Covid-19 Tweet Sentiment Analysis

Mohammad Waseq-ul Islam , Md. Humaion Kabir Mehedi , Mohammed Julfikar Ali Mahbub and Annajiat Alim Rasel

Department of Computer Science and Engineering
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{mohammad.waseq.ul.islam, humaion.kabir.mehedi ,mohammed.julfikar.ali.mahbub }@g.bracu.ac.bd
annajiat@bracu.ac.bd

***Abstract*—With the current pandemic still at large and new variants emerging, the age of social media and Artificial Intelligence (AI) has made it accessible to monitor the spread of news. *Twitter* is a massive platform where people have been rampantly checking and spreading news of the current world affair since its beginning back in 2006. Negative and positive tweets may now be filtered for future study in various sectors of Natural Language Processing (NLP) thanks to new technology and research. One such technology and application of NLP includes sentiment analysis. In this paper, with the help of machine learning classification, we had to perform sentiment analysis and predict the emotion of a tweet (Positive, Neutral, or Negative). The performance metric of Stochastic Gradient Descent showed promising result accuracy, f1-score and recall of about 80%.**

***Index Terms*—AI, pandemic, social media, NLP, Twitter, sentiment analysis, classification, Machine Learning**

I. Introduction

As digital data and computer technologies have proliferated rapidly in the social sciences, the conflict between "qualitative" and "quantitative" methods has become more heated in a variety of ways [1]. It has recently been suggested that the seemingly contradictory research methodologies might be harmonized with the increased availability of internet data. The quali-quantitative approach, on the other hand, relies on the network graph. And one such phenomenon includes the spread of news on *Twitter*, which uses the traditional quali-quantitative method but remains elusive towards the network graphs. *Twitter* has been massive since its beginning back in 2006.

This is mostly due to the fact it gives the people the freedom to write, to communicate without using any email, platforms to promote any business [2] The sophistication has allowed the massive growth of the audience on *Twitter*. Nevertheless, with popularity comes the need for security and supervision, and technology has made it possible to analyze big data.

Because of the internet and social media, information has become more accessible and easy. However, this comes at a heavy cost; being held responsible for propagating hatred or false information or remarks on the internet. Profanities have proliferated on the internet in recent years, and given the sudden pandemic across the world, Covid-19, people have been relying mostly on these social platforms for their day to day to day news and updates. Sentiment analysis has been a use of NLP to evaluate and systematically identify effective expressions such as emotion, opinion, appraisals, or attitude towards anything. Positive or negative expressions can be analyzed through various means of Machine Learning [3]. In this paper, we have implemented sentiment analysis of *Twitter* for Covid-19 using different Machine Learning models.

Section II of this study briefly covers prior work in the domains of sentiment analysis. Section III describes the dataset that was used to train the classifiers. Section IV demonstrates the methodologies. The results and research are shown in Section V while future work discussion with a conclusion are in section VI

II. Related Works

To examine the issue of sentiment analysis, researchers have been experimenting with text classification techniques as well as related areas in order to produce an effective solution for identifying emotions in words and expressions. K. Wang et al [4] conduct sentiment analysis in the realm of peer review of scientific publications, which has never been done before. Using a reviewer's peer review text and an automated method, they were the first to attempt to anticipate an article's recommendation or conclusion. Aside from that, the job required detecting feelings, such as negative and positive, in phrases. The ICLR open reviews were used to create two assessment datasets, and the findings confirmed the efficacy of their suggested approach. In a variety of experimental conditions, the model significantly outperforms a few other models.

Yousif et al. [5] investigates sentiment analysis on scientific citation due to the abundance of scientific publications available currently. This paper describes the process of scientific citation sentiment analysis, as well as recently proposed methodologies that are presented, assessed, and criticised. They also discuss adjacent disciplines like citation function categorization and citation recommendation, both of which have recently received a lot of interest. The use of the classical machine learning models leave the scope of using deep learning models for future works which would be more reliable and work efficiently.

