Analyzing COVID-19 Sentiments on Twitter: An Effective Machine Learning Approach

Irfan Qutab[1]; Muhammad Aqeel[2]; Unaiza Fatima[3;] Imtiaz Ahmed[4]
School of Software[1,2,3]; School of Electronics and Information[4]
Northwestern Polytechnical University Xi'an, China

Abstract:- The COVID-19 pandemic has brought about a surge in online discussions and social media activity, making it crucial to analyze public sentiment towards the virus and related topics. This thesis focuses on Sentiment Analysis of COVID-19 data on Twitter, employing Multinomial Logistic Regression as the primary classification algorithm. This research explores Sentiment Analysis of COVID-19 data on Twitter using Multinomial Logistic Regression. It constructs a tweet dataset reflecting various sentiments—positive, negative, and neutral. The data undergoes preprocessing, and a Sentiment Analysis model is built, with 70% of data for training and 30% for testing. The model uses Count-Vectorizer, Tf-idf for feature extraction, and Multinomial Logistic Regression to classify tweets. The study achieves state-of-the-art results with a high accuracy of 95.14%, demonstrating the effectiveness of this approach. The results offer valuable insights into public sentiment during crises, aiding in decision-making and communication strategies.

***Keywords**:- Sentiment Analysis, COVID-19, Twitter, Multinomial Logistic Regression, Social Media.*

I. INTRODUCTION

Sentiment Analysis reveals a semantic comprehension of user-generated content, such as blog and social network posts, where readers express their opinions and sentiments [1]. Sentiment Analysis, often known as "opinion mining," is a technique for understanding the motivations behind online discussions to learn more about people's opinions, beliefs, and feelings [2]. Feelings such as joy, anger, sadness, love, fear, and anxiety are all examples of emotions [3]. Sentiment Analysis provides valuable insight into the public's perspective on a topic while monitoring social media. An opinion or viewpoint is defined as an expression of a person's inner state of mind. Personal feelings, whether favorable or bad, are what a person is trying to portray when they write or speak [4]. By combining techniques from NLP, TA, and CS, Sentiment Analysis can determine the author's emotional stance on a topic from their written or spoken words. The tool aids in identifying whether a text expresses positive, negative, or neutral emotions. Sentiment Analysis is a powerful tool for understanding public opinion, especially that of your target market.

The respiratory system is the primary target of the coronavirus, one of the worst illnesses. Historically, two epidemics of coronary viruses—severe acute respiratory syndrome (SAR-CoV) and Middle East respiratory disease (MERS)—have been recognized as agents that represent a significant risk to public health (MERS-CoV). The coronavirus, which can infect both humans and animals, is thought to have originated in Wuhan, China [5]. On March 11, 2020, the World Health Organization proclaimed a global pandemic due to the spread of the Covid19 virus. (The World Health Organization, 2020). When compared to SARS-CoV, the virus responsible for the 2003 SARS pandemic that swept through Singapore, Hong Kong, Taiwan, Canada, and other countries, SARS-CoV-2 is more infectious. As the number of confirmed cases of COVID-19 continues to climb rapidly, efforts to prevent and control the spread of this virus are of the utmost importance [6].

Sentiment Analysis has gained a lot of attention in recent years, and it has been implemented in nearly every industry. Most software and hardware are designed to work with speakers of English and other Indo-European languages. People use the Internet for a variety of reasons, including gathering information and actively contributing content in the form of comments and ideas on social media platforms. You may see the latest reactions and responses. There is evidence from polls and surveys that people are affected by the opinions and feedback of those who read them [7].

In this research a Sentiment Analysis model is developed using Multinomial Logistic Regression for classification of tweets. The process involved tokenizing data, extracting features using Count-vectorizer and Tf-idf, and classifying tweets into negative, positive, and neutral categories. The model's effectiveness was evaluated using metrics like accuracy, precision, recall, and F1-score.

An introduction to sentiment classification and the current study is presented in this part. Literature review is elaborated in Section 2. There are several processes involved in constructing these models, which are explained in Section 3. Test results of the models are described in section 4. The conclusions and suggestions for further research have also included at the end.

II. LITERARURE REVIEW

The massive death toll from the COVID-19 epidemic threatens the global population's health, food supply, and employment. In addition to a high mortality toll, the COVID-19 outbreak has threatened worldwide food supplies, occupational safety, and public health. A variety of social media platforms and financial markets reacted differently to COVID-19 dynamics such as death rates, transmission characteristic, period of national virus, and early casualties. During the lockdown, many took to social media to express their frustrations, which aided in the global dissemination of news of the pandemic. Considering this terrible scenario, it is critical to examine tweets by looking at trending terms associated with the pandemic [8]. The Researchers discussed a cross-language sentiment study of tweets from European users during the COVID-19 outbreak. Interactions on social media are chaotic and distracting, and there is too much data being generated to process it all by hand. This led the authors to the conclusion that automated processes are required to reliably gather useful data. The authors analyze the sentiments expressed in tweets that were collected during Europe's initial COVID-19 epidemic [9].

Exploring the essence of any sentence might reveal contextual polarity. It can be detected by first recognizing the sentence’s neutral or polar meaning [10]. For sentiment analysis on English Twitter data, tree kernel model and feature-based model performed well [11]. Sentiment analysis is a tool that may be used in a variety of different areas than product reviews. It may be used, for example, to extract helpful information from newspaper articles [12]. It may be used to estimate stock market changes based on the environment in the area [13]. The opinions of political candidates can be evaluated through sentiment analysis. Aside from opinions gleaned through political discussions, predictions of election outcomes may also be made using social media reviews of political parties and politicians [14].

Sentiment classification model for roman Urdu text was proposed. In which text is classified into four categories named as politics, sports, education and religious [15]. Roman Urdu text sentiment analysis was analyzed and evaluated. They described the systems which are developed for Roman Urdu sentiment classification. A comparison of researchers' results was also made [16]. N-grams identify text patterns, while TF-IDF highlights important terms for sentiment classification. Combining these methods improves accuracy and reduces computational complexity in sentiment analysis [17].

