrophe or not. If this is the case, make a prediction of 1. If the condition is not met, the anticipated outcome would be a value of zero. The outcomes are also represented in the form of Learning Rate and Confusion Matrices in the proposed research.**

Keywords*— *Artificial Intelligence, Deep Learning, Tweet Classification Analysis, Model Training, Classification, *Learning Rate, Deep Learning*

I. INTRODUCTION

This theory presents an innovative methodology for improving catastrophe prediction via the use of machine learning (ML) and natural language processing (NLP) tools to classify tweets. In recent times, social media platforms have emerged as very significant resources for accessing up-to-date information during times of calamities. The objective of this theory is to use the potential of Twitter data by constructing a resilient classification model capable of discerning between tweets pertaining to disasters and those unrelated to disasters.

Social media platforms, namely Twitter, have become crucial mediums for the dissemination of information in times of crises. The vast amount of data produced renders human analysis unfeasible, hence requiring the use of automated methodologies. The primary objective of this idea is to create a tweet categorization system that demonstrates proficiency in accurately discerning and categorising tweets pertaining to calamitous events.

The first phase is the acquisition of a comprehensive and inclusive dataset of tweets that encompasses a wide range of both disaster and non-disaster situations, with the aim of ensuring diversity and representativeness. The provided dataset is used as a means to train the machine learning model, facilitating its acquisition of knowledge about patterns and correlations across various categories of tweets.

The data obtained from tweets in its raw form is necessarily characterised by noise and lacks a systematic format. Natural Language Processing (NLP) approaches are often used in the preparation of textual data. These techniques include many tasks such as tokenization, stemming, and the elimination of stop words. Moreover, the elimination of extraneous symbols, emoticons, and hyperlinks contributes to the improvement of the input data's quality.

Methods such as TF-IDF (Term Frequency-Inverse Document Frequency) or word embeddings (such as Word2Vec or GloVe) are used to transform tweets into numerical vectors that effectively represent the semantic significance of words.

The tweet classification task involves the evaluation of several machine learning methods, including Naive Bayes, Support Vector Machines (SVM), recurrent neural networks (RNNs), and transformers such as BERT. The selected model should demonstrate a high level of accuracy, precision, and recall when differentiating between tweets related to disasters and those that are not.

The model undergoes training using a dataset that has been labelled, and its performance is assessed via validation using a distinct collection of tweets that were not included in the training process. The process of iterative refinement is used in order to enhance the performance of the model, taking into account several criteria such as the risk of overfitting and the ability to generalise to novel data.

After the model has attained a level of performance that meets the desired criteria, it may be seamlessly included into a monitoring system that operates in real-time. The present system engages in ongoing analysis of incoming tweets, categorising them into two distinct groups: those pertaining to disasters and those unrelated to such events. The tweets that have been categorised may thereafter be used to provide prompt notifications and improve the effectiveness of early warning systems.

The efficacy of the suggested methodology is assessed using measures such as precision, recall, F1-score, and accuracy. The validation of the proposed system's superiority may be achieved by comparative assessments with current methodologies and models.

II. LITERATURE

Paul, N.R. and colleagues offer two innovative hybrid deep neural network models. The first model integrates Convolutional Neural Networks (CNN) with Gated Recurrent Units (GRU), whereas the subsequent model combines CNN with SkipCNN. In order to demonstrate the efficacy of our proposed models in recognising crisis-related information and identifying various forms of information essential for humanitarian relief, they assessed their performance on four distinct datasets given by CrisisNLP [1-2]. In this study, Gopi, A.P. et al. provide a proposed methodology for determining opinion scores and drawing conclusions. The process of categorising views is referred to as opinion mining, while the process of determining the sentiment scores associated with those opinions is known as sentiment analysis [3-4]. Vyas et al. used the Twitter dataset pertaining to the RUW, which was obtained from Kaggle.com. The RUemo framework has the capability to extract a total of 27 unique emotions expressed by Twitter users. These emotions are further categorised using machine learning methods. The researchers attained commendable levels of testing accuracy while using multilayer perceptron and logistic regression machine learning approaches for the job of multiclass emotion categorization [5-6]. Alduailaj, A.M. et al. conducted a study with the objective of using machine learning techniques in the Arabic language to facilitate the automated identification of cyberbullying [7-8]. Rahman et al. conducted a comparative analysis of single-layer architectural models and multi-tier models. The findings of the Multi-tier model show a modest improvement in comparison to the single-layer design [9-10]. A classification of tweet polarity was conducted by Ainapure, B.S. et al. using the VADER and NRCLex tools [11-12]. Alqarni, A. et al. conducted a study that focused on the sentiment analysis of Arabic tweets, specifically in relation to the COVID-19 outbreak in Saudi Arabia [13-14]. In their study, Kanan et al. elucidated the notion of Big Data analytics in the context of business on social media. They used three prominent Natural Language Processing techniques, namely stemming, normalisation, and stop word removal, to streamline the understanding of this idea [15-16]. Sadigov et al. examined the effects of the pandemic on the field of education and the subsequent emotional changes experienced by users. The study focused on the use of machine learning models to analyse the significant transformations that occurred inside the education system [17-18]. Ali et al. successfully identified intraclass groups and individuals with significant influence. The model that has been built demonstrates a high level of effectiveness in detecting hate tweets and identifying communities associated with such content. Moreover, this approach has promise for monitoring social media platforms in order to spot any indications of impending unrest or threats [19-20].

