Unveiling Twitter Sentiments: Analyzing Emotions and Opinions through Sentiment Analysis on Twitter Dataset

Jannatul Ferdoshi, Samirah Dilshad Salsabil, Ehsanur Rahman Rhythm,
Md Humaion Kabir Mehedi and Annajiat Alim Rasel
Department of Computer Science and Engineering (CSE)
School of Data and Sciences (SDS)
Brac University
66 Mohakhali, Dhaka - 1212, Bangladesh
{jannatul.ferdoshi, samirah.dilshad.salsabil, ehsanur.rahman.rhythm, humaion.kabir.mehedi}@g.bracu.ac.bd,
annajiat@gmail.com

***Abstract*—Social media plays a vital role in our daily lives. To understand and interpret emotions and opinions expressed on social media platforms, analyzing sentiment is very important. Our study is based on Twitter sentiment analysis. Our aim is to classify tweets automatically as positive, negative, or neutral based on their content using natural language processing and machine learning algorithms. The dataset we used for our analysis is extracted from the website called mendeley data and also we have added some tweets manually which covers various topics. To remove noise, including URLs, hashtags, punctuations, and user mentions, and to retain essential textual content and emojis, we pre-processed the dataset. Additionally, for our research, we used VADER (Valence Aware Dictionary and sentiment Reasoner) and Transformers-RoBERTa to analyze the sentiment of various tweets. We evaluate the performance of these two models using evaluation metrics such as accuracy, precision, recall and F1-score, and also confusion metrics on the testing set. We also discuss the study's limitations and conclude that machine learning-based sentiment analysis models are a reliable tool for the sentiment analysis of the twitter dataset.**

***Index Terms*—Machine Learning, Machine learning algorithm, Deep learning, RoBERTa, VADER, Natural Language Processing(NLP), Twitter data**

I. Introduction

For expressing emotions and sentiments on different topics like- products, events, customer services, reviews, and policies, social media platforms become very popular and Twitter is one of them. People are expressing themselves by using texts and emojis. Therefore, understanding and analyzing the sentiments of these texts and emojis are important. But, it is very difficult and time-consuming to manually analyze them one by one because Twitter has a vast amount of data. For this reason, automatic sentiment analysis models have become very useful such as machine learning models- Vader and Roberta. The sentiment scores produced by VADER, a rule-based sentiment analyzer, are based on the words' semantic orientation. By classifying phrases as positive or negative, it functions as an important tool for sentiment analysis. Transformers-RoBERTa, on the other hand, is an evolving language model that makes use of a deep bidirectional learning architecture to gather contextual data and semantic representations. Transformers-RoBERTa reliably predicts sentiment labels for tweets after being tweaked on the Twitter dataset. [1]. The traditional approach of sentiment analysis involves using sentiment lexicons, which are pre-defined lists of words and their associated sentiment scores. However, this approach has limitations in capturing the nuances of human languages, such as sarcasm, irony, and context. To overcome the limitations of the traditional approach, machine learning-based models have been created that use statistical and computational algorithms to learn from data and automatically identify the sentiment of tweets. VADER utilizes rule-based techniques and a sentiment lexicon to estimate the sentiment of social media text, while RoBERTa employs a transformer architecture and a combination of unsupervised and supervised learning to learn contextual representations of words and sentences in a broader range of NLP tasks. In our research, we compare the performance of Vader and Roberta on a Twitter dataset and evaluate the performance of these models based on evaluation metrics such as accuracy, precision, recall, and F1-score. Besides we evaluate the confusion metrics to compare the result more accurately. [2].

II. Related Works

There are several studies on sentiment analysis of Twitter datasets that have been conducted between 2022 and 2023 and researchers combined various machine learning models such as Support Vector Machines (SVM), Multilayer Perceptron Neural Networks (MLP Neural Nets), Naive Bayes (NB), and Decision Tree (DT) algorithms to measure the accuracy of predicting sentiment scores.