III. METHODOLOGY

We have divided our framework into six different phases. We have used twitter dataset that is labeled as positive, negative and neutral. After performing the necessary data preprocessing and feature extraction, we trained our models on the training set. Using our test dataset that was initially split out from the preprocessed collected comments, we evaluated each model’s performance using recall, precision, and F1-Score, accuracy, and Confusion matrix. To validate the model’s performance K-fold cross validation is used. Figure 1 represents the process flow of our model, with subsequent subsections explaining it in better detail.

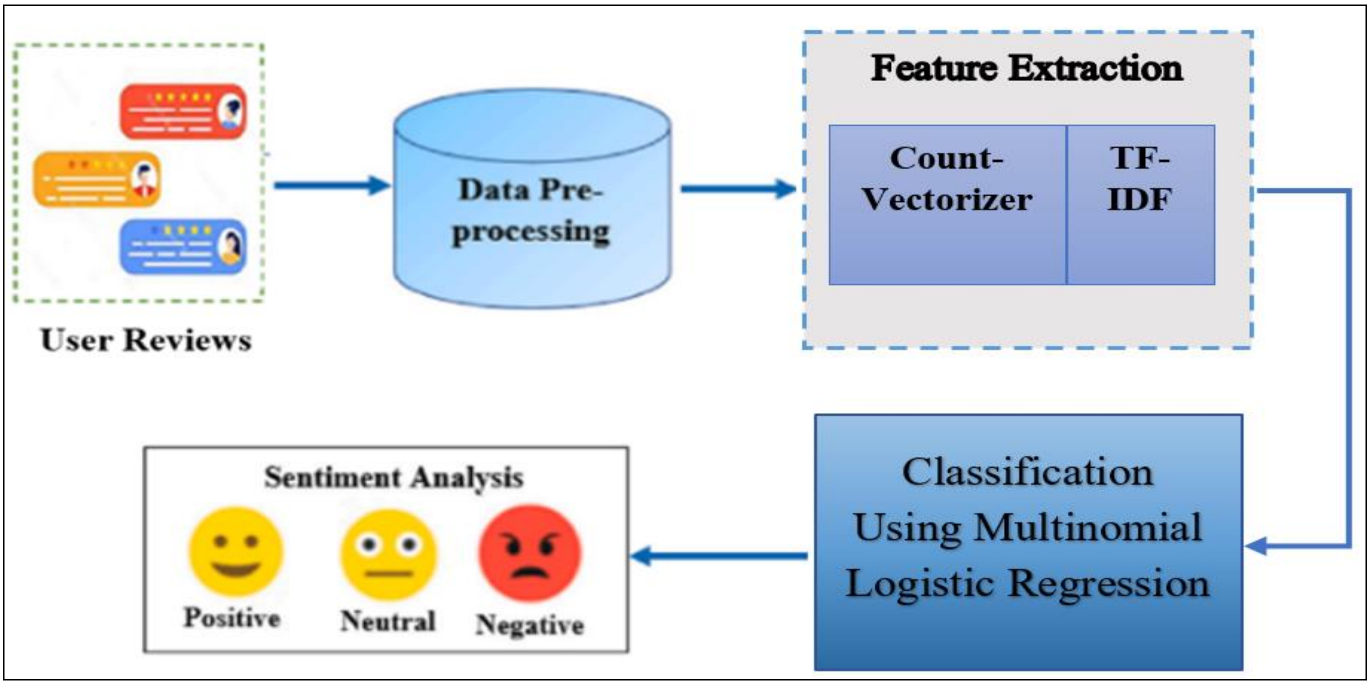

Fig 1: Methodology for Sentiment Classification of Covid-19 Comments

A. Dataset

In our research on Sentiment Analysis of Covid-19 comments on Twitter, we utilized a dataset specifically focused on the Covid-19 pandemic. While Twitter offers various methods for data mining, such as the Twitter API, Tweepy, and web scraping tools, our dataset was obtained from Git-hub, providing a valuable resource for our classification. The dataset comprised a total of 139,328 tweets, each representing feedback or comments shared by users on Twitter. These tweets were labeled as negative, positive, or neutral, with 104,222 negative comments, 9,753 positive comments, and 25,353 neutral comments. This diverse dataset allowed us to capture a broad range of sentiments expressed by individuals regarding Covid-19 on Twitter.

B. Data Pre-Processing

This process involves removing punctuation, numbers, and stop words, with Python libraries like NLTK and WordNet Lemmatizer employed to normalize and extract key terms.

- **Data Auto Correction:** Auto-correction addresses text inaccuracies, fixing typos to enhance sentiment categorization. Techniques include lowercasing to ensure consistency and removing stop words to focus on meaningful content, improving classification accuracy.
- **Stemming:** Stemming simplifies words by removing suffixes, reducing them to their base forms, like "running" to "run." This process helps in reducing the number of unique words, enhancing the efficiency of sentiment analysis algorithms.
- **Lemmatization:** Lemmatization refines text by converting words to their dictionary form, such as changing "running" and "ran" to "run." This standardization improves classification accuracy by grouping similar words together and focusing on their core meanings.
- **Tokenization:** Tokenization splits the text into individual words or units, accommodating Twitter-specific features like hashtags and mentions. This step enables detailed analysis of how people express sentiments about COVID-19 on Twitter.

C. Training and Testing

The dataset of 139,328 COVID-19 tweets was split into 70% for training (97,530 tweets) and 30% for testing (41,799 tweets). The training set was used to teach the Sentiment Analysis model the patterns associated with various sentiments. The testing set then evaluated the model's accuracy and ability to generalize to new, unseen data, providing a robust assessment of the model's effectiveness in classifying the emotional states of the tweets.

D. Pipelining

Pipelining is method in which the computer processor performs different tasks in a logical way through programming instructions in multiple steps. This processing can be continued, in order and overlapped manners. A pipeline consists of a series of stages that are linked together to form a pipe. The instructions come in through one side and leave through the other. The pipeline's goal is to create a number of phases that can be jointly cross-validated using varying parameter values. Pipeline must use fit and transform method. Generally, pipeline implements fit-transform() and fit-predict(). In our work we have constructed a pipeline for converting words into vectors, features extraction by using count vectorizer and Tf-idf and fitting the model to predict the results.