- An Introduction to Exploratory Data Analysis of Textual Data is done in this work.
- Text data cleaning is a crucial step in the preprocessing of data, particularly in the context of textual information.
- Proficiency in training FastAI ULMFIT on personalised textual data is Acquired.
- This research aims to provide guidance on the implementation of a language model using custom data.
- Proficiency in the use of Transfer Learning as a means to enhance accuracy is obtained.
- This guide aims to provide comprehensive instructions and strategies for beginners to effectively deploy a Text Classification model and achieve enhanced accuracy.

The study encompasses several subtopics. The next portion of the research paper presents a detailed description of the dataset used for analysis, followed by a later section on Data Validation. Section 5 provides an explanation of the AWD_LSTM language model, which is used for the classification of tweets. The findings are shown in Section 6 of the presentation. Moreover, Section 7, situated at the culmination of the presentation, encompasses a comprehensive compilation of references.

III. INPUT DATASET

The dataset consists of a compilation of tweets that have been labelled and annotated as either being linked to disasters or not related to disasters. The tweets are derived from a wide range of geographical regions and include a variety of catastrophe categories, such as earthquakes, floods, wildfires, and storms, among others. The dataset has been specifically curated to include the intricacy and fluctuation of language used in real-time situations occurring during crisis events on social media platforms, with a special focus on Twitter.

The dataset obtained from Kaggle has been carefully adjusted to guarantee an equitable distribution of both catastrophe and non-disaster tweets. The dataset comprises tweets that have pertinent phrases, hashtags, or mentions linked to different categories of disasters. Furthermore, tweets unrelated to disasters cover a broad spectrum of subjects in order to convey the multitude of common linguistic expressions seen on the site.

In order to improve the resilience of the model, the dataset is subjected to comprehensive preprocessing, which involves eliminating extraneous elements such as useless symbols, hyperlinks, and emojis. The process of tokenization and stemming is used to standardise the text, while feature extraction methods such as TF-IDF or word embeddings are performed to turn the textual data into numerical vectors that are compatible with machine learning algorithms.

The primary objective of this dataset is to provide a thorough and authentic portrayal of the linguistic patterns seen in tweets during both disaster and non-disaster situations. This dataset seeks to support the creation and refinement of a robust classification model that can accurately anticipate the occurrence of disasters as shown in Fig. 1.

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1

Fig. 1. Dataset Features Extracted into a CSV File

IV. Data Validation

The procedure of data validation plays a crucial role in guaranteeing the integrity and dependability of the dataset used for the purpose of training and assessing machine learning (ML) and natural language processing (NLP) algorithms designed for tweet categorization. The process of data validation is characterised by its iterative nature, whereby any detected flaws during the validation phase must be resolved prior to continue with the subsequent steps of model training and assessment. A well verified dataset serves as the fundamental basis for a dependable and efficient machine learning and natural language processing model designed for the categorization of tweets in the context of catastrophe prediction as shown in Fig. 2.

