Kalpa Publications in Computing
Volume 19, 2024, Pages 144–156
Proceedings of 6th International Conference on Smart Systems and Inventive Technology

Exploring the Effectiveness of Various Machine LearningModels in Analyzing Sentiment from Twitter Data

Jishnu Prakash K[1], Liz Maria Liyons[2], Anaswer Ajay[3], Akshaj Vadakkath Joshy[4], and Sarath S[5]

Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham,
Amritapuri, India jishnuprakash668@gmail.com[1], lizmarialiyons@gmail.com[2], anaswer.impluse@gmail.com[3], akshajvj2003@gmail.com[4], saraths@am.amrita.edu[5]

Abstract. This research assesses sentiment analysis on Twitter posts about Twitter tweets, comparing machine learning techniques. Four different models are used to classify and their efficacy is evaluated. Addressing balanced datasets, SVM excels in balanced sentiment scenarios. Using metrics like accuracy and recall, the study offersinsights for decision-making in marketing and social studies. Emphasizing machine learning effectiveness, the research suggests improvements for sentiment analysis in diverse domains, particularly in understanding positive and negative Twitter tweets.

Keywords: Twitter Tweets Classification · Sentiment Analytics · Machine Learning Methods · Comparative Analysis.

1 Introduction

The advent of social media has transformed communication, offering a powerful platform for individuals to articulate opinions, exchange experiences, and to participate in having meaningful conversations. Among the myriad of platforms, twitter stands out as a space which is dynamic and is able to capture the thoughts and sentiments of the masses with about 280-character messages. The sentimental analysis of the various different twitter postshas emerged as the most pivotal research area, unveiling important insights into the attitude as well as the collective mood of the users. [7] The research paper takes on a comprehensive journey of sentimental analysis applied to different twitter posts. The research paper study is divided into three phases, starting with detailed pre-processing to address the unique and important characteristics of Twitter language. Following that, a carefully designed and curated feature vector that includes relevant aspects of tweets serves as the foundation of the analysis. Finally we used different classifying models to classify tweets to negative and positive class. This research aims to compare how different machine learning techniques understand sentiments in Twitter discussions using two different datasets: balanced and unbalanced datasets, where the distribution of positive and negative classes differ. We're looking at how specific information about a topic affects the features used in the analysis. This includes dealing with challenges like the brief nature of tweets and the use of emoticons, slang, and misspellings. By creating and studying a dataset tailored to this topic, we want to understand how different classifiers impact sentiment analysis.

As studies turn to perceptual classification and use a variety of classifiers, including logistic regression, Multinomial Naive Bayes, Support Vector Machine(SVM), and Gaussian Naive Bayes, and we tested our models using parameters like accuracy, recall, and F1 scores. This study analyzes several models making use of pre-processed datasets and a range of evaluation criteria in order to determine the extent to which different machine learning models execute when analyzing sentiment from Twitter data. The present research includes an in-depth examination of studies that use four supervised machine learning methods to examine sentiment on the two distinct Twitter datasets. The intent is to assess which sentiment analysis model is the most productive, offering valuable details on the best algorithms for dependable and precise sentiment classification from Twitter data. The aim is to strengthen sentiment analysis in the fast-paced field of social media analytics for dynamic user emotions.

2 Related Work

Many researches on Sentimental Analysis [10] have been conducted in past year. An unsupervised learning system for review categorization based on semantic orientation is presented in Paper [3]. By examining the average emotion of sentences that contain adjectives or adverbs, it makes predictions about whether or not a review is recommended. With the use of mutual information and both positive and negative reference terms, the algorithm determines semantic orientation. Paper[8] explores the optimization of K Nearest Neighbour for sentiment analysis, with a focus on improving performance overall and classification accuracy.