E. Feature Extraction

The process through which the dataset can be converted into the bundle of features is called "extraction of features". Feature is a pattern entity whose quantity can be determined by one of these characteristics [18]. Choosing the purpose for collecting the features is important because machine learning approach highly depends on its features. TF-IDF Vectorizer is used for Feature Extraction.

- **Count-Vectorizer:** Count-Vectorizer is a text preprocessing tool in Natural Language Processing (NLP) that converts a collection of text documents into a matrix of token counts. It tokenizes the text by splitting it into individual words (or tokens) and then builds a vocabulary of unique words across the dataset. Each document is represented as a vector of word counts, where each word corresponds to a column in the matrix. Count-Vectorizer is commonly used for feature extraction in text classification tasks, helping models understand the frequency of words within the text.
- **TF-IDF:** Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical method used to evaluate how important a word is within a document relative to a collection of documents. It combines the frequency of a word in a specific document with the inverse frequency of that word across the entire dataset. TF-IDF is widely used in Natural Language Processing (NLP) for feature extraction in Sentiment Analysis. By reducing the weight of common, less informative words, TF-IDF improves the accuracy of sentiment classification models.

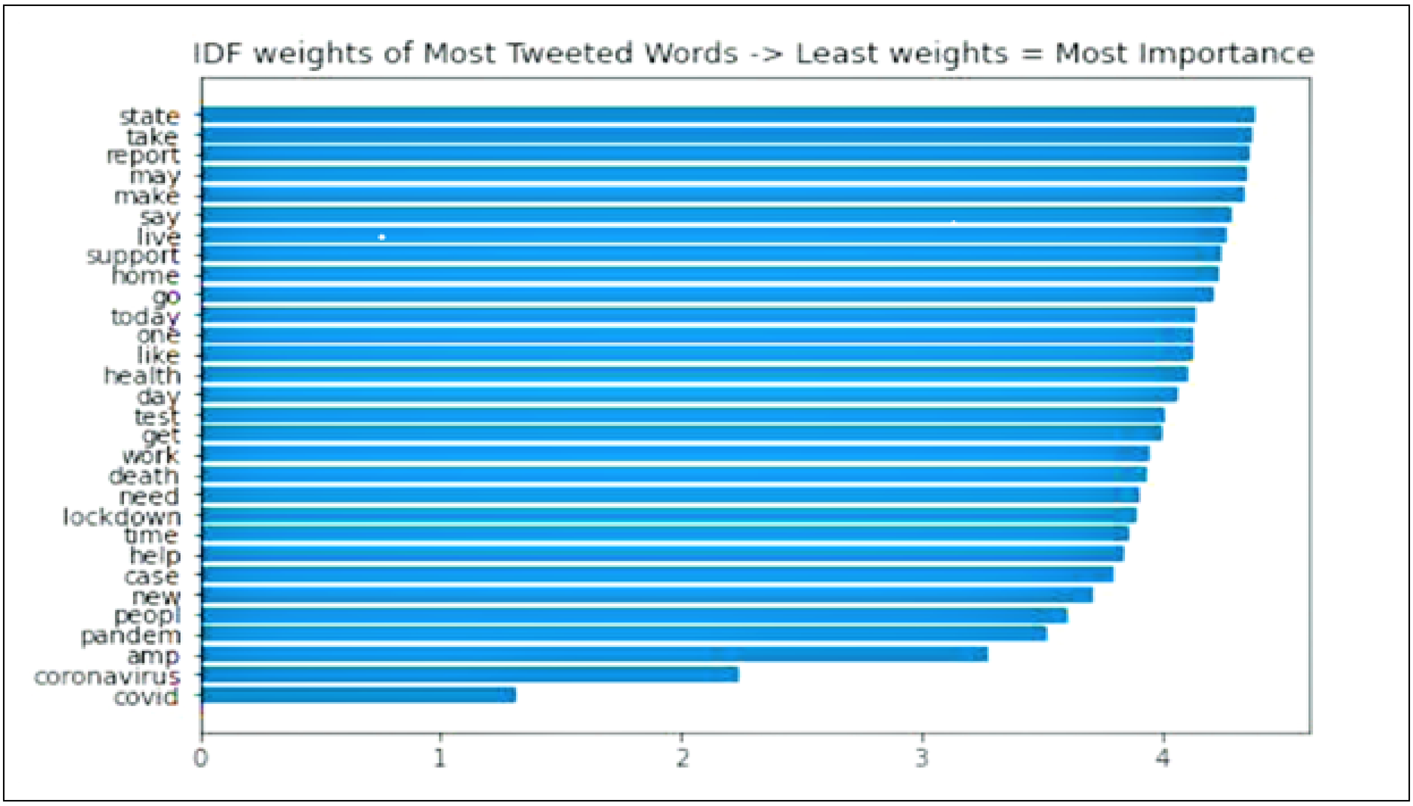

Fig 2: IDF Weights of Most Tweeted Words

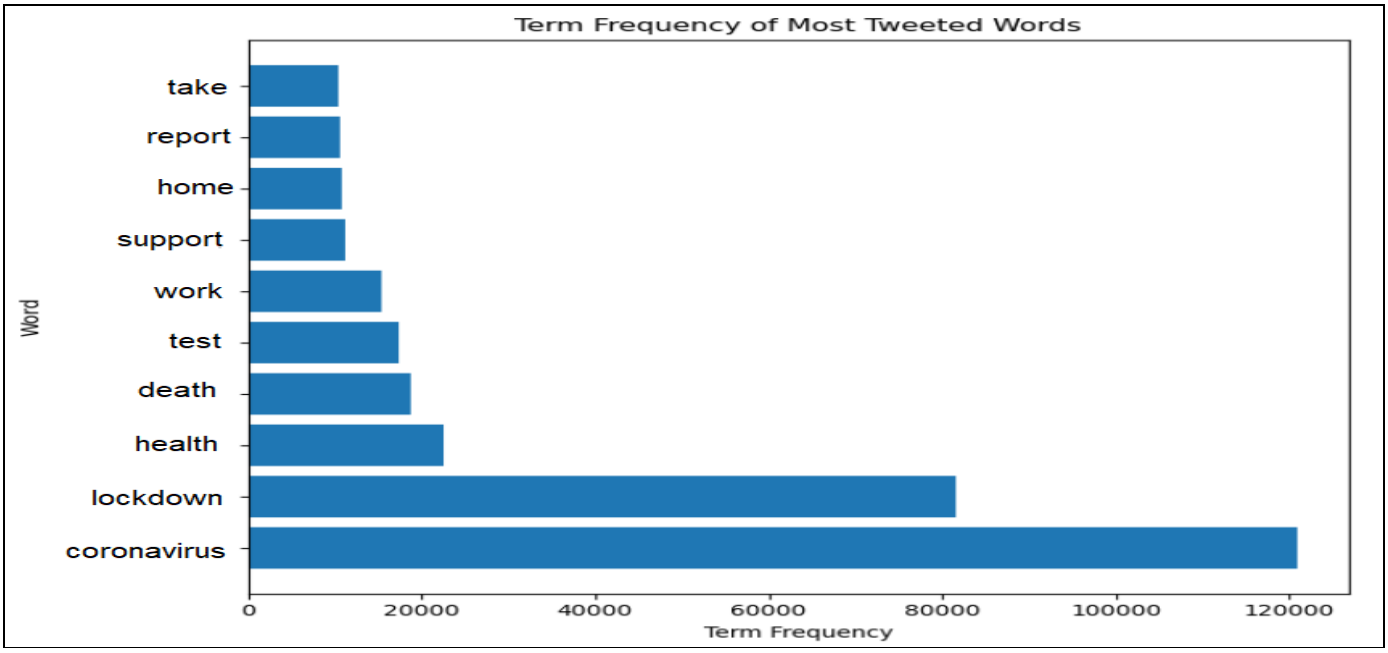

Fig 3: Term Frequency of Most Tweeted Words

F. Logistics Regression

Logistic Regression is a binary classifier. By using Logistic Regression data can be classified into two classes i.e. positive or negative, yes or no, mail is spam or not etc. It is also used to measure the probability of an outcome of a binary event and to resolve classification problems. It works with dependent variables and independent variables. The independent variables can affect the dependent variables.

- *Logistic/Sigmoid Function:*

$$h_\theta(X) = g(\theta^T . X) \tag{1}$$

$$g(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

$$where\, z = \theta^T . x$$

- *To Calculate Cost of the Function:*

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} (y)^i \; log h_\theta(x^i) + (1 - y^i) log\left(1 - h_\theta(x^i)\right) \tag{3}$$

- *To Calculate Gradient Descent:*

