COVID Sentiment Analysis of Social Media Data Using Enhanced Stacked Ensemble

E. Prabhakar[1], P.S Aiswarya[2], A. Jamuna Banu[3], M. Kowsalya[4] ,P. Kanimozhi[5]

[1]*Assistant Professor, Dept. of Computer Science and Engineering(CSE), Nandha College of Technology, TamilNadu, India*
[2,3,4,5]*Final Year CSE, Dept. of CSE, Nandha College of Technology, TamilNadu, India*

---***---

Abstract - *COVID-19 triggered a global public health crisis and a slew of other concerns, including an economic downturn, job losses, and mental anguish. This epidemic affects people all around the world, causing anxiety, stress, worry, fear, repugnance, and poignancy in addition to the sickness. During this time, social media involvement and interaction skyrocketed, allowing people to share their thoughts and opinions on the aforementioned health issues. We can assess the people's ideas and attitudes on health status, concerns, panic, and awareness related to social issues using user-generated content on social media, which can help establish health intervention techniques and plan effective campaigns based on popular impressions. On the COVID-19 tweets dataset, we look at user sentiment at different time intervals to help with hot topics on Twitter. This research contributes a novel way of analysing social media on Twitter. It provides new information on the impact of sentiment polarity in COVID-19 tweets on tweet responses.*

Key Words: Sentiment Analysis, COVID-19, Twitter Data, Stacked Ensemble, Public Opinion, Social Media Data.

1. INTRODUCTION

Twitter, for example, has gone a mainstream venue for numerous people to share their opinions on current events as a series of tweets [1]. It has a large, global impact on popular perceptions of current events [2, 3]. The pandemic's illness burden has unavoidable consequences for population health and well-being, health resource use, social dynamics, global economies, and health technology development. Stakeholders from all around the world (governments, non-profit organizations, and healthcare groups) have been working hard to combat the COVID-19 outbreak.

Surprisingly, the COVID-19 pandemic has sparked a flood of brief informal texts on Twitter, reflecting public fears, concerns, and a new type of discrimination [4]. As a result, it's never been more important to comprehend and draw conclusions from the vast number of social media messages, such as Tweets. This knowledge will aid in the development of public health campaigns or social media events to refute misinformation and combat the global fear and stigma surrounding COVID-19.

Our effort covers the concept of assessing active users' various attitudes, such as positive, negative, and neutral sentiments, toward trending topics linked to COVID-19 at a specific time interval. This study focuses on people's positive, negative, and neutral feelings on COVID-19's top-k trending sub-topics on Twitter. The primary contributions of our research include Evaluating the effectiveness of existing algorithms and proposing a model that successfully classifies emotions.

The remainder of the paper is laid out as follows. Related work is presented in Section 2. The proposed mechanism for determining feelings is detailed in Section 3. Section 4 discusses the findings of the experiments. Finally, Section 5 brings the work to close-by outlining future research prospects.

2. LITERATURE SURVEY

By using the power of Natural language processing (NLP) to analyze the sentiment that is being transmitted in the particular data, the notion of opinion mining or sentiment analysis has been employed and developed for diverse analyses over time [5]. This section summarizes past empirical research in this field.

The BERT model is used in the paper [6] to perform Sentiment Analysis on Twitter data. The location of individual tweets was utilized to categorize the data presented in this article. The BERT model for emotion categorization was used to train the data, and the model's performance was assessed using the SVM classifier.

Using COVID-19 Twitter data, a model for assessing the influence of COVID-19 on stocks was developed in the article [7]. With an accuracy of 86.24 percent, this model was trained using supervised learning. The goal of this study was to assist businesses in predicting stock prices, developing new marketing tactics, and tracking the company's progress following the Coronavirus epidemic.

The most talked about subjects on Twitter during and after the initial wave of the COVID-19 outbreak are examined in Paper [8]. Topic extraction was done with Latent Dirichlet Allocation (LDA), while sentiment analysis was done with a Lexicon-based technique. The interests of everyone on numerous themes during the initial wave of the epidemic were well represented in this presentation. A dataset of 600,000 tweets in English was used, with 80 percent of the data being used to train the model and the remaining 20% being used to test the model. The paper used sentiment analysis to illustrate people's feelings about the most popular topics.

Paper [9] concentrates its investigation on Twitter data from all Indian states from November 2019 to May 2022. Sentiment analysis was effectively used in the obtained dataset in this article, and it was concluded that the Indian people's general sentiment was positive. Because of the increased amount of positive COVID19 instances, certain states had a higher number of tweets than others.

The paper [10] compares the sentiment of all of the COVID-19 tweets on Twitter. The sentiment of the tweets was determined using VADER sentiment analysis, BERT sentiment analysis, and Logistic Regression. The paper [11] uses two datasets: textual tweets from six different countries in April 2020, and tweets from the top ten politicians on the topic of Coronavirus. The research finishes with findings that aid in understanding sentiment and the variances between each country's sentiments. The top emotions among respondents from the six countries were found to be "trust," "fear," and "anticipation."

The Twitter data from April 30, 2020, was utilized in the paper [12] for sentiment analysis, which was done using term weighting, TF-IDF, and Logistic Regression. This algorithm was able to correctly classify the sentiment of tweets with an accuracy of 94.71 percent. This literature evaluation provided us with crucial information regarding earlier research in this sector. This clarified how we should proceed with our own planned method.

3. PROPOSED METHODOLOGY

To forecast the sentiment of COVID-19 tweets, a new model is proposed. The tweets were manually tagged after they were pulled from Twitter.