Recent research on the use of deep neural networks (DNN), convolutional neural networks (CNN) [6], and others to solve various sentiment analysis problems such as sentiment classifi-

cation, cross-lingual problems, textual and visual analysis, and product review analysis using deep neural networks (DNN), CNN, and others is highlighted and summarized by A. Yousif et al. [5] in their paper. The challenge was to use deep learning to overcome the shortage of labelled data in NLP for sentiment analysis.

To identify the mood of distinct tweets, B. Duncan (2015) [7] created a deep learning neural feedforward network model based on deep learning neural feedforward networks. 74.15 percent of the time, the accuracy rate was achieved.

According to earlier research, sentiment prediction is generally done with Nave Bayesian and neural networks to categorize texts with a small amount of data; however, to improve accuracy and handle large amounts of data, deep neural networks with deep learning architecture are used.

This paper's contribution is to use the inherently interpretable models for large data to monitor sentiment analysis (Positive, Negative, or Neutral) with respect to the pandemic, Covid-19.

III. Dataset

In order to get the training set of tweets with both positive and negative keywords, the *Twitter* API is utilized. In the testing sample of tweets, the positive and negative terms are the same. The dataset consists of five categories and the target label; username and screen name in numeric to avoid privacy concerns, location, dates of the tweet, the original tweet and the sentiment that is going to be classified. A total of 41157 instances in the training set, and 3798 instances in the test set. The set was sufficient for training the machine learning models and learning the differences in the sentiments of the tweets involved.

IV. Method

A. Preprocessing and Exploratory Data Analysis

The architecture of the suggested approach for sentiment analysis of tweets has been illustrated. Preprocessing phases include text cleaning and preprocessing the texts. The tweets are trained in the training step. The sentiment test, which is performed on the test dataset, is the last stage. And then the metrics help to better understand the models.

Text cleaning is one of the most fundamental phases in text classification for sentiment analysis because it removes tokens or other elements that are difficult to comprehend or evaluate the meaning of. There may be white spaces, punctuation, stop words, and other components in texts or phrases. These characters do not provide a lot of information and are tough to decipher for sentiment analysis. Words like "the," "is," and other English stop words, for example, convey nothing about the context, the elements listed in the texts, or the connections between them. [8]

There are several steps in preprocessing the texts for normalizing the sentences. These are:

- Deleting emojis and other non-characters for every of sentences.
- Removing retweets and symbols like '@'.
- Eliminate the hyperlinks from the statements, sentenes you see on Twitter.
- Create a list of tokenized sentences.

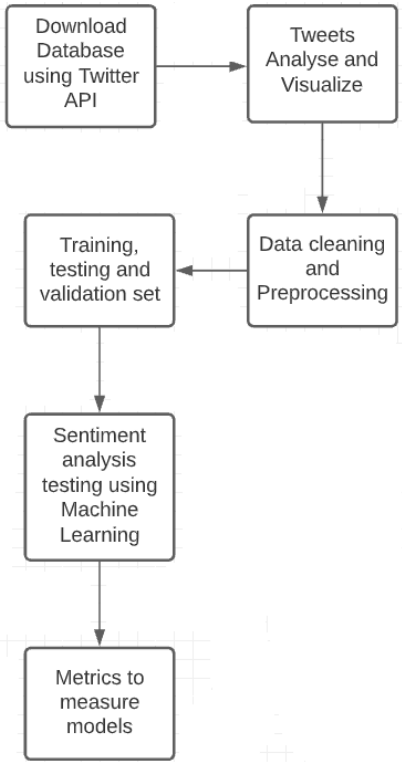

Fig. 1. Workflow of Sentiment Analysis

For a better understanding of the dataset, visualization such as Wordcloud was also performed using the *WordCloud* library [9]. In order to understand the significance of the categories, figures were drawn which includes barplot, donutchart. The percentage or ratio of tweets per location is visualized, and it seemed that London had the highest *Twitter* users in this dataset. Additionally, a barplot illustrates the counts of different number of tweet tokens it and the barplot after it was cleaned. Tweet tokens less than four were discarded because they do not give any information or important in the dataset.