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{4}$$

After derivation of $J(\theta)$ the equation becomes:

$$\theta_j = \theta_j - \alpha \sum_{i=1}^{m} (h_\theta \, [\, (x)^i - (y)^i \,] \, x_j^i \tag{5}$$

(Update all θ_j simultaneously) and α is the learning rate.

G. Multinomial Logistic Regression

To classify data where there are more than two possible outcomes, researchers have developed Multinomial Logistics Regression which is also called Soft-max Regression [19]. In other words, it is a method for determining the probability of a set of outcomes associated with a categorically separated variable that are independent to one another.

- *For k classes Formula is Written Below:*

$$P\,(\, y = c|x; \theta_1, \theta_2, \theta_3, \ldots, \theta_k \,) = \frac{\exp(\theta_c^T x)}{\sum_{j=1}^{k} exp\;(\theta_j^T x)} \tag{6}$$

Where $\theta_1, \theta_2, \theta_3 \ldots \theta_k$ are the parameters.

- *Soft-Max Function:*

$$y|x = \theta_1, \theta_2, \ldots, \theta_k \sim \text{Multinomial}(\emptyset_1, \emptyset_2, \ldots, \emptyset_k) \tag{7}$$

$$\text{where } \emptyset_j = h_{\theta_j}(x)$$

- *Classification by Multinomial Logistic Regression Formula for More Than Two Classes:*

$$h_{\theta_c}(x) = \frac{\exp(\theta_c^T x)}{\sum_{j=1}^{k} exp\;(\theta_j^T x)} \tag{8}$$

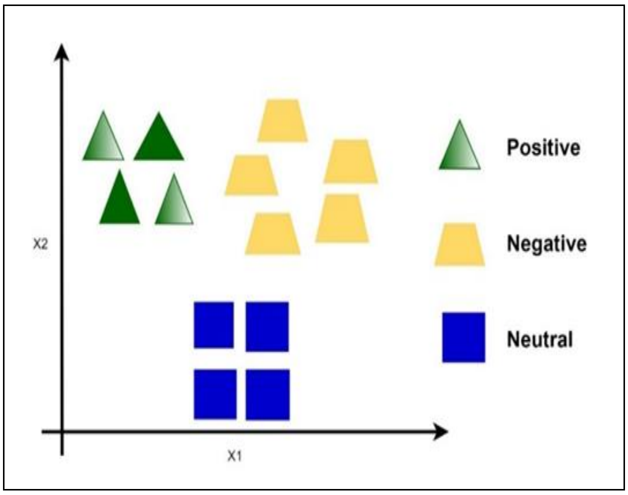

Fig 4: Data Classification in MLR

IV. EXPERMENTAL RESULTS

A. Evaluation Metrics

We have evaluated the efficiency of the classifier by using precision, recall and F1-score. Confusion Matrix of the model is also represented to show the performance of the model.

B. Accuracy

For evaluating classification model, accuracy is a key metric. The magnitude to which a model is successful in its predictions is a basic definition of its accuracy. There is a formal definition of accuracy which looks like this:

$$Accuracy = \frac{No. of \; correct \; predictions}{Total \; no. of \; predictions} \tag{9}$$

For binary classification accuracy in terms of negatives and positives is determined as follows:

$$\text{Accuracy} = \frac{TN + TP}{TP + TN + FP + FN} \tag{10}$$

C. Precision & Recall

The terms "precision" & "recall" are frequently employed in the context of information extraction. Precision is the record numbers that have been retrieved, and recall is the total record numbers that have been recovered. Since Precision and recall varies inversely to each other, this emphasizes the significance of having a reliable classification system to provide context for their differences.