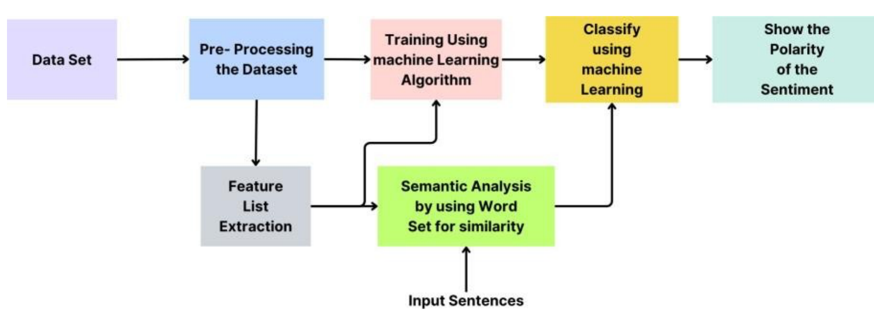

Fig. 1. Architecture

The increasing usage of social networking sites and the value of sentiment analysis in understanding public attitudes expressed on a range of topics via these platforms are examined in Paper [12]. The review explores a variety of sentiment analysis methodologies, including learning-based and lexicon-based strategies. The study, which focuses on data from Twitter, identifies particular problems and obstacles that need to be resolved in order to successfully apply these strategies. Comprehending the attitude on social media sites such as Twitter is essential for assessing public opinion on certain subjects of interest, and this paper provides an invaluable summary of approaches and difficulties in this regard. The effect of text pre-processing on sentiment analysis on Twitter is examined in this paper [13]. Classification performance is greatly enhanced by extending acronyms and substituting negations, particularly when using Naive Bayes classifiers. We have two different approaches to deal with sentimental analysis of text. That is Symbolic and Machine learning methods. The below given section explains the two methods mentioned before.

2.1 Method 1 - Symbolic Techniques

Symbolic techniques in unsupervised sentiment classification leverage existing lexical resources to analyze and interpret textual data. These approaches often involve the utilization of predefined dictionaries, semantic networks, and linguistic rules to identify sentiment-bearing words and phrases. By relying on symbolic representations, researchers aim to capture the intricate nuances of sentiment in a language, allowing for a more nuancedunderstanding of sentiment within the text. However, these techniques may face challenges in handling context-dependent sentiments and may require continuous refinement to adaptto evolving language patterns.

2.2 Method 2 - Machine Learning techniques

We will give the training and testing set for machine learning classification. In training set, wewill give sample feature and its corresponding class. The machine learning model learns from this set to understand patterns and relationships. Once trained, we test the model's knowledge using the data test set that is not known before to the model.[5] This gives us feedback on how accurately the model can generalize its learning to new, unseen data, by understanding the underlying patterns.

3 SUGGESTED APPROACH

As shown in Figure 1, we have collected twitter tweets for twitter using Twitter API client to CSV file. Generally the tweet text will be short or misspelled word and even text can sometime use slang. Then, we introduce three steps in order to classify texts based on sentiments. [9] The initial step is to make data ready for processing which is termed as data pre-processing. Then using the important characteristic in dataset we created vector offeature. The third and last step will be the classification of tweet text to its corresponding class using machine learning classifiers.

3.1 Objective

The main objective of this research paper is to compare the efficacy of various machine learning techniques in analysing the sentiment of Twitter posts. The study aims to investigate the impact of domain specific information on feature vector selection, considering challenges posed by short length tweets, emoticons, slang, and misspellings. By creating and analyzing a dataset specific to this research seeks to identify the influence of different classifiers on sentiment classification performance within this domain. The proposed two- phase feature extraction approach aims to enhance the efficiency of sentiment analysis for Twitter content, contributing valuable insights for future applications in social media analytics. This study uses a two-phase feature extraction strategy to solve special issues in Twitter sentiment analysis. Firstly, general features are extracted to identify common sentiments and language patterns, which sets the groundwork for deciphering wider sentiment trends in tweets. Phase two concentrates on domain-specific data, including subtleties unique to the Twitter language such as hashtags, emojis, and user mentions. This focused approach improves sentiment analysis by taking into account the intricacies of social media conversation, guaranteeing accurate sentiment interpretation in the fluid and concise form of Twitter material.

1.1 Data collection

The data for our dataset is collected from twitter using API client of twitter and is stored as csv files and fed into workspace using pandas library of python. Collection of these data from Twitter is arranged in csv with proper column name. The dataset used for training the model has a target attribute which helps in understanding the emotion of the respective.

1.2 Pre-processing of Tweet text

To avoid misspelling and slang usage in text, the pre-processing step is done before feature extraction.[6] The Python NLTK (Natural Language Toolkit) library contains the most popular text preparation methods. We also use the regex module of python to perform certain important pre-processing steps. For stemming we used the porter stemmer algorithm. The only way to improve quality of data is through data pre-processing. The pre- processing is a main step as the pre-processed text is the input to the further process thatshould be done for the completion of the prediction.

1.3 Classification of Text

The classification is performed using Logistical Regression, Multinomial Naive Bayes, Support Vector Machine(SVM) and Gaussian Naive Bayes. After comparing the performance a detailed observation is given below.

2 METHODS FOR CLASSIFICATION

Different varieties of classification algorithm are used in order to classify the tweet text to its corresponding class based on sentiment. We chose Support Vector Machines (SVM), Naive Bayes, and Logistic Regression for our sentiment analysis on Twitter. For binary and multi-class classification, logistic regression functions well, providing meaningful coefficients for feature relevance and acceptability for sparse twitter data.