3.1 Exploratory Data Analysis

Columns like "UserName" and "Date" provide no useful information for our investigation. As a result, we don't use these features while creating models.

3.2 Data Pre-Processing

Text data pre-processing is crucial since it qualifies the raw text for mining. The goal of this stage is to remove noise such as punctuation (.,?,", etc.), special characters (@, percent, &, $, etc.), numbers (1,2,3, etc.), tweeter handle, links (HTTPS: / HTTP:), and phrases that have little weight in connection to the text.

3.3 Vectorization

A count vectorizer or a TF-IDF vectorizer can be used. Count Vectorizer creates a sparse matrix of all words in a document and the number of times they appear. TFIDF refers to the term frequency-inverse document frequency and is a numerical statistic that determines how important a word is in a corpus or collection of documents. The TF–IDF value increases in proportion to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the term, which helps to account for the fact that some terms appear more frequently than others in general.

3.4 Building Classification Models

Ordinal Multiclass classification is the problem at hand. Because there are five different sorts of attitudes, we must train our models so that they can correctly categorize the test dataset. I'll use Naive Bayes, Logistic Regression, Random Forest, XGBoost, Support Vector Machines, CatBoost, and Stochastic Gradient Descent to create several models.

3.5 Enhanced Stacked Ensemble

The Enhanced Stacked Ensemble method is a supervised ensemble machine learning algorithm that integrates the finest selection combination of classifiers using the stacking model.

3.6 Evaluation

This paper compares the performance of various existing algorithms with proposed algorithms during the evaluation process.

4. EXPERIMENTAL RESULTS

The collected dataset consists of 41157 records and 6 attributes. All tweets came only from March and April month of 2020. There are 5 classes. The positive class consists of 11422 records, the Negative class consists of 9917, and the Neutral class consists of 7713, whereas the Extremely Positive has 6624 records and the Extremely Negative has 5481 records. Accuracy is used to measure the performance of the classification algorithms. The test accuracy is calculated for all existing algorithms and proposed algorithms.

Table -1: Accuracy

Algorithm	Accuracy
Support Vector Machines	85
Random Forest	83
Naive Bayes	79
XGBoost	74
Logistic Regression	86
Stochastic Gradient Decent	86
CatBoost	85
Enhanced Stacked Ensemble	88

Test accuracy of various algorithms is presented in Table -1. Naïve Bayes and XGBoost provide less than 80% accuracy. The performance of existing and proposed algorithms is illustrated in Table -1. From the result, it is clear that the results of the enhanced stacked ensemble outperform all other existing algorithms.

5. CONCLUSION

People use social media as one of their primary sources of information. Unlike traditional media, social media allows multiple organizations to communicate differing viewpoints on the same occurrence, and readers can instantly respond to these tweets. Social media, notably Twitter, has become one of the most essential, shared channels for revealing and tracking the COVID-19 pandemic trend because of its simplicity and involvement. As a result, social media (e.g., Twitter) has become an increasingly public medium for comprehending people's thoughts and reactions to current events with various narratives. Along with the COVID-19 trend, it is also necessary to research the readers' feelings. This study offers a systematic approach to understanding people's reactions to negative and good COVID-19 pandemic reports.

REFERENCES

[1] Monachesi, Paola, and Saskia Witte born. "Building the sustainable city through Twitter: Creative skilled migrants and innovative technology use." Telematics and Informatics 58 (2021): 101531.

[2] Gjerstad, Peder, et al. "Do President Trump's tweets affect financial markets?" Decision Support Systems 147 (2021): 113577.

[3] Gruzd, Anatoliy, and Philip Mai. "Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter." Big Data & Society 7.2 (2020): 2053951720938405.

[4] Li, Jessica, and H. Raghav Rao. "Twitter as a rapid response news service: An exploration in the context of the 2008 China earthquake.", The Electronic Journal of Information Systems in Developing Countries 42.1 (2010): 1-22.

[5] Apoorva Shete, Rohan Pradyuman, Sheetal Gondal, "Sentiment Analysis of COVID-19 Vaccine Tweets", International Research Journal of Engineering and Technology (IRJET), Volume 8, Issue 8, 2021.

[6] Singh, M., Jakhar, A.K. & Pandey, S. Sentiment analysis on the impact of coronavirus in social life using the BERT model. Soc. Netw. Anal. Min. 11, 33 (2021).

[7] Yuvraj Jain, and Vineet Tirth, "Sentiment Analysis of Tweets and Texts Using Python on Stocks and COVID-19", International Journal of Computational Intelligence Research, ISSN 0973-1873, Volume 16, Number 2 (2020), pp. 87-104.

[8] Manal Abdulaziz, Alanoud Alotaibi, Mashail Alsolamy, and Abeer Alabbas, "Topic-based Sentiment Analysis for COVID-19 Tweets" International Journal of Advanced Computer Science and Applications (IJACSA), 12(1), 2021.

[9] T. Vijay, A. Chawla, B. Dhanka and P. Karmakar, "Sentiment Analysis on COVID-19 Twitter Data," 2020, 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), 2020, pp. 1-7.

[10] A. J. Nair, V. G, and A. Vinayak, "Comparative study of Twitter Sentiment on COVID - 19 Tweets," 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1773-1778.

[11] G. Matosevic and V. Bevanda, "Sentiment analysis of tweets about COVID-19 disease during the pandemic," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 1290-1295.

[12] Imamah and F. H. Rachman, "Twitter Sentiment Analysis of Covid-19 Using Term Weighting TF-IDF and Logistic Regression," 2020 6th Information Technology International Seminar (ITIS), 2020, pp. 238-242.