- *In a Classification Task*

$$\text{Precision} = \frac{True\ Positive}{False\ Positive + True\ Positive} \tag{11}$$

$$\text{Recall} = \frac{True\ Positive}{False\ Negative + True\ Negative} \tag{12}$$

D. F1-Score

F1 is similar calculation of the check it is also called (F-measurement or F-score). Precision p is commonly measured as the proportion of accurately identified positive outcomes, which are divided by proportion of all samples categorized as positive, while recall r is the proportion of accurately identified positive results, which are divided by proportion of all examples categorized as positive.

$$F1-Score = \frac{2*(precison*recall)}{precison+recall} \tag{13}$$

Table 1: Classification Result Using MLR

Test Accuracy: 0.9514581688557142				
Test Report				
	Precision	Recall	f1-score	support
Neu	0.97	0.97	0.97	31390
Neg	0.91	0.92	0.92	7487
Pos	0.88	0.84	0.86	9299
Accuracy			0.95	41799
macro avg	0.92	0.91	0.91	41799
weighted avg	0.95	0.95	0.95	41799

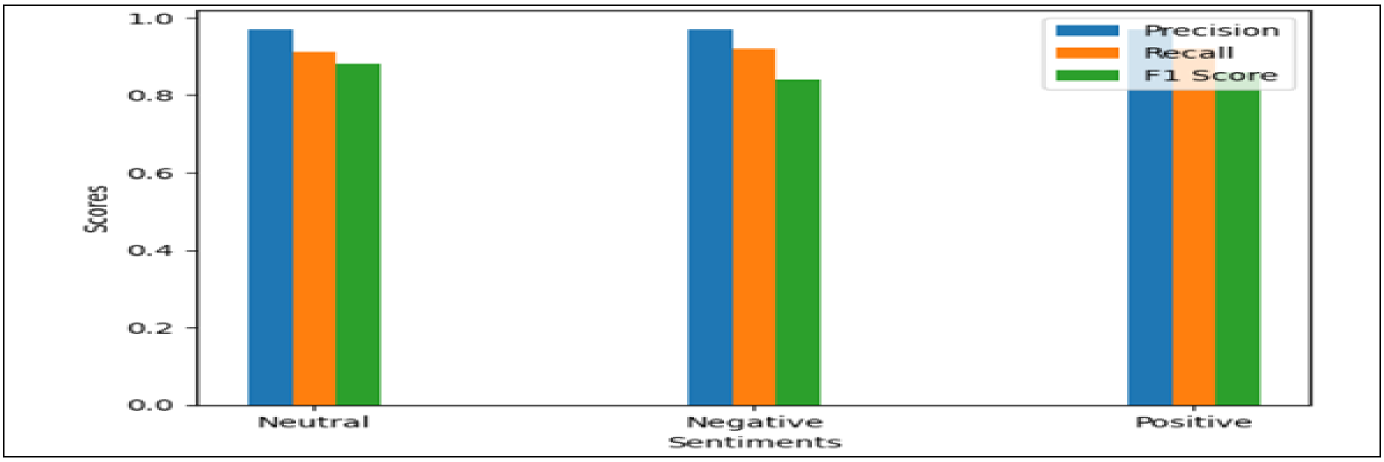

Fig 5: Metrics Comparison for Different Sentiments

E. Confusion Matrix

The confusion matrix, also known as the error matrix in machine learning and classification, assumes a matrix structure that highlights errors. Its purpose is to assess the classifier's effectiveness. By depicting both predicted and actual values, the confusion matrix offers a visual representation of disparities. This evaluation draws on insights from the confusion matrix, illustrated in Figure 6. Which encompasses metrics like True Positive (TP, True Negative (TN), False Negative (FN) and False Positive (FP). Correct predictions are positioned along the diagonal for visualization.

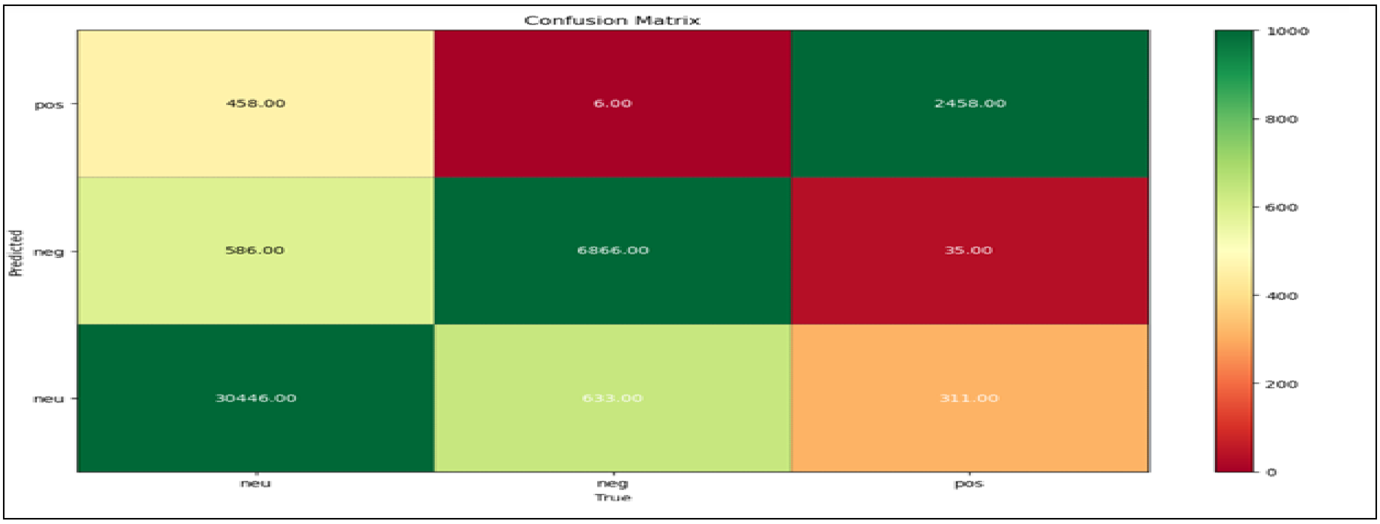

Fig 6: Confusion Matrix

F. K-Fold Cross Validation

Cross-validation is a resampling method used on limited data basis to test machine learning models. The method has a parameter k that is the same as the number of classes to be subdivided for a data set. As such, the procedure is also named k-fold cross validation. If a value for k is selected, it may be used rather than k in the model relation of k = 10 times the cross validation. The initial data is broken into a subset (fold) by k-fold cross-validation, each with the same magnitude, and is used by k-fold to measure the system validity.