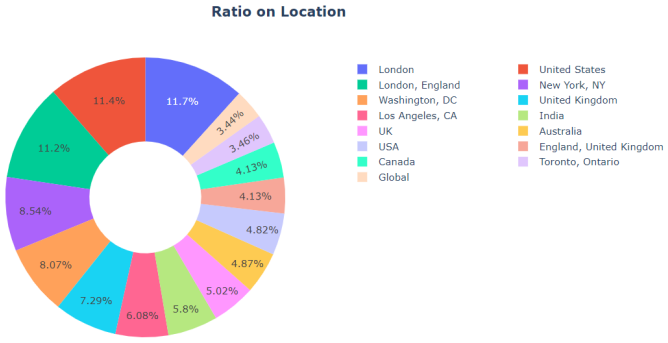

Fig. 2. Tweet user percentage across different locations

The preprocessing steps also include discarding data that has less significance than others, a deep cleaning procedure. In this deep cleaning, tweets with less than four words/tokens were discarded leaving 40935 instances. The remaining 222 instances were removed as a cleaning measure for efficient machine learning classification.

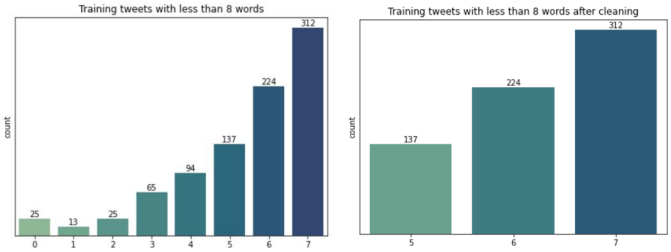

Fig. 3. Barplot showing the counts of different number of tweet tokens and the tweet tokens after it was cleaned

B. Predictive model development

There needs to be modifications to the data before feeding it to the learning models involved. However, there are numerous machine learning algorithms which are incapable of working directly with labelled data. They demand numerical input and output variables as the standard variables in the algorithm [10].

Generally, the transformation of categorical to numerical terms is done so that there is efficient implementation of the predictive machine learning algorithms.

The categorical value of the target label, sentiment, was examined in this paper and transformed to three independent classes using One-hot-encoding: Extremely positive and positive encoded "2," Neutral as "1," and Extremely negative and negative encoded as "0." This aids the machine learning algorithms in producing better outcomes.

1) XGBoost: Lately, XGboost has become one of the most used machine learning models by researchers for prediction. XGBoost is a toolkit for distributed gradient boosting that has been optimized for speed, versatility, and portability [11]. It is an open-source library that implements fast and highly optimized gradient boosted decision trees. XGBoost models require more knowledge and model tuning to get the best accuracy than techniques like Random Forest.

2) Gradient Boosting: It's important to use gradient boosting when constructing predictive models since it's really effective. The model is built using a collection of inferior prediction models, such as decision trees [12].It is based on the idea that the total prediction error is reduced when the best prospective future model is combined with past models. The primary premise of error reduction is to determine the intended results for the following model.

3) Logistic Regression: Rather predicting a continuous variable like size, logistic regression predicts whether something is true or untrue. Linear Regression, on the other hand, can handle both continuous and discrete data in the same way as logistic regression does. However, while logistic regression is most often employed with dichotomous dependent variables, it may also be conducted with outcome variables that have three or more classifications [13].

4) Stochastic Gradient Descent: Stochastic gradient descent (SGD) is a prominent and widely used technique in a variety of Machine Learning algorithms [14]. For each iteration SGD takes one data point at random from the whole data collection. This in consequence drastically reduces computation of the algorithm. A common practice in SGD is to use a small number of data for points for the sample at each step rather than just one. This is known as 'mini-batch' gradient descent. The purpose of this mini-batch is to create a balance between the benefits of gradient descent and the speed of SGD.

