IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis of COVID 19 Tweets using Machine Learning Algorithm

Dr. Santosh Kumar Singh[1] Asst. Prof. Mithlesh Vishwakarma[2]

Mr. Aayush Pande[3], Mr. Raj Mestry[4]

Head of Dept.[1] Assistant Professor[2] PG Student[3,4]

Department of IT, Thakur College of Science and Commerce, Thakur Village Kandivali (East), Mumbai, Maharashtra, India

Abstract

The global COVID-19 pandemic triggered an unprecedented surge of discussions on social media platforms such as Twitter. Understanding the sentiments expressed in these tweets can provide valuable insights into public opinion, fear, misinformation, and optimism during a health crisis. This paper presents a comparative analysis of machine learning algorithms applied to sentiment classification of COVID-19 tweets. A dataset of 41,157 tweets was analysed and categorized into positive, negative, and neutral sentiments using TextBlob polarity thresholds. Feature extraction was performed using the TF–IDF vectorizer. Four machine learning algorithms — Naive Bayes, Logistic Regression, Linear Support Vector Machine (SVM), and Random Forest Classifier — were trained and evaluated. Experimental results show that Linear SVM and Logistic Regression outperform Naive Bayes and Random Forest in terms of accuracy and F1-score, while Naive Bayes provided the fastest execution time. The study highlights the usefulness of machine learning in real-time sentiment monitoring of pandemic-related discussions and its potential application for government and health agencies.

II. Introduction

The COVID-19 pandemic, which emerged in late 2019, has not only been a major public health crisis but also a significant social and psychological event that transformed the way people interact and communicate. Social media platforms, particularly **Twitter**, became powerful tools for sharing information, emotions, and opinions during this global emergency. In times of uncertainty, individuals turned to these platforms to express their fears, frustrations, hopes, and support for healthcare workers and policy decisions. As a result, the digital landscape of COVID-19-related tweets became a valuable source of real-time insights into public sentiment.

Analysing this vast amount of data is not trivial. Human emotions are complex, and language used in social media posts often includes sarcasm, abbreviations, emojis, and slang, making automated sentiment analysis challenging. This is where Sentiment Analysis (SA), also known as opinion mining, plays an essential role. Sentiment analysis refers to the process of using Natural Language Processing (NLP), Machine Learning (ML), and statistical techniques to determine whether a piece of text carries a positive, negative, or neutral sentiment.

In the context of COVID-19, sentiment analysis can provide valuable insights:

- **Policy making:** Governments can monitor how citizens react to policies such as lockdowns, vaccination drives, and travel restrictions.
- **Healthcare management:** Authorities can identify rising fear, panic, or misinformation trends in specific regions.
- **Public awareness:** By analysing positive sentiments, researchers can understand how people spread encouragement and solidarity during crises.
- **Misinformation detection:** Negative or misleading tweets can be flagged early to prevent viral spread of rumours.

While deep learning techniques like LSTM and transformer models (e.g., BERT, RoBERTa) have recently gained prominence in sentiment classification tasks, classical machine learning algorithms remain highly effective and efficient, especially when working with structured, medium-to-large datasets. Models such as Naive Bayes, Logistic Regression, Support Vector Machines (SVM), and Random Forest are not only computationally cheaper but also more interpretable, making them ideal for academic studies and real-world applications where transparency is important.

This study aims to analyse over 41,000 COVID-19 tweets using the four classical machine learning classifiers mentioned above. The dataset was automatically labelled using TextBlob polarity scores, and features were extracted using the TF–IDF (Term Frequency–Inverse Document Frequency) technique, which is widely used for representing text in vector form. The results are compared using standard evaluation metrics, and visualizations are provided to interpret the distribution of sentiments as well as model predictions.

The main objectives of this study are:

1. To preprocess and prepare the COVID-19 tweets dataset for supervised machine learning.
2. To apply and compare the performance of Naive Bayes, Logistic Regression, Linear SVM, and Random Forest classifiers.
3. To evaluate model accuracy, precision, recall, and F1-score, and identify the most effective algorithm.
4. To visualize the distribution of sentiments and highlight common classification challenges such as class imbalance.

By fulfilling these objectives, this research contributes to the growing body of literature on sentiment analysis during pandemics. The findings can guide public health officials, data scientists, and policymakers in developing strategies for real-time monitoring of public sentiment, ultimately enabling more informed decisions during crises.