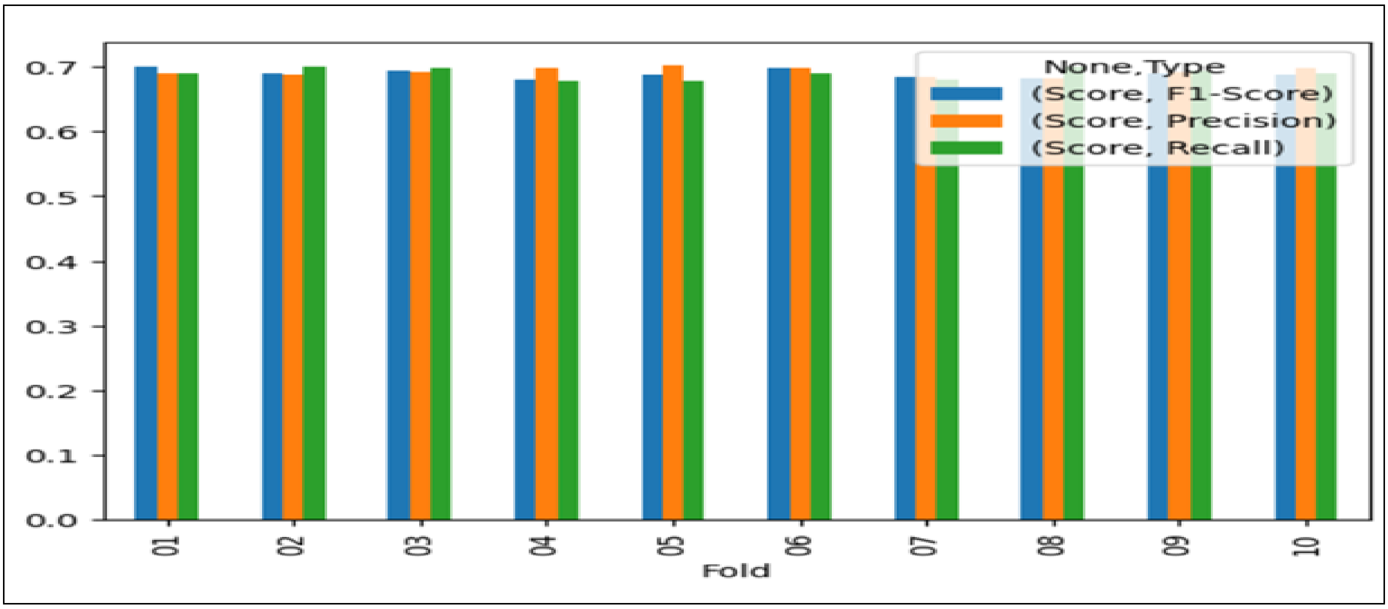

Fig 7: Result Dependent on K-Fold Number

In the second experiment, we examine how well Multinomial Logistic Regression works when we vary fold by a factor of 2–20. However, results are unaffected by different fold values. The study found that the significance of Folding has an effect on the accuracy of each operation. Up to 99% accuracy has been reached in multinomial functional classification. Performance of each approach is independent of fold size.

G. Learning Curve

To calculate the efficiency of the model, learning curves are used. The output of a training set and test set typically indicates the "risk or cost / score vs size". In the case of classifiers, we can always use score or 1-score versus the training set & testing set size to draw the learning curves. The learning curves are useful in further iteration back to determine how much data we might need to optimally train the model.

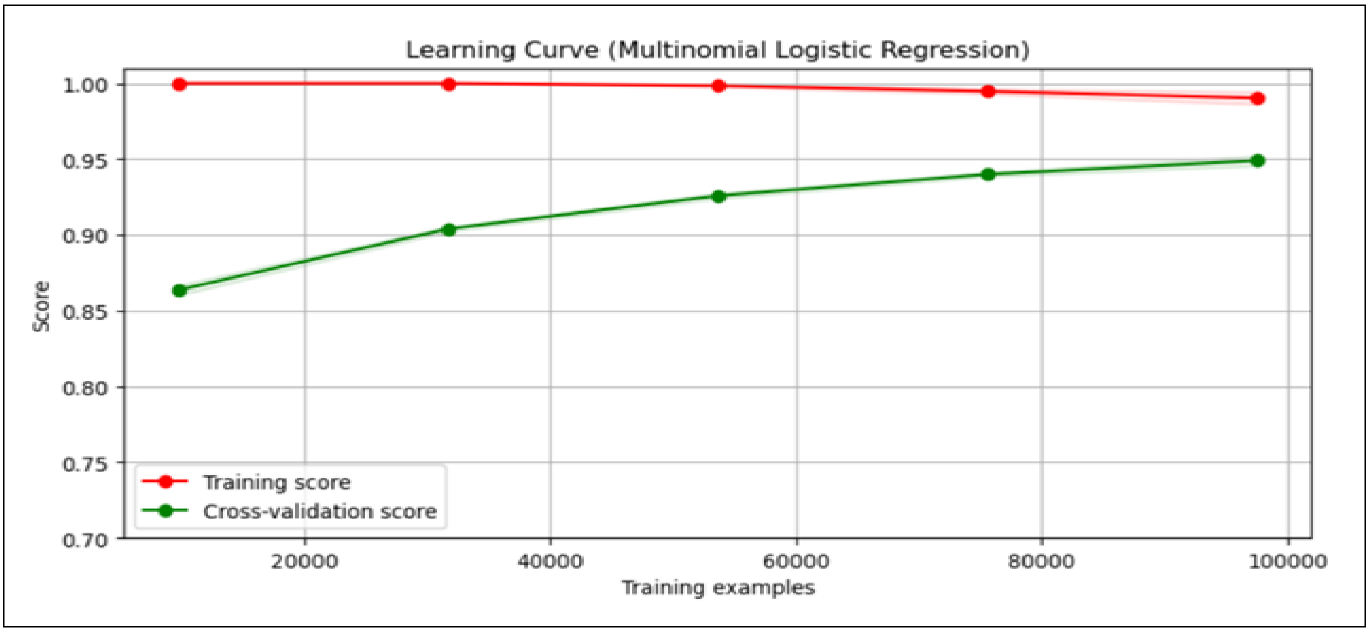

Fig 8: Training & Testing Curve

H. Testing Scenarios of Different Data Samples

The experiment aimed to evaluate the performance of the proposed model by using different ratios of training and testing samples. The table showcases the ratios along with the corresponding amounts of data allocated for training and testing. Table II show the sampling ratios.