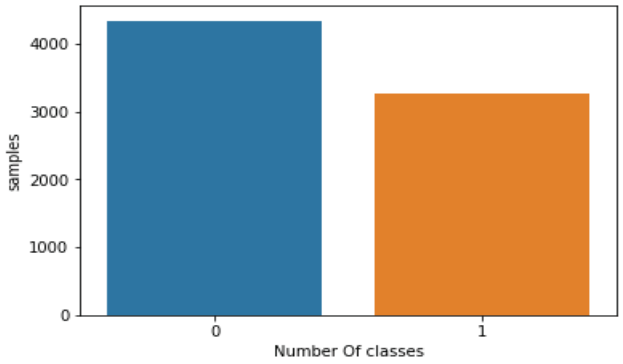

Fig. 2. Class Distributions of Tweets

V. Language Model AWD_LSTM For Classification of Tweets

The AWD_LSTM, also known as the ASGD Weight-Dropped LSTM, is an advanced recurrent neural network (RNN) structure specifically developed for the purpose of natural language processing (NLP) applications. In the realm of classifying tweets for the purpose of predicting disasters, the AWD_LSTM presents a robust and efficient method for collecting the contextual details and subtleties included within tweet content. In conclusion, the AWD_LSTM language model demonstrates a robust solution for the classification of tweets in the context of disaster prediction, owing to its sophisticated LSTM architecture and effective regularisation approaches. By using transfer learning and meticulous fine-tuning, the model can proficiently exploit the distinctive linguistic attributes of Twitter data, hence enhancing precision in real-time identification and categorization of tweets pertaining to disasters.

VI. Results

A. Hyperparameter Tuning for Model Training for Tweets Classification

The optimisation of machine learning models, particularly for tweet classification in catastrophe prediction, necessitates the critical process of hyperparameter tweaking. The procedure entails methodically fine-tuning the hyperparameters of the model in order to identify the optimal configuration that maximises both prediction accuracy and generalisation. Within the framework of tweet classification, the hyperparameters assume a crucial function in determining the behaviour of the model and exerting an influence on its capacity to differentiate between information pertaining to disasters and content unrelated to disasters. In summary, the process of hyperparameter tuning plays a crucial role in optimising the model for the purpose of classifying tweets in the context of catastrophe prediction. By conducting a methodical examination of the hyperparameter space and conducting thorough evaluations, the tuning procedure seeks to improve the predictive accuracy and generalisation capacities of the model, hence enhancing the usefulness of the catastrophe prediction system as shown in Fig. 3.

epoch	train_loss	valid_loss	accuracy	time
0	0.418936	0.457862	0.795007	00:03
1	0.421566	0.468393	0.792816	00:04
2	0.438573	0.476650	0.789750	00:04
3	0.453029	0.564869	0.779676	00:03
4	0.470692	0.537868	0.727113	00:03
5	0.493346	0.509943	0.777486	00:03
6	0.487005	0.484410	0.774420	00:03
7	0.481655	0.514257	0.769601	00:04
8	0.487110	0.491979	0.765659	00:03
9	0.486289	0.545819	0.734560	00:03
10	0.474185	0.471837	0.781428	00:04
11	0.464478	0.494066	0.787560	00:04
12	0.455770	0.465380	0.787122	00:03
13	0.451577	0.471351	0.784494	00:03
14	0.447491	0.458496	0.792816	00:03
15	0.439185	0.460489	0.786246	00:03
16	0.430277	0.452752	0.796759	00:03
17	0.427531	0.461279	0.792378	00:03
18	0.422353	0.456802	0.795007	00:03
19	0.419916	0.458118	0.795445	00:03

Fig. 3. Hyperparameter Tuning Depiction

B. Learning Rate Graph Analysis

The learning rate is a crucial hyperparameter that significantly influences the training process of machine learning models, especially those used in the categorization of tweets for the purpose of catastrophe prediction. The learning rate is a crucial factor in the optimisation algorithm's iterative process of updating model parameters to minimise the loss function, since it influences the magnitude of the step taken. In the domain of tweet classification, the careful selection of a suitable learning rate is of utmost importance in order to strike a harmonious equilibrium between model convergence and the prevention of convergence-related challenges. In summary, the careful selection of an optimal learning rate plays a pivotal role in the training process of models used for tweet categorization in the context of catastrophe prediction. By engaging in a methodical process of investigation and assessment, it is possible to adjust the learning rate in order to optimise the efficiency of model convergence, enhance its capacity to generalise to novel data, and improve the accuracy and dependability of the catastrophe prediction system as shown in Fig. 4.