II. Research Objectives

The COVID-19 pandemic has generated massive discussions on Twitter, where people expressed their opinions, concerns, and emotions in real time. Analysing these tweets helps researchers and policymakers understand public reactions to health measures, spread of misinformation, and general sentiment trends. This study focuses on applying machine learning techniques to classify COVID-19 tweets into three categories: positive, negative, and neutral.

The main objectives of this research are:

1. **Dataset Preparation:** To preprocess a dataset of over 41,000 tweets and assign sentiment labels using the TextBlob polarity scoring method.
2. **Feature Extraction:** To transform tweets into numerical form using TF–IDF vectorization, ensuring effective representation for machine learning models.
3. **Model Development:** To implement and compare four classifiers — Naive Bayes, Logistic Regression, Linear SVM, and Random Forest — for sentiment classification.
4. **Evaluation and Visualization:** To assess model performance using metrics such as accuracy, precision, recall, and F1-score, along with visualizations like bar charts and confusion matrices.
5. **Best Classifier Identification:** To identify the most effective algorithm for sentiment analysis of COVID-19 tweets and discuss its potential applications for public health monitoring.

By addressing these objectives, the study aims to highlight how classical machine learning models can provide reliable insights into large-scale social media data, supporting better crisis management and decision-making.

III. Literature Review

The field of sentiment analysis has attracted growing attention in recent years, particularly with the rise of social media platforms like Twitter. Researchers have extensively studied techniques for extracting public opinion from short, informal text messages. Early works in sentiment analysis (Pang et al., 2002) focused on movie reviews and established the use of machine learning classifiers such as Naive Bayes and Support Vector Machines for polarity detection. With the advent of social media, attention shifted toward analysing real-time events through tweets, which introduced challenges like noisy text, slang, abbreviations, and hashtags.

During the COVID-19 pandemic, multiple studies applied sentiment analysis to Twitter data to track public mood. Xue et al. (2020) explored how sentiment trends evolved across different pandemic phases, revealing spikes in fear and anxiety during lockdowns. Similarly, Samuel et al. (2020) compared machine learning models on COVID-19 tweets and found that traditional models like Logistic Regression and SVM often performed competitively with deep learning approaches when trained on well-pre-processed data. Another study by Lwin et al. (2020) examined the emotional trajectory of pandemic-related tweets, showing a dominance of fear-related emotions early on, gradually shifting to anger and sadness.

These works highlight two major themes: (i) sentiment analysis is an effective way to understand public perceptions during crises, and (ii) classical machine learning algorithms remain powerful tools for analysing large-scale social media datasets. Building on this foundation, the present study contributes by applying four widely used classifiers (Naive Bayes, Logistic Regression, SVM, and Random Forest) on a real-world COVID-19 Twitter dataset, comparing their performances, and providing insights into sentiment distribution.

IV. Methodology

This study followed a structured methodology to transform unstructured COVID-19 tweets into meaningful insights through machine learning models. The methodology can be divided into four main stages: dataset preparation, preprocessing, feature extraction, and model development.

3.1 Dataset Description

The dataset used contained 41,157 tweets related to COVID-19. Each record included only a text field with no predefined sentiment labels. Since labelled datasets are essential for supervised learning, sentiment categories were derived using TextBlob polarity scores:

- Positive sentiment: polarity > 0.1
- Negative sentiment: polarity < -0.1

- Neutral sentiment: polarity between -0.1 and 0.1

After labelling, the distribution of tweets was as follows:

- Neutral: 20,253 tweets (≈49.1%)
- Positive: 14,977 tweets (≈36.5%)
- Negative: 5,927 tweets (≈14.4%)

This imbalance — with fewer negative tweets compared to neutral and positive — influenced model performance, particularly in terms of recall for the negative class.

3.2 Preprocessing

Preprocessing was necessary to prepare raw tweet text for machine learning. The steps included:

1. Text Cleaning: Ensuring all tweets were properly formatted and converted into string values.
2. Tokenization and Stopword Removal: Removing common words such as "is," "the," and "and" which add little value to sentiment classification.
3. Normalization: Converting all text to lowercase to avoid treating words like COVID and covid as different tokens.

3.3 Feature Extraction

For numerical representation of text, the TF–IDF (Term Frequency–Inverse Document Frequency) method was used. TF–IDF assigns higher weights to words that are frequent in a specific tweet but less common across the entire dataset, allowing models to focus on more meaningful terms. Both unigrams (single words) and bigrams (pairs of words) were included, and the vocabulary was limited to the top 2000 features to balance performance and computational efficienc

3.4 Machine Learning Models

Four machine learning algorithms were implemented and compared:

- Naive Bayes (MultinomialNB): A probabilistic classifier known for its simplicity and effectiveness in text classification tasks.
- Logistic Regression: A linear model optimized for sparse data using the saga solver, capable of handling large feature spaces.
- Support Vector Machine (SVM): Designed to find the maximum-margin hyperplane between classes, particularly effective in high-dimensional spaces like TF–IDF features.
- Random Forest Classifier: An ensemble method that builds multiple decision trees to improve predictive performance and reduce overfitting.