Table 2: Sampling Ratios for Training and Testing

Data	Data Training Amount	Data Testing Amount
50%:50%	69664	69664
60%:40%	83596	55732
70%:30%	97530	41799
80%:20%	111462	27866
90%:10%	125395	13933

The bar graph compares the performance of precision, recall, and F1-score across different training and testing sample splits in a sentiment analysis model. As the training data proportion increases from 50:50 to 90:10, the performance metrics show consistent improvement, indicating that a higher proportion of training data enhances the model's accuracy in classifying sentiments. This suggests that a well-balanced and larger training dataset is crucial for achieving optimal results in sentiment analysis.

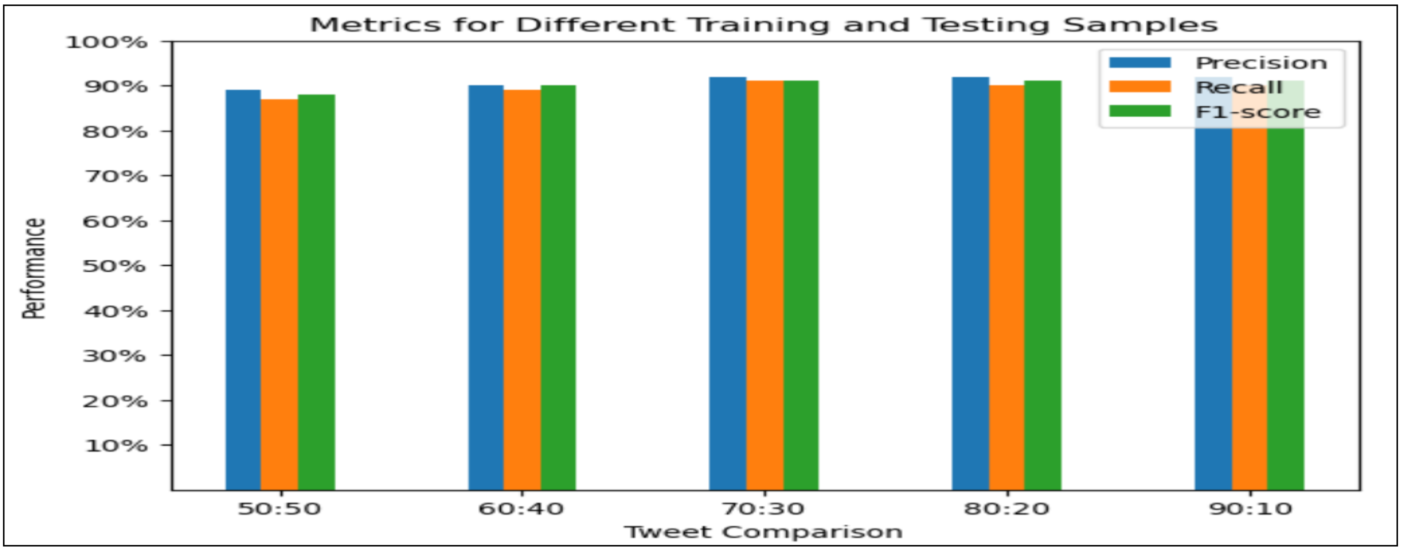

Fig 9: Result Based on Comparison of Data

I. Comparison of Previous Studies with Proposed Model

Previous studies on sentiment analysis of COVID-19 Twitter data used various machine learning models, including Naïve Bayes, SVM, Logistic Regression, and deep learning models, achieving accuracies up to 94%. However, Multinomial Logistic Regression was not used. Different approaches, such as hybrid models with BERT and SVM, LSTM-based models, and Random Forest with TF-IDF, yielded varying accuracy rates from 80.9% to 91.2%. Other studies explored sentiment polarization, misinformation spread, and the impact of social networks, using models like BERT, Graph Convolutional Networks, and multi-modal approaches. The results are compared with the proposed model in Table 3.

Table 3: Comparison of Proposed Model with the State of Art Models

Model	Precision	Recall	F1-Score	Accuracy
LR	0.91	0.92	0.90	0.94
BERT & SVM	0.87	0.85	0.86	0.87
Random Forest	0.79	0.76	0.77	0.81
SVM	0.85	0.84	0.83	0.86
Graph Based DL Model	0.89	0.87	0.88	0.90
Proposed Model	0.93	0.92	0.93	0.95

The bar graph presents a comparison between the proposed model and previous studies in terms of accuracy. With an accuracy of 95.14%, the proposed model surpasses the performance of the other models, establishing its effectiveness in Sentiment Analysis. The graph effectively illustrates the varying levels of accuracy. Overall, the graph serves as compelling evidence, highlighting the superior performance of the proposed model when compared to previous studies. It reinforces the significance and credibility of the proposed model in addressing Sentiment Analysis tasks.