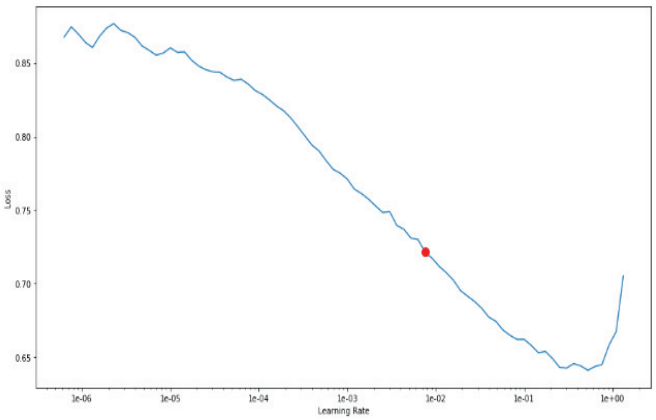

Fig. 4. Learning Rate Curve Analysis

C. *Confusion Matrix Illustration for the Classification of Tweets*

The use of a confusion matrix is a fundamental technique for assessing the effectiveness of machine learning models, specifically within the domain of tweet categorization for the purpose of predicting disasters. The analysis offers a thorough examination of the model's prognostications, emphasising the quantities of accurate positive outcomes, accurate negative outcomes, inaccurate positive outcomes, and inaccurate negative outcomes. Every individual feature inside the matrix provides vital insights into the model's capacity to effectively categorise tweets, hence playing a crucial role in evaluating the model's effectiveness in predicting disasters. In essence, the confusion matrix offers a comprehensive and practical evaluation of the performance of a tweet categorization machine. Through the process of analysing the matrix and extracting essential metrics, professionals have the ability to improve and optimise the model in order to achieve more precision and dependability in the forecast of disasters as shown in Fig. 5.

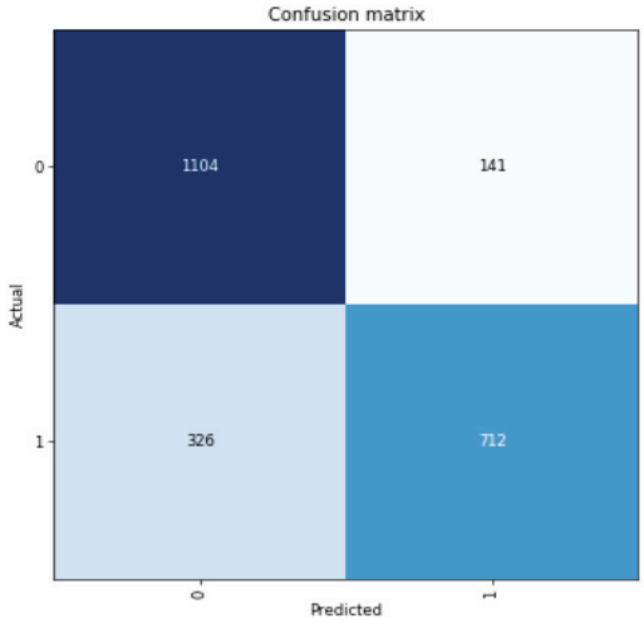

Fig. 5. Confusion Matrix Illustration for Tweet Classification

VII. Conclusion

In summary, this theoretical framework establishes the groundwork for a novel use of machine learning (ML) and natural language processing (NLP) in the realm of catastrophe forecasting by means of categorising tweets. The model under consideration exhibits the capacity to substantially enhance the efficiency and precision of discerning pertinent data under exigent circumstances, consequently augmenting the overall efficacy of catastrophe management and response endeavours. The main goal of this study is to create a system that is both extremely efficient and accurate in classifying tweets in real-time. This system aims to improve early warning systems and enhance disaster management. The effectiveness of the model will be evaluated by using assessment measures like as precision, recall, and accuracy. This will establish the model as a valuable tool for enhancing disaster prediction capabilities. The main aim of this study is to predict if a given tweet is related to a genuine disaster or not. Given the circumstances, please provide a projection of one. In the event that the condition is not satisfied, the expected result would be a numerical value of zero. The outcomes are also represented in the form of Learning Rate and Confusion Matrices in the proposed research.

References

[1] Paul, N.R., Sahoo, D. and Balabantaray, R.C., 2023. Classification of crisis-related data on Twitter using a deep learning-based framework. Multimedia Tools and Applications, 82(6), pp.8921-8941.