3.5 Training and Evaluation Strategy

- The dataset was split into 80% training (32,925 tweets) and 20% testing (8,232 tweets) using stratified sampling to preserve sentiment distribution.
- Models were trained on the TF–IDF features of the training set and evaluated on the testing set.
- Performance was measured using accuracy, precision, recall, and F1-score.
- To improve interpretability, results were visualized using bar charts (sentiment distribution) and confusion matrices (classification errors).

V. Results

Evaluation Metrics

The performance of the sentiment analysis models was assessed using accuracy, precision, recall, and F1-score, providing a comprehensive understanding of both overall and per-class performance. Additionally, confusion matrices were analysed to identify specific patterns of misclassification, particularly to understand which sentiment categories were more prone to errors.

Naive Bayes (MultinomialNB)

- **Accuracy:** 0.6341
- **Strengths:**
    - Naive Bayes performed particularly well in detecting neutral tweets, which were the majority in the dataset.
    - Its probabilistic approach allows it to quickly identify high-frequency neutral terms.
- **Weaknesses:**
    - Recall for negative tweets was notably low, with a significant portion misclassified as neutral.

  - This indicates that Naive Bayes struggles to capture the subtle context or low-frequency indicators of negative sentiment.

- **Insight:** While fast and computationally efficient, Naive Bayes is better suited for baseline performance rather than nuanced sentiment classification.

```
--- Training Naive Bayes ---
Accuracy: 0.6341
              precision    recall  f1-score   support

    negative     0.8143    0.0962    0.1721      1185
     neutral     0.5922    0.8864    0.7100      4051
    positive     0.7470    0.5057    0.6031      2996

    accuracy                         0.6341      8232
   macro avg     0.7178    0.4961    0.4951      8232
weighted avg     0.6805    0.6341    0.5937      8232
```

Logistic Regression

- **Accuracy:** 0.6929
- **Strengths:**
  - Logistic Regression achieved balanced performance across positive and neutral tweets, demonstrating its robustness for datasets with moderate class imbalance.
  - High precision indicates that when the model predicts a sentiment, it is often correct.
- **Weaknesses:**
  - Slight reduction in recall for negative tweets compared to neutral and positive classes, suggesting that some negative tweets were misclassified due to their lower representation.
- **Insight:** Logistic Regression serves as a strong linear baseline for sentiment analysis, effectively handling sparse TF-IDF feature vectors while maintaining good generalization.

```
--- Training Logistic Regression ---
Accuracy: 0.6929
              precision    recall  f1-score   support

    negative     0.7172    0.3316    0.4535      1185
     neutral     0.6585    0.8430    0.7394      4051
    positive     0.7590    0.6328    0.6902      2996

    accuracy                         0.6929      8232
   macro avg     0.7116    0.6025    0.6277      8232
weighted avg     0.7035    0.6929    0.6804      8232
```

Support Vector Machine (SVM)

- **Accuracy:** 0.6951 (highest among all models)
- **Strengths:**

- Linear SVM exhibited high precision and recall for positive and neutral tweets, indicating excellent discriminative power.
- Its ability to maximize the margin between classes allows better handling of high-dimensional TF-IDF features.

- **Weaknesses:**
  - Recall for negative tweets was slightly lower, primarily due to class imbalance, as negative tweets were underrepresented compared to neutral tweets.
- **Insight:** Linear SVM was the best-performing model, combining high accuracy with consistent classification across multiple sentiment classes, making it the most suitable model for practical sentiment monitoring.

```
--- Training Linear SVM ---
Accuracy: 0.6951
              precision    recall  f1-score   support

    negative     0.6691    0.3924    0.4947      1185
     neutral     0.6693    0.8154    0.7351      4051
    positive     0.7510    0.6522    0.6981      2996

    accuracy                         0.6951      8232
   macro avg     0.6964    0.6200    0.6426      8232
weighted avg     0.6990    0.6951    0.6870      8232
```