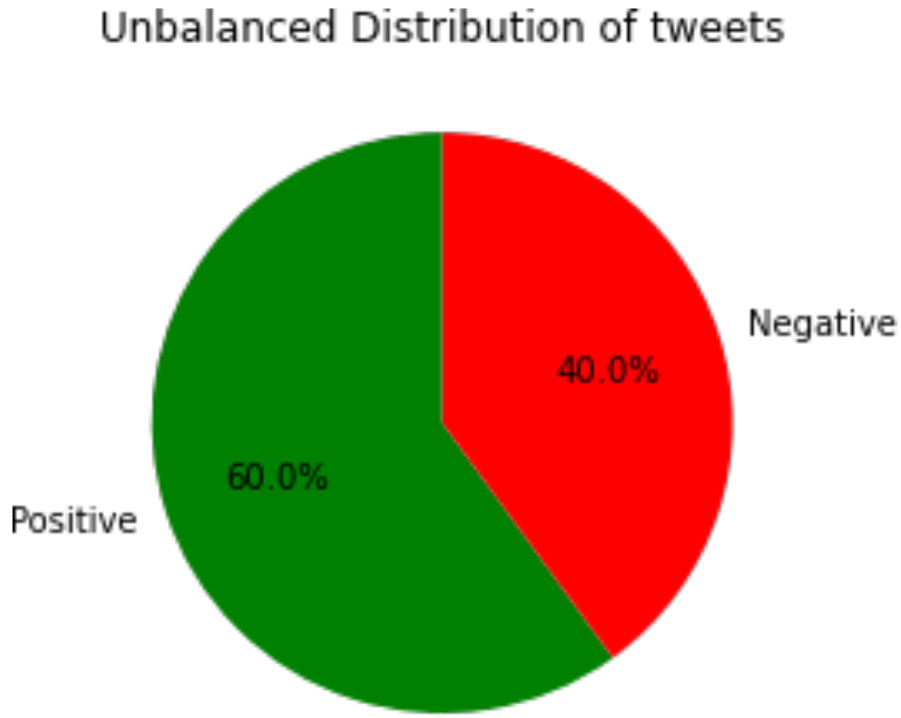

Fig. 2. Distribution of positive and negative samples

Large Twitter datasets are readily handled by Naive Bayes, which performs outstandingly with categorical characteristics like hashtags and emoticons and delivers meaningful probabilistic interpretations. Because of its resilience to noise and outliers, support vector machines (SVM) are a good fit for the high-dimensional data used in sentiment analysis. Theyare especially useful for correcting imbalances in sentiment classes. Real-time analysis of Twitter streams is made better by the interpretability of Logistic Regression and Naive Bayes models alongside their handling efficiency. Further enhancing the overall productivity isSVM's the capacity for high accuracy, particularly with suitable parameter adjustments. A complete examination of Twitter sentiment analysis will be guaranteed by a broad range of classifiers that allow for vast comparison, benchmarking, and analysis of tasks that pertain to the topic at hand.

3 EVALUATION

We use accuracy and f1 score to determine the consistency and application of algorithms to particular model. The Eq 1 refers to the accuracy formula and Eq 2 depicts recall formula. We can find F1 score which is derived from accuracy and recall and is depicted in the Eq 3. From TABLE I and TABLE II, we can see the performance metrics for models in the case of both balanced as well unbalanced datasets. To provide a further detailed explanation about the metrics, we are looking at both datasets as a component of our research into sentiment analysis on Twitter data.

Model	Accuracy	Precision	Recall	F1-Score
Gaussian Naive Bayes	0.597500	0.627097	0.596271	0.570952
Logistic Regression	0.745000	0.745185	0.745000	0.744952
Multinomial Naïve Bayes	0.758125	0.770389	0.767891	0.767529
Support Vector Machine	0.768125	0.768124	0.768110	0.768114

Table 1. Performance Metrics for Models in Unbalanced Dataset

BALANCED DATASET:

As demonstrated in Figure 4, the evaluation of multiple machine learning models on the balanced dataset (referred to as the "Balanced Dataset" in our work) suggests differing degrees of accuracy. With an AUC value of 0.77, SVM emerges triumphant indicating its efficacy in recognizing crucial patterns. With an AUC score of 0.75, multinomial naive bayes come in close second, showing its adeptness at handling the evenly distributed positive and negative tweets. While logistic regression performs excellently in sentiment analysis, exhibiting an AUC score of 0.74, Gaussian Naive Bayes only obtains a score of 0.6. A balanced representation of sentiments in the dataset is displayed in Figure 3 using the balancedTwitter sentiment analysis, which exhibits the equal amount of positive and negative distribution of Twitter tweet samples. The balanced dataset's confusion matrix can be observed in Figure 7, and it is clear from this data that SVM achieves acceptable levels of recall, accuracy, precision, and F1-Score. This illustrates how well the model accommodates abalanced distribution of positive and negative feelings. Outstanding outcomes are demonstrated as well by Gaussian Naive Bayes, Multinomial Naive Bayes, and Logistic Regression, which help to provide an extensive understanding of the sentiment analysis properties of the balanced dataset. Figure 8's performance study underlines theseconclusions even more by contrasting the various machine learning models applied to the balanced dataset employing measures for accuracy, precision, recall, and F1-Score among theseveral machine learning models that were implemented on the well-balanced dataset. SVM shines out as being particularly useful in the situation.