[2] Bharany, S., Badotra, S., Sharma, S., Rani, S., Alazab, M., Jhaveri, R.H. and Gadekallu, T.R., 2022. Energy efficient fault tolerance techniques in green cloud computing: A systematic survey and taxonomy. Sustainable Energy Technologies and Assessments, 53, p.102613.

[3] Gopi, A.P., Jyothi, R.N.S., Narayana, V.L. and Sandeep, K.S., 2023. Classification of tweets data based on polarity using improved RBF kernel of SVM. International Journal of Information Technology, 15(2), pp.965-980.

[4] Sharma, B. and Koundal, D., 2018. Cattle health monitoring system using wireless sensor network: a survey from innovation perspective. IET Wireless Sensor Systems, 8(4), pp.143-151.

[5] Vyas, P., Vyas, G. and Dhiman, G., 2023. Ruemo—the classification framework for russia-ukraine war-related societal emotions on twitter through machine learning. Algorithms, 16(2), p.69.

[6] Singh, P.K. and Sharma, A., 2022. An intelligent WSN-UAV-based IoT framework for precision agriculture application. Computers and Electrical Engineering, 100, p.107912.

[7] Alduailaj, A.M. and Belghith, A., 2023. Detecting arabic cyberbullying tweets using machine learning. Machine Learning and Knowledge Extraction, 5(1), pp.29-42.

[8] Reshan, M.S.A., Gill, K.S., Anand, V., Gupta, S., Alshahrani, H., Sulaiman, A. and Shaikh, A., 2023, May. Detection of Pneumonia from Chest X-ray Images Utilizing MobileNet Model. In Healthcare (Vol. 11, No. 11, p. 1561). MDPI.

[9] Rahman, H., Tariq, J., Masood, M.A., Subahi, A.F., Khalaf, O.I. and Alotaibi, Y., 2023. Multi-tier sentiment analysis of social media text using supervised machine learning. Comput. Mater. Contin, 74, pp.5527-5543.

[10] Gill, K.S., Anand, V. and Gupta, R., 2023, August. An Efficient VGG19 Framework for Malaria Detection in Blood Cell Images. In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-4). IEEE.

[11] Ainapure, B.S., Pise, R.N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M.S. and Bizon, N., 2023. Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. Sustainability, 15(3), p.2573.

[12] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2023, May. Smart Shoe Classification Using Artificial Intelligence on EfficientnetB3 Model. In 2023 International Conference on Advancement in

Computation & Computer Technologies (InCACCT) (pp. 254-258). IEEE.

[13] Alqarni, A. and Rahman, A., 2023. Arabic Tweets-based Sentiment Analysis to investigate the impact of COVID-19 in KSA: A deep learning approach. Big Data and Cognitive Computing, 7(1), p.16.

[14] Yang, M., Kumar, P., Bhola, J. and Shabaz, M., 2021. Development of image recognition software based on artificial intelligence algorithm for the efficient sorting of apple fruit. International Journal of System Assurance Engineering and Management, pp.1-9.

[15] Kanan, T., Mughaid, A., Al-Shalabi, R., Al-Ayyoub, M., Elbes, M. and Sadaqa, O., 2023. Business intelligence using deep learning techniques for social media contents. Cluster Computing, 26(2), pp.1285-1296.

[16] Sharma, S., Gupta, S., Gupta, D., Juneja, S., Gupta, P., Dhiman, G. and Kautish, S., 2022. Deep learning model for the automatic classification of white blood cells. Computational Intelligence and Neuroscience, 2022.

[17] Sadigov, R., Yıldırım, E., Kocaçınar, B., Patlar Akbulut, F. and Catal, C., 2023. Deep learning-based user experience evaluation in distance learning. Cluster Computing, pp.1-13.

[18] Sharma, R. and Kukreja, V., 2022, March. Amalgamated convolutional long term network (CLTN) model for Lemon Citrus Canker Disease Multi-classification. In 2022 International conference on decision aid sciences and applications (DASA) (pp. 326-329). IEEE.

[19] Ali, M., Hassan, M., Kifayat, K., Kim, J.Y., Hakak, S. and Khan, M.K., 2023. Social media content classification and community detection using deep learning and graph analytics. Technological Forecasting and Social Change, 188, p.122252.

[20] Sasubilli, S.M., Kumar, A. and Dutt, V., 2020, June. Machine learning implementation on medical domain to identify disease insights using TMS. In 2020 International Conference on Advances in Computing and Communication Engineering (ICACCE) (pp. 1-4). IEEE.