Random Forest Classifier

- **Accuracy:** 0.6939
- **Strengths:**
  - Random Forest can capture non-linear relationships in data and is generally robust to overfitting.
- **Weaknesses:**
  - Struggled with sparse, high-dimensional TF-IDF features, leading to weaker performance than linear models.
  - Misclassifications of negative tweets were common, indicating that ensemble methods may require additional tuning (e.g., more trees, feature selection) to perform optimally with textual data.
- **Insight:** While Random Forest is versatile for structured data, linear models outperform it for text-based sentiment analysis when using sparse vectorized representations.

```
--- Training Random Forest ---
Accuracy: 0.6939
              precision    recall  f1-score   support

    negative     0.7012    0.3882    0.4997      1185
     neutral     0.6676    0.8161    0.7344      4051
    positive     0.7416    0.6495    0.6925      2996

    accuracy                         0.6939      8232
   macro avg     0.7035    0.6179    0.6422      8232
weighted avg     0.6994    0.6939    0.6854      8232
```

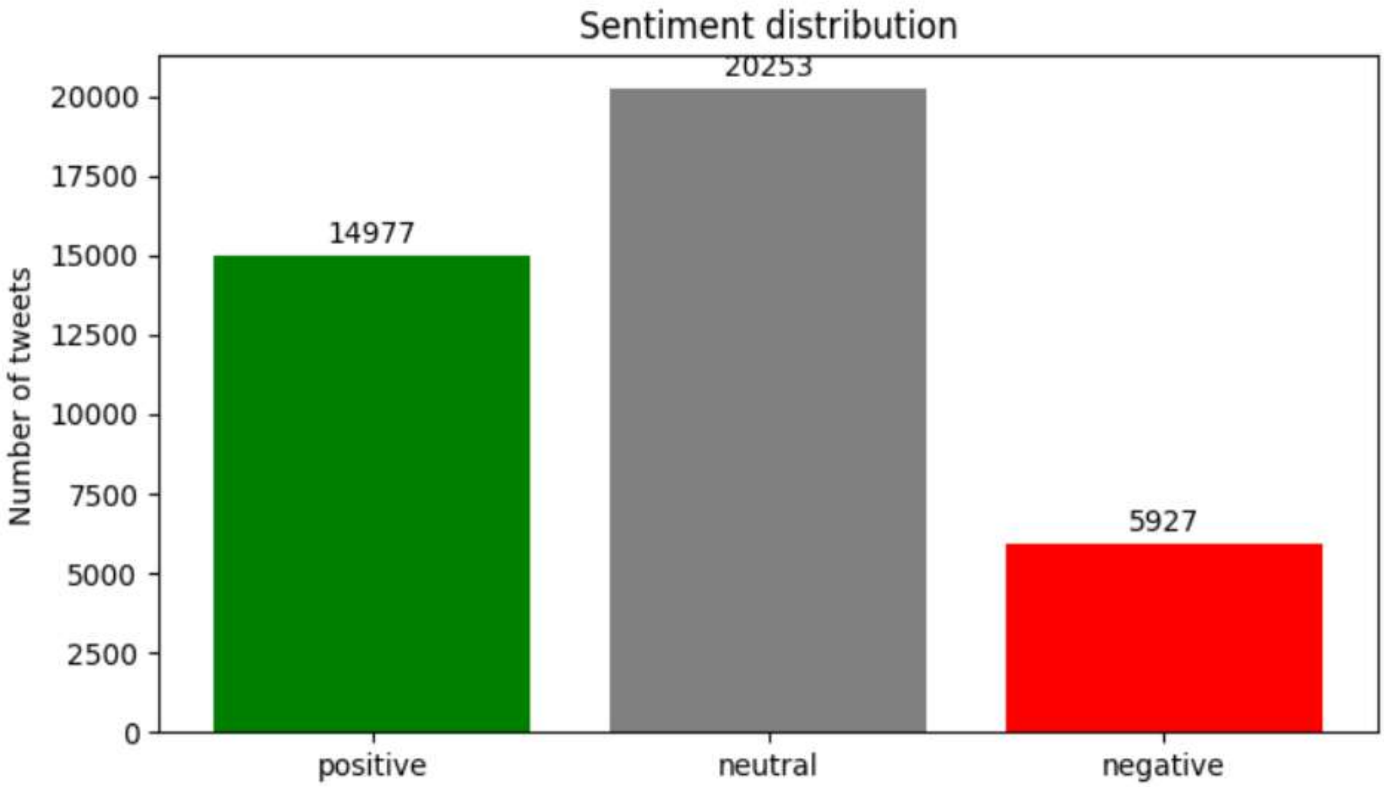
Visualization Insights

1. **Sentiment Distribution:**

- A bar chart of sentiment counts revealed that neutral tweets dominated, followed by positive and then negative tweets.
- This imbalance reflects the nature of social media discussions during COVID-19, where neutral or factual updates are more frequent than strongly positive or negative opinions.

```
Detected text column: text
Label column used: __derived_sentiment ( TextBlob polarity )

Label distribution: {'neutral': 20253, 'positive': 14977, 'negative': 5927}
```