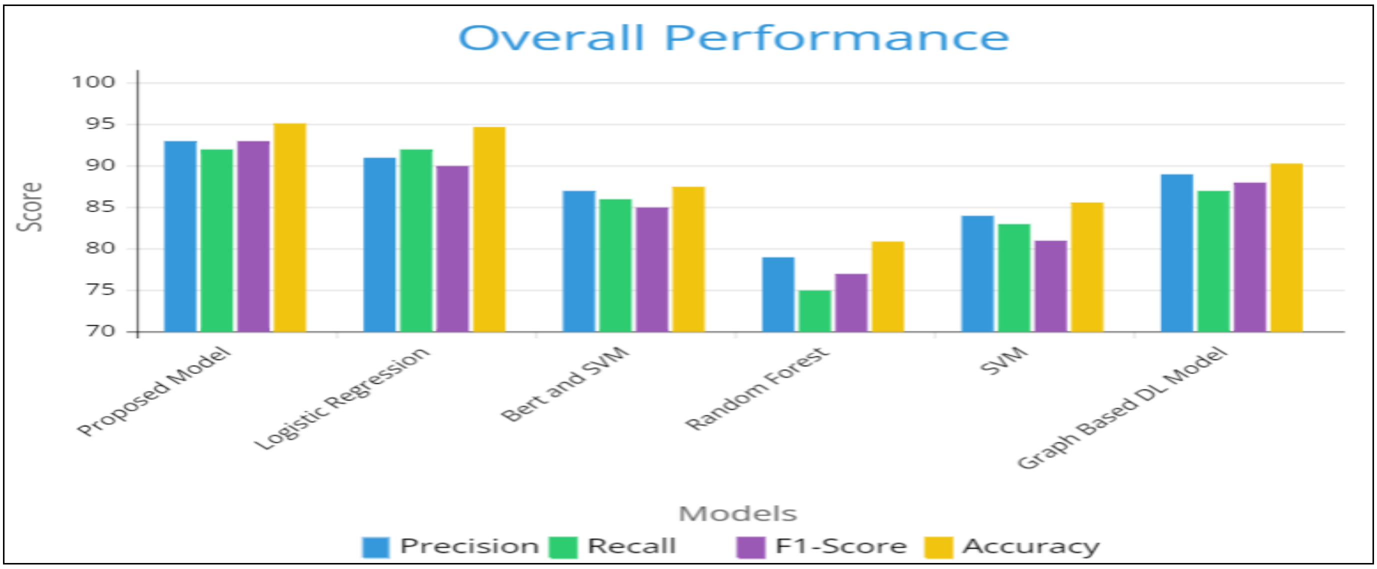

Fig 10: Comparison of Proposed Model with Baseline Models

V. CONCLUSION

This research classified COVID-19-related sentiments on Twitter into positive, negative, and neutral categories using Multinomial Logistic Regression. The dataset consisted of 139,328 comments, which were preprocessed and divided into 70% for training and 30% for testing. A pipeline approach, incorporating techniques like Count-vectorizer and Tf-idf, was used for feature extraction. The model achieved state-of-the-art results and achieved 95.14% accuracy, with consistency verified through K-Fold cross-validation. Changes in fold values from 2 to 20 did not affect the results, ensuring reliability in the model's performance.

ACKNOWLEDGMENT

I wish to convey my sincere appreciation to my research fellows for their invaluable guidance and support during the progression of this research endeavor. Their expertise and encouragement have played a pivotal role in shaping the outcome of this study.

REFERENCES

[1]. E. Doğan and B. Kaya, "Deep learning based Sentiment Analysisand text summarization in social networks," International Artificial Intelligence and Data Processing Symposium (IDAP),IEEE., pp. 1-6, 2019 .

[2]. W. P. Ramadhan, S. A. Novianty and S. C. Setianingsih, "Sentiment Analysisusing multinomial logistic regression," International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC),IEEE, pp. 46-49, 2017.

[3]. Agarwal, B. Xie, I. Vovsha, O. Rambow and R. J. Passonneau, "Sentiment Analysisof twitter data," In Proceedings of the workshop on language in social media, pp. 30-38, 2011.

[4]. Tyagi and N. Sharma, "Sentiment Analysisusing logistic regression and effective word score heuristic," International Journal of Engineering and Technology (UAE), vol. 7(2.24), pp. 20-23, 2018.

[5]. D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang and Z. Peng, "Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China," vol. 323(11), pp. 1061-1069, 2020.

[6]. Y. Y. Zheng, Y. T. Ma, J. Y. Zhang and X. Xie, "COVID-19 and the cardiovascular system," Nature reviews cardiology, vol. 17(5), pp. 259-260, 2020.

[7]. W. Medhat, A. Hassan and H. Korashy, "Sentiment Analysisalgorithms and applications: A survey," Ain Shams engineering journal, vol. 5(4), pp. 1093-1113, 2014.

[8]. K. Chakraborty, S. Das and A. K. Kolya, "Sentiment Analysisof covid-19 tweets using evolutionary classification-based LSTM model.," In Proceedings of Research and Applications in Artificial Intelligence, no. Springer, Singapore, pp. 75-86, 2021.

[9]. Kruspe, M. Häberle, I. Kuhn and X. X. Zhu, "Cross-language Sentiment Analysisof european twitter messages duringthe covid-19 pandemic," arXiv:2008, p. 12172, 2020.

[10]. Wilson, T., Wiebe, J., & Hoffmann, P. (2005, october). Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of human language technology conference and conference on empirical methods in natural language processing, 347-354.

[11]. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on language in social media (LSM 2011), 30-38.

[12]. Xu, T., Peng, Q., & Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. Knowledge-Based Systems, 35, 279-289.

[13]. Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowledge-Based Systems, 41, 89-97.

[14]. Maks, I., & Vossen, P. (2012). A lexicon model for deep sentiment analysis and opinion mining applications. Decision support systems, 53(4), 680-688.

[15]. Qutab, I., Malik, K. I., & Arooj, H. (2020). Sentiment Analysis for Roman Urdu Text over Social Media, a Comparative Study. arXiv preprint arXiv:2010.16408.

[16]. Qutab, I., Malik, K. I., & Arooj, H. (2022). Sentiment Classification Using Multinomial Logistic Regression on Roman Urdu Text. International Journal of Innovations in Science & Technology, 4(2), 223-335.

[17]. T. Sahni, C. Chandak, N. R. Chedeti and M. Singh, "Efficient Twitter Sentiment Analysisusing subjective distant supervision," International Conference on Communication Systems and Networks (COMSNETS), IEEE, pp. 548-553, 2017.

[18]. Croft, W. B., Metzler, D., & Strohman, T. (2010). Search engines: Information retrieval in practice (Vol. 520). Reading: Addison-Wesley.

[19]. Vruniotis, V. (2017). Machine Learning Tutorial: The Multinomial Logistic Regression (Softmax Regression)| Datumbox. Blog. datumbox. com. Np.