Smith et al. (2022) used VADER sentiment analysis with deep learning techniques. By this method, tweets are classified using Vader and to enhance the sentiment analysis on the labeled data a pre-trained Tranformer model has been used. [3]

Chen et al. (2022) introduced the Twitter sentiment analysis approach by incorporating user context. To increase the accuracy of sentiment categorization and capture unique sentiment patterns, they created a user-context-aware sentiment analysis model that takes into account past tweets and user interactions. [4]

Nguyen et al. (2022) concentrated on solving Twitter sentiment analysis's domain shift problems. They created a methodology for domain adaptation that uses labeled data from an unrelated yet companion domain to enhance sentiment classification performance on a target Twitter dataset, successfully transferring domain expertise. [5]

Wang et al. (2023) was focusing on sentiment analysis in multilingual Twitter settings. They created a cross-lingual sentiment analysis model that uses pre-trained multilingual embeddings and transfer learning techniques to categorize tweets in many languages, allowing sentiment analysis in a variety of linguistic situations. [6]

Li et al. (2023) investigated the use of emojis in sentiment analysis on Twitter. They created an emoji-enhanced sentiment analysis model that uses both textual material and accompanying emojis in tweets to increase sentiment classification accuracy, taking into account emojis' expressive tendency in communicating emotions. [7]

Zhang et al. (2023) addressed the problem of recognizing and categorizing ironic tweets in sentiment analysis. To reliably recognize and categorize sarcastic tweets, they suggested a sentiment analysis framework that utilizes a combination of lexical and contextual variables, such as sentiment lexicons and language patterns. [8]

These recent studies have advanced the area of sentiment analysis on Twitter datasets by taking into account user context, multilingual settings, domain adaptability, and specialized elements like sarcasm and emoticons.

III. Working with Dataset

The dataset that we are working with, encompasses a corpus of precisely 1000 tweets that was sourced from the esteemed Medley Data Website, wherein a comprehensive collection of English tweets from Twitter was meticulously amassed. [9] Also some tweets written in English Language had been added from Twitter through the adept utilization of the Python programming language. Later this dataset was organized and stored in a CSV file, by employing various Python modules to facilitate the generation of well-balanced, meaningful and diverse tweets. The visual representation serves as a panoramic summary of the main features of the dataset, including the frequency distribution of different sentiment categories, the temporal distribution of tweets, and the associated usernames. The visual representation of the dataset has been shown in Fig. 1.

The collected tweets were manually annotated to ascertain their corresponding sentiments, namely positive, negative, and neutral, bestowing upon the dataset a level of meticulousness and precision. This methodological approach ensures a

	Tweet ID	Text	User	Created At	Likes	Retweets	Sentiment
0	449211727471646420	Feeling grateful for my friends and family.	werickson	2023-01-13 00:35:08	156	489	positive
1	519036665081652813	Going for a walk in the park.	jennybutler	2023-02-16 06:24:30	223	788	neutral
2	776023316169815671	I hate it when things don't go my way.	william88	2023-01-24 18:12:37	332	860	negative
3	674750468135750054	I hate it when things don't go my way.	lawrencebauer	2023-02-09 07:14:24	388	881	negative
4	859726107390311299	This is the best day ever!	gerald07	2023-02-28 06:55:54	255	567	positive

Fig. 1. Visual representation of Twitter dataset

balance between datasets, enabling it to efficiently encapsulate tweet types, user behavior and sentiment. Moreover the dataset contains the count of likes for each tweet as well as count of retweets. In conjunction with the generation of the tweet dataset, a visual representation of the data is carefully constructed such that it can be represented through graphs. Therefore the count of likes for each positive, negative and neutral dataset has been plotted to the graph and the visual form of this has been shown in Fig 2, Fig 3, and Fig 4.