C. Model Evaluation Metrics

In the course of creating classifications predictions, there are four types of outcomes that might occur: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The classification metrics for model evaluation used to assess the performance of the models are as follows: accuracy, precision, recall-score, and f1-score. Accuracy, precision, and recall-score are the classification metrics for model evaluation used to evaluate the performance of the models. To begin, the model must be focused on the True Positive and True Negative aspects of the situation. The accuracy of the models is defined as the proportion of true predictions made by the models. According to the formal definition of accuracy, this is what it means:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

The precision score is the second factor to consider. Precision is defined as the percentage of correctly recognized positives out of all expected positives. The formula is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Third, recall score or sensitivity is the percentage of positives you properly recognized out of all positives. The following equation shows the calculation.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

There is another tradeoff in classification, besides the bias-variance tradeoff, that is sometimes overlooked. This is the tradeoff between precision and recall [15].

The F1-Score was then computed. It's known as the harmonic mean of the model's precision and recall. The computation yielded the following result:

$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

V. RESULTS

We conducted tests to assess the prediction models (advanced machine learning techniques) and find the most essential component for sentiment analysis on *Twitter* for the specified dataset on Covid-19 using various data visualization methods.

Table I
PERFORMANCE METRICS

Models	Accuracy	Precision	Recall	F1-score
XGBoost	0.753	0.761	0.753	0.754
GradientBoosting	0.767	0.773	0.767	0.768
Logistic Regression	0.789	0.789	0.789	0.789
SGD	0.80	0.801	0.799	0.80

Table I summarizes the results of the evaluations of the accuracy, precision, recall, and F1-score of the machine learning and model performance. Precision is a measure of quality, whereas recall is a measure of quantity. More relevant outcomes are produced by algorithms with higher accuracy, whereas algorithms with higher recall provide the majority of relevant outcomes (whether or not irrelevant ones are also returned) [15]. Due to a high proportion of false negatives and false positives, accuracy and recall ratings are mostly inaccurate [16].

The analysis of the machine learning and models performance that have been evaluated on accuracy, precision, recall and F1-score are summarized in Table I. Precision is used for indicating the quality of the performance of the algorithm, whereas recall is the quantity. An algorithm that yields high precision gives more relevant results than unrelated ones, but an algorithm with high recall predominantly yields the of relevant outcomes (whether or not those that are not needed are also returned)[15]. Nevertheless, precision and recall scores are mostly unreliable due to the high rate of false negatives and positives high at times [16].

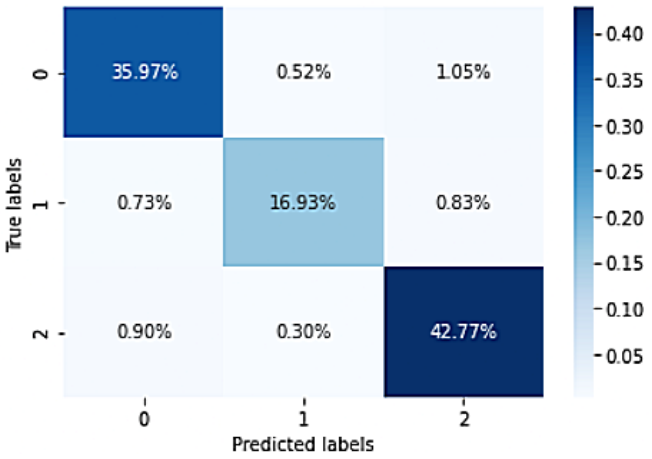

Fig. 4. Confusion Matrix for Stochastic Gradient Descent

Ergo, the F1-score captures the best of the total performance of the algorithms because it takes both precision and recall into account. So taking all into account, XGBoost had the lowest score of about 75.4%, while Gradient Boosting had similar of 76.8%. Logistic Regression and Stochastic Gradient Descent had overall 78.9% and 80.0% respectively. The performance could be improved with a bigger dataset or using inverse random sampling [17]. Regardless of discrepancy in the precision and recall, the values were still analyzed and scrutinized. The table reveals that, while the accuracy value for total data was almost comparable for all of the methods, Stochastic Gradient attained the greatest precision value of 80.1%. The recall ratings for the dataset procured by the considered algorithms were approximately similar, with each approach scoring around 80%. Nevertheless, XGBoost algorithm scored the lowest in the recall performance matrix, 75.3%, while SGD yielded the higher of about 79.9%. Figure 4 shows the confusion matrix for SGD where the 0, 1 and 2 labels stand for 'Negative', 'Neutral' and 'Positive' sentiment respectively.