Table 2. Performance Metrics for Models in Balanced Dataset

Model	Accuracy	Precision	Recall	F1-Score
Gaussian Naive Bayes	0.597500	0.627097	0.596271	0.570952
Logistic Regression	0.745000	0.745185	0.745000	0.744952
Multinomial Naive Bayes	0.758125	0.770389	0.767891	0.767529
Support Vector Machine	0.768125	0.768124	0.768110	0.768114

UNBALANCED DATASET:

A comprehensive examination of the unbalanced dataset which can be seen by a scenario where positive tweets outnumber negative tweets in Twitter data offered by the API is presented in Figure 2. Also, referring to the Figure 5 which shows the ROC curve, it is clear that the AUC ratings of the multiple machine learning models provide insights into the degree to which they operate. AUC Ratings: Using logistic regression, tweets are capable of being discovered with an AUC of 0.55. SVM does remarkably well at classifying thoughts in an asymmetrical distribution, with an amazing AUC of 0.63. Despite Multinomial Naive Bayes succeeds smoothly in multidimensional sentiment analysis (AUC of 0.60), Gaussian Naive Bayes delivers decent sentiment prediction (AUC of 0.56 in the imbalanced dataset). A more thorough look at the confusion matrix, a key instrument for obtaining model performance measures, can be found in Figure 6. The evaluation criteria provide a detailed insight of each model's effectiveness and include accuracy, precision, recall, and F1-score. Performance study: Having reference to Figure 9, this study demonstrates howthe four machine learning models that were applied on the unbalanced dataset have been assessed in terms of accuracy, precision, recall, and F1-score. Ultimately, the evaluation of the data set with imbalances reveals how various machine learning models perform effectively when managing situations when the vast majority of tweets are positive.

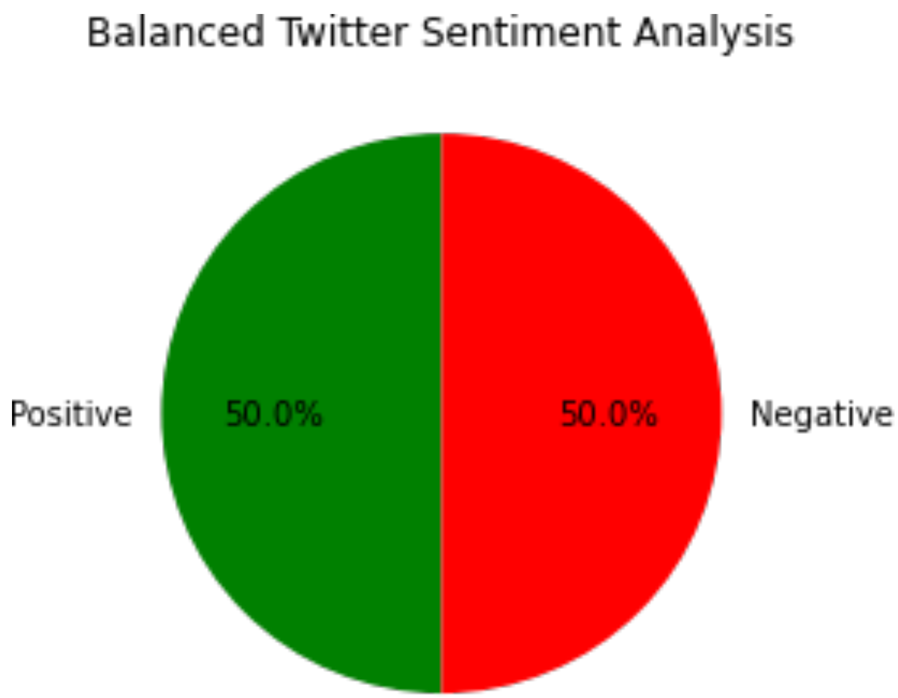

Fig. 3. Distribution of positive and negative samples

As an outcome, our sentiment analysis on the balanced Twitter dataset shows the Support Vector Machine (SVM) model's superior performance. SVM emerges up as a particularly effective model with an AUC score of 0.77 and demonstrated success in handling a balanced distribution of positive and negative tweets. The sentimental analysis's equal positive and negative tweets division adds further proof to the robustness of the model. SVM is the preferred choice for dependable sentiment analysis in a balanced dataset situation owing to the accompanying confusion matrix and performance analysis, which further supports the model's superiority over other approaches.