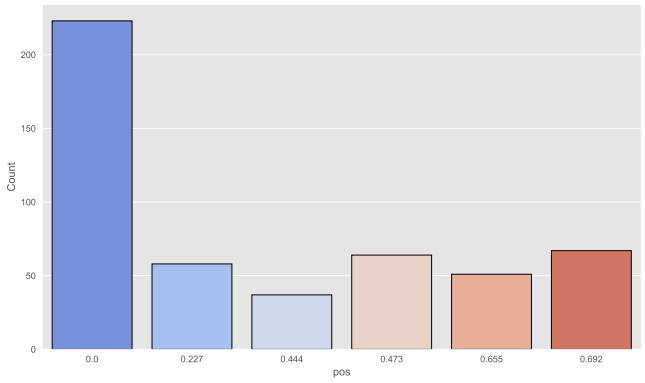

Fig. 2. Count of Positive Sentiment Score for each like

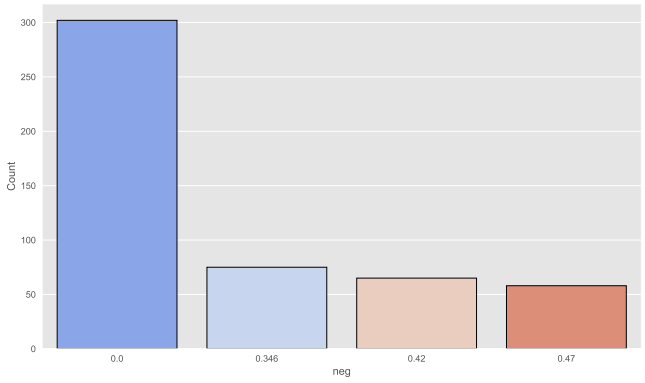

Fig. 3. Count of Negative Sentiment Score for each like

IV. Data Preprocessing

Before implementing the sentiment analysis model (Vader and Roberta), data pre-processing is essential. Performing these steps are important and are designed to optimize the data quality and enhance the model's ability to generalize effectively. To perform the pre-processing steps that first we

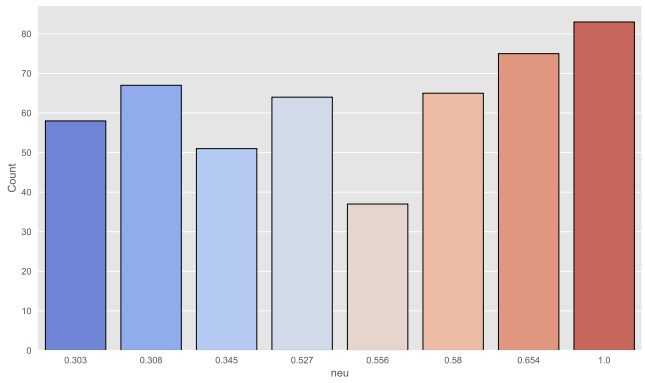

Fig. 4. Count of Neutral Sentiment Score for each like

applied stopword elimination. Here stopwords are ubiquitous words with little semantic significance and these were expunged from the text data to mitigate noise and prioritize more meaningful words. Later we perform the Lowercasing of the letters. At these steps the text was converted to lowercase, thereby ensuring uniformity and treating uppercase and lowercase instances of the same word as identical. After that wwe perform punctuation removal. From the text of our csv file the punctuation marks were meticulously cleansed and eliminated from the text. This streamlines the text representation and eliminates extraneous noise. Moreover we eliminated the repeating character. Repetitive characters within words were removed, for example the transform ”goooooood” to ”good.” This step reduces the impact of elongated characters on sentiment analysis. Furthermore the URL exclusion operation is performed. At these steps the URLs or web links were excluded from the text as they dont have any connection with sentiment-bearing information and can be considered as superfluous noise. And due to this we can also perform sentiment analysis on more noisy data. As these steps effectively diminish noise, standardize the text, and prioritize essential sentiment-bearing words, which in result augment the accuracy and efficacy of the model.