VI. CONCLUSION

In this paper, sentiment analysis of Covid-19 Tweets were analyzed using different machine learning models. Our findings see that the deep learning model, Stochastic Gradient Descent performed most efficiently in the given dataset while XGBoost performed the least impressive. Computationally, SGD is faster and more efficient since it processes only one sample each time. For larger datasets as well, it can converge faster for the optimized paramaters involved. There are plenty of deep learning models that could be used for better results [18] [19]. Transformers such as roBERTa [20] has promises in this field of analyzing the sentiments to get better scores. In future, we hope to get better understanding using deeper neural networks where other languages may also be involved in tweets that would require further analysis and more computational power.

REFERENCES

[1] D. Moats and E. Borra, "Quali-quantitative methods beyond networks: Studying information diffusion on twitter with the modulation sequencer," *Big Data Soc.*, vol. 5, no. 1, p. 205395171877213, 2018.

[2] C. Soren Gordhamer, "The 5 keys to twitter's success," Nov. 2009, accessed: 2021-12-19. [Online]. Available: https://www.huffpost.com/entry/the-5-keys-to-twitters-su_b_296119

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. J. Passonneau, "Sentiment analysis of twitter data," in *Proceedings of the workshop on language in social media (LSM 2011)*, 2011, pp. 30–38.

[4] K. Wang and X. Wan, "Sentiment analysis of peer review texts for scholarly papers," in *The 41st International ACM SIGIR Conference on Research amp; Development in Information Retrieval*, ser. SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 175–184. [Online]. Available: https://doi.org/10.1145/3209978.3210056

[5] A. Yousif, Z. Niu, J. K. Tarus, and A. Ahmad, "A survey on sentiment analysis of scientific citations," *Artificial Intelligence Review*, vol. 52, no. 3, pp. 1805–1838, 2019.

[6] M. Krommyda, A. Rigos, K. Bouklas, and A. Amditis, "Emotion detection in twitter posts: a rule-based algorithm for annotated data acquisition," in *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. Los Alamitos, CA, USA: IEEE Computer Society, dec 2020, pp. 257–262. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CSCI51800.2020.00050

[7] B. Duncan and Y. Zhang, "Neural networks for sentiment analysis on twitter," in *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*. IEEE, 2015, pp. 275–278.

[8] Karthe, "Infographic : Cleaning text data python," 2015, accessed: 2021-12-20. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/06/quick-guide-text-data-cleaning-python/

[9] R. Atenstaedt, "Word cloud analysis of the bjgp."

[10] J. Brownlee, "Why one-hot encode data in machine learning?" Jul. 2017, accessed: 2021-12-20. [Online]. Available: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[12] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, no. 21, 2013.

[13] L. G. Grimm and P. R. Yarnold, *Reading and understanding multivariate statistics*. American Psychological Association, 1995.

[14] L. Bottou, *Stochastic Gradient Descent Tricks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 421–436. [Online]. Available: https://doi.org/10.1007/978-3-642-35289-8_25

[15] P. Huilgol, "Precision vs recall," 2020, accessed: 2021-12-22. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

[16] Q. Adibur Rahman Adib, H. K. Mehedi, M. S. Sakib, K. Patwary, M. Hossain, and A. A. Rasel, "A deep hybrid learning approach to detect bangla fake news," 10 2021, pp. 442–447.

[17] M. A. Tahir, J. Kittler, and F. Yan, "Inverse random under sampling for class imbalance problem and its application to multi-label classification," *Pattern Recognition*, vol. 45, no. 10, pp. 3738–3750, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320312001471

[18] A. M. Ramadhani and H. S. Goo, "Twitter sentiment analysis using deep learning methods," in *2017 7th International annual engineering seminar (InAES)*. IEEE, 2017, pp. 1–4.

[19] D. Goularas and S. Kamis, "Evaluation of deep learning techniques in sentiment analysis from twitter data," in *2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML)*. IEEE, 2019, pp. 12–17.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.