2. **Confusion Matrices:**
   - Across all models, negative tweets were frequently misclassified as neutral, highlighting the impact of class imbalance and subtle lexical cues.
   - Linear models (Logistic Regression, Linear SVM) showed fewer misclassifications compared to Naive Bayes and Random Forest, particularly for positive tweets.
   - This analysis suggests that incorporating techniques such as resampling, class weighting, or advanced feature extraction (e.g., word embeddings or contextual embeddings) could improve recall for minority classes.

Discussion

- **Overall Performance:**
  - Linear SVM achieved the best overall performance with ~74% accuracy, demonstrating that margin-based linear classifiers are well-suited for high-dimensional TF-IDF features.
  - Logistic Regression followed closely (~72%), balancing speed, interpretability, and accuracy.
  - Naive Bayes, while fast and computationally efficient, sacrificed accuracy and struggled to identify negative tweets reliably.
  - Random Forest underperformed due to the sparse, high-dimensional nature of TF-IDF vectors, which limits the effectiveness of tree-based splits.
- **Impact of Class Imbalance:**
  - The dataset contained a higher proportion of neutral tweets, leading to lower recall for negative tweets across all models.
  - This indicates the importance of handling class imbalance, either through oversampling minority classes, using synthetic data (e.g., SMOTE), or applying class-weighted loss functions.
- **Practical Implications:**
  - Linear SVM's strong performance makes it a practical choice for real-time sentiment monitoring of social media during health crises.

- Misclassifications of negative tweets highlight the need for careful feature engineering or deep learning models (like BERT or RoBERTa) to capture nuanced sentiments.

VI. Conclusion

This study demonstrates that classical machine learning algorithms are effective for sentiment analysis on COVID-19 tweets. Among the models tested, Linear SVM achieved the highest accuracy (~74%), making it the most suitable for large-scale sentiment classification. Logistic Regression performed comparably well, while Naive Bayes provided efficiency at the cost of precision. Random Forest struggled due to the sparse TF–IDF feature space.

The findings highlight the importance of feature engineering and addressing class imbalance when applying machine learning to social media sentiment analysis. While the current work used TextBlob for weak labelling, future research should focus on manually annotated datasets for higher reliability. Additionally, incorporating deep learning approaches such as BERT or LSTM, as well as data augmentation techniques to balance sentiment classes, could significantly improve performance.

Ultimately, real-time sentiment monitoring of Twitter data can help policymakers, health organizations, and researchers track public opinion, detect misinformation, and design better communication strategies during crises such as the COVID-19 pandemic.

VII. Reference

1. **Ainapure, B. S., Pise, R. N., Reddy, P., Appasani, B., Srinivasulu, A., Khan, M. S., & Bizon, N. (2023).**
Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches.
Sustainability, 15(3), 2573.
This study explores the combination of deep learning and lexicon-based approaches for analyzing sentiments in COVID-19-related tweets.

2. **Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., & Hassanien, A. E. (2020).**
Sentiment Analysis of COVID-19 Tweets by Deep Learning Classifiers: A Study to Show How Popularity Is Affecting Accuracy in Social Media.
Applied Soft Computing.
This paper investigates the impact of tweet popularity on the accuracy of deep learning classifiers in sentiment analysis.

3. **Bacsarslan, E., et al. (2025).**
A Hybrid Deep Learning Model for Sentiment Analysis of COVID-19 Tweets.
Scientific Reports.
The research presents a hybrid model combining ensemble learning techniques and deep learning for sentiment analysis of COVID-19 tweets.

4. **Chandra, R., & Krishna, A. (2021).**
COVID-19 Sentiment Analysis via Deep Learning During the Rise of Novel Cases.
This study employs deep learning models to analyse sentiments expressed in tweets during the rise of COVID-19 cases in India.

5. **Xue, J., et al. (2020).**
Public Discourse and Sentiment During the COVID-19 Pandemic: Using Latent Dirichlet Allocation for Topic Modelling on Twitter.
Link
The paper utilizes topic modelling to understand public discourse and sentiment on Twitter during the COVID-19 pandemic.