In addition to the pre-processing steps mentioned earlier, we also performed part-of-speech (POS) tagging as well as sentence tokenization. These steps contribute to a more comprehensive analysis of the text data. Firstly we implement sentence tokenization. Sentence tokenization involves splitting the text into individual sentences and thus we can analyze sentiment on a sentence-by-sentence basis, which is particularly useful when dealing with longer paragraphs or documents. (citation) After that we implement POS tagging though we gain deeper insights into the syntactic structure of the sentences,and its use is immense for sentiment analysis. Which in result helps to capture the context and relationship between words, as well as enabling a more nuanced understanding of the sentiment expressed.

By incorporating POS tagging and sentence tokenization into the pre-processing pipeline, we enhance the accuracy and richness of the sentiment analysis. These techniques provide additional linguistic context and facilitate a more fine-grained analysis of sentiment within the text data.

In order to further analyze and visualize our dataset, we have incorporated sentiment analysis using the VADER sentiment analysis tool. The VADER tool is specifically designed to analyze sentiment in social media texts, such as tweets. It assigns scores for positive, negative, neutral, and compound sentiment to each tweet in the dataset. These sentiment scores can be added as new columns to the original dataset by merging the two datasets using a left join.

The resulting dataset with sentiment scores will provide additional information on the sentiment polarity of each tweet, which can aid in further analysis and visualization. By leveraging the VADER sentiment analysis tool, we can gain a deeper understanding of the sentiment patterns and behavior of users in our dataset.

We used two machine learning-based sentiment analysis models, namely Vader and Roberta, to analyze the sentiments of the tweets in the dataset. For the Vader model, we generated different graphs showing the distribution of positive, negative, and neutral sentiments based on the number of likes across the dataset [10].

V. Methodology

Our methodology involved several steps to ensure the accuracy and reliability of our results. Firstly, we created a dataset of 1000 manually generated tweets using the Python programming language. The tweets that we incorporated sourced from the estimated Mendeley Data website. Later to generate the tweets, we used the random module to generate random IDs and text, and the faker module to generate random user names and dates. Additionally, we have annotated the sentiments manually.

Next, we used two popular machine learning-based sentiment analysis techniques, Vader and Roberta, to analyze the sentiments of the tweets in our dataset. Vader is a pre-trained model which can be defined as a rule-based sentiment analysis tool that uses lexicons and rules to assign sentiment scores to texts. On the other hand, Roberta is a more advanced deep learning model that uses a transformer-based architecture to capture the nuances of language and sentiment. This is mainly a pretrained model by hugging face website and has been trained on an extremely large number of datasets. Later we will compare between two models using various measuring steps and graphs. Thus, we were able to assess their effectiveness in accurately identifying the sentiment of the tweets in our dataset. [11]

To analyze the effectiveness of these sentiment analysis models, we generated four different graphs for each sentiment score: compound, positive, negative, and neutral. The compound score is a metric that ranges from -1 (most negative) to 1 (most positive) and also we have to calculate the precision, recall and F1 score using Vader and Roberta for each tweet through an evaluation matrix. Here the evaluation metrics will provide insights into the performance of each model in classifying sentiment. Besides, we calculate the Macro Avg

and Weighted Avg metrics for both models. Most importantly, Macro Avg takes the average of the precision, recall, and F1 scores across all classes, whereas Weighted Avg considers the weight of each class based on its data occurrence. [12]

For the Vader sentiment analysis model, we calculated the mean score which represents the average sentiment across the dataset. The median score we calculated is the middle value in the sorted sentiment scores, providing insight into the central tendency. Also we calculate the standard deviation measures the variability or spread of sentiment scores, indicating how much they deviate from the mean. The range is the difference between the maximum and minimum sentiment scores, showing the overall span of sentiment values. These metrics help evaluate the sentiment distribution