Accuracy = (TN+FN)/(TN+FN+TP+FP) □ (Eq 1)
Recall = TP/(TP+FN) □(Eq 2)
F1 Score= (2 x Precision x Recall)/(Precision + Recall) □ (Eq 3)

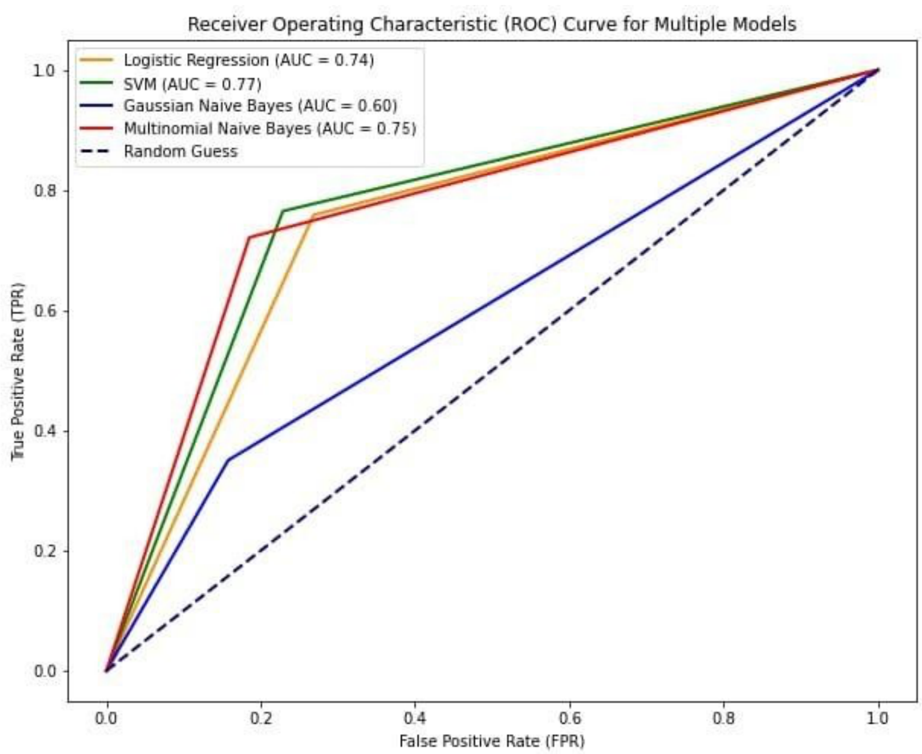

Fig. 4. ROC Curve for different models in Balanced Dataset

4 MODEL SELECTION

From graph [Figure 4] reveals that the Support Vector Machine model outperform other models for the balanced dataset. Support Vector Machines (SVMs) are often more effective for sentiment analysis on datasets with an equal distribution of instances across different sentiment classes. [4] This effectiveness is attributed to SVMs' capability to identify a hyper plane that maximizes the separation between classes. In a balanced dataset, where positive and negative sentiments are equally represented, SVMs are better in finding a distinct boundary. The key mechanism of SVM is determining a hyper plane that can maximizesthe margin between data points of different classes.

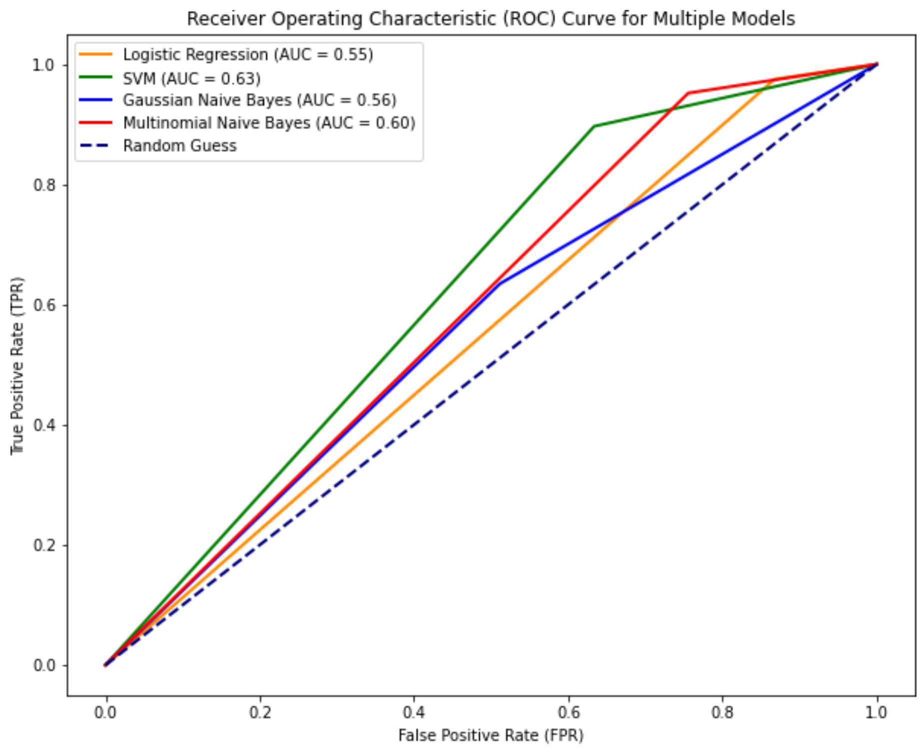

Fig. 5. ROC Curve for different models in Unbalanced Dataset

In balanced datasets, the equal representation of sentiments facilitates the SVM in achieving a clear separation. Challenges arise in imbalanced datasets, where one sentiment class significantly outweighs the other. [2] In such cases, SVMs may exhibit bias towards the majority class, aiming to minimize classification error by predominantly assigning instances to the majority class. It's important to consider that while SVMs are well-suited for balanced datasets in sentiment analysis, the choice of the machine learning algorithm should take into account various factors, including dataset size, nature, feature representation, and specific characteristics of the sentiment analysis task.

5 CONCLUSION

A number of techniques, including symbolic and machine learning approaches, are available to identify emotions in text. Symbolic approaches are usually thought of as being less efficient and more complex than machine learning methods. Notably, these strategies have been advantageous for the area of sentiment analysis on Twitter. It's uncommon to run into challenges sorting through a large number of keywords in tweets to find the ones that are emotionally charged. Moreover, handling misspellings and informal speech introduces more complexities. In order to address these issues, a two-step feature extraction procedure along with optimal pre-processing is used to develop a robust feature vector.

First, distinctive features connected to Twitter are extracted and incorporated into the feature vector. These Twitter-specific attributes are subsequently eliminated from the tweets, allowing room for a second phase of feature extraction that is comparable to the oneused in regular text analysis. The combined set of these features obtained is the final feature vector. For assessing the classification accuracy of the feature vector, a number of classifiers such as Logistic Regression, Multinomial Naive Bayes, Support Vector Machine, and Gaussian Naive Bayes have been used. These classifiers are crucial in determining how effectively the feature vector functions in Twitter sentiment analysis. Among the four models evaluated, Support Vector Machine (SVM) is the top performing technique, showcasing the highest predictive accuracy. SVM model had demonstrated exceptional promise in classifying the tweets into corresponding classes. The research provides highly valuable insights into choosing the most fitting predictive models for understanding and predicting sentiments expressed on the Twitter platform. The collaborative use of Support Vector Machine (SVM) demonstrates the potential to increase our comprehension of public sentiment, enabling more efficient decision-making in various domains.

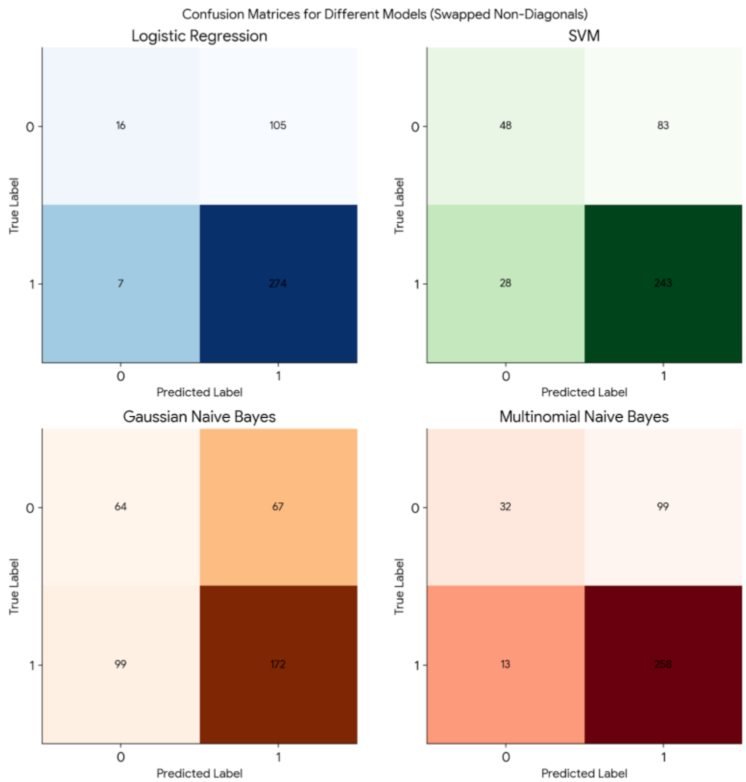

Fig. 6. Confusion Matrix of Unbalanced Dataset

The findings of the research can be applied in real-world scenarios. By evaluating how various machine learning models perform in sentiment analysis on Twitter, we may be enabled to create applications that automatically detect and assess sentiments in large amounts of data.Leveraging the most of some models' comprehensibility, computational effectiveness, and ability to cope with skewed datasets could enhance the efficacy and efficiency of sentiment analysis techniques. This type of data is valuable to developers and businesses arranging to use sentiment analysis on Twitter as it allows them to select models which fulfill their individual goals and requirements. This increases the precision and dependability of classification of sentiment in real-world scenarios.

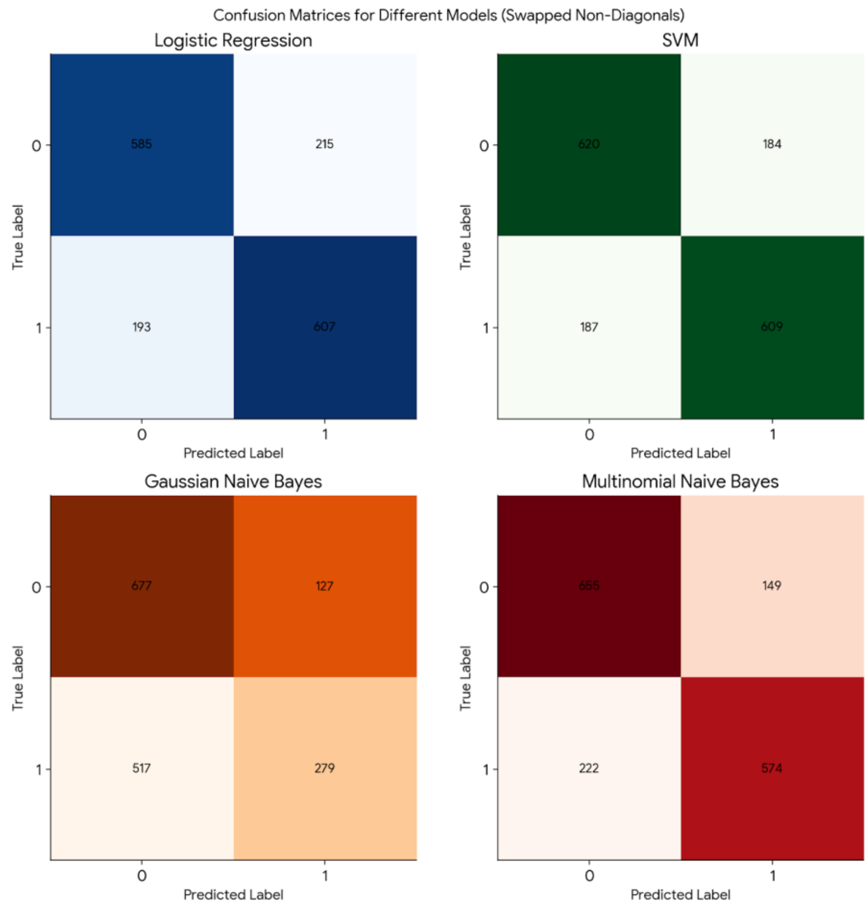

Fig. 7. Confusion Matrix of Balanced Dataset

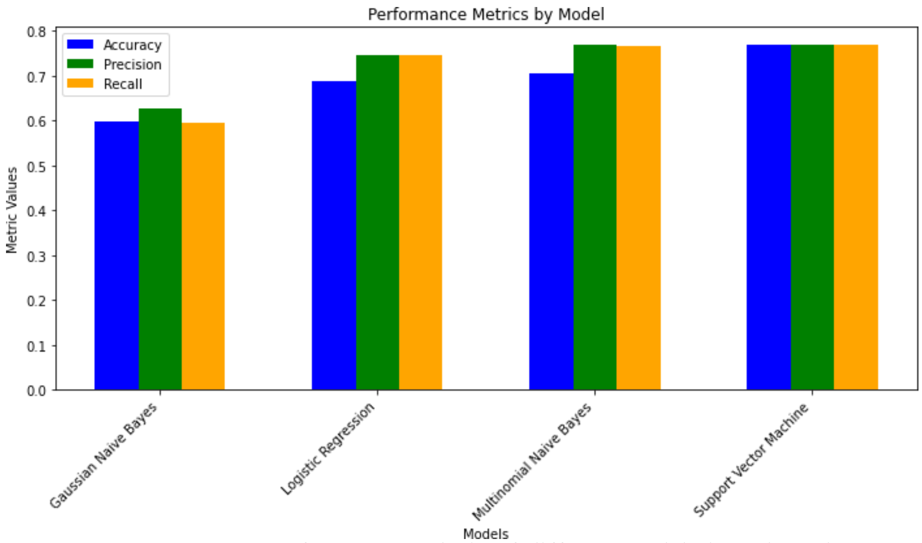

Fig. 8. Performance analysis of different models in Balanced Dataset

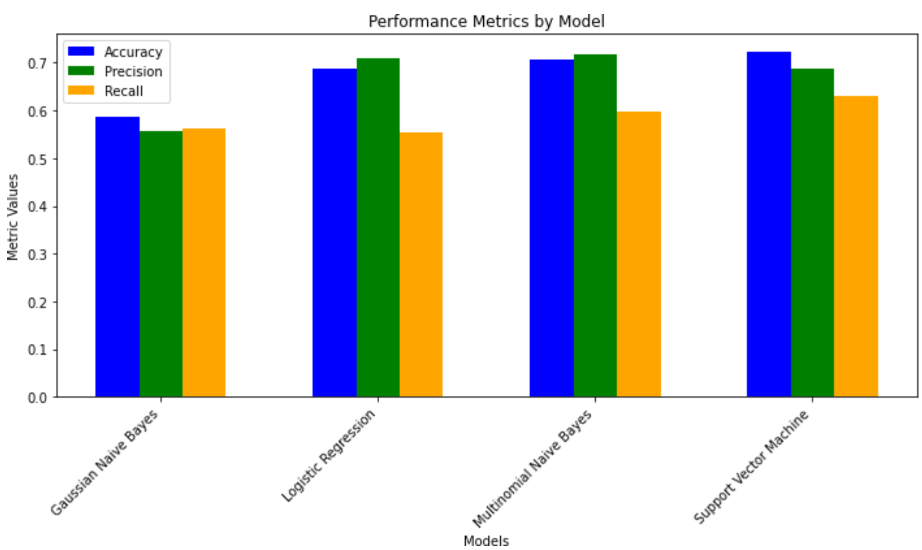

Fig. 9. Performance analysis of different models in Unbalanced Dataset

References

1. A. Ravindran, G. G. Krishna, Sagara and S. S., ”A Comparative Analysis of Machine Learning Algorithms in Detecting Deceptive Behaviour in Humans using Thermal Images,” 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2019, pp. 0310-0314, doi: 10.1109/ICCSP.2019.8697911.
2. R. S. Pillai and D. L. R, ”A Survey on Citation Recommendation System,” 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies(ICICICT), Kannur, India, 2022, pp. 423-429, doi: 10.1109/ICICICT54557.2022.9917887.
3. P.D. Turney,” Thumbs Up or Thumbs Down? Semantic Orientation Applied to UnsupervisedClassification of Reviews,” Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp. 417-424, July 2002.
4. Gayathri, R.G. and Nair, J.J., 2015. Towards efficient analysis of massive networks. International Journal of Applied Engineering Research, 10(69), pp.222-227.
5. Chaitanya, S. Sarath, Malavika, Prasanna and Karthik, ”Human Emotions Recognitionfrom Thermal Images using Yolo Algorithm,” 2020 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 2020, pp. 1139-1142, doi: 10.1109/ICCSP48568.2020.9182148.
6. Y. K. M, S. C K, T. A R, S. B. Kumar and G. Sarath, ”A Twitter-based Software VulnerabilityAlert Framework using Natural Language Processing,” 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 702-707, doi:10.1109/ICCES57224.2023.10192794.
7. Jasmine Bhaskar, Sruthi, K., and Prof. Prema Nedungadi, “Enhanced sentiment analysis ofinformal textual communication in social media by considering objective words and intensifiers”, in IEEE International Conference on Recent Advances and Innovations in Engineering(ICRAIE), 2014, Jaipur, 2014
8. Huq, Mohammad Rezwanul, Ali Ahmad, and Anika Rahman. ”Sentiment analysis on Twitter data using KNN and SVM.” International Journal of Advanced Computer Science and Applications 8, no. 6 (2017).
9. Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. ”Sentiment analysis algorithms and ap-plications: A survey.” Ain Shams engineering journal 5, no. 4 (2014): 1093-1113.
10. El Abdouli, Abdeljalil, Larbi Hassouni, and Houda Anoun. ”Sentiment analysis of moroccan tweets using naive bayes algorithm.” International Journal of Computer Science and Information Security (IJCSIS) 15, no. 12 (2017).
11. Vishwanathan, S. V. M., and M. Narasimha Murty. ”SSVM: a simple SVM algorithm.” In Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN’02 (Cat.No. 02CH37290), vol. 3, pp. 2393-2398. IEEE, 2002.
12. S. Bhuta, A. Doshi, U. Doshi and M. Narvekar, ”A review of techniques for sentiment analysis Of Twitter data,” 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2014, pp. 583-591, doi: 10.1109/ICI- CICT.2014.6781346. keywords: Niobium; Noise measurement; Sentiment Analysis; Supervised Learning; Social Networks,
13. Z. Jianqiang and G. Xiaolin, ”Comparison Research on Text Pre-processing Methods on Twitter Sentiment Analysis,” in IEEE Access, vol. 5, pp. 2870-2879, 2017, doi: 10.1109/AC- CESS.2017.2672677. keywords: Twitter; Sentiment analysis; Terminology; Support vector machines; Uniform resource locators; Organizations; Analytical models; Twitter; sentiment analysis; text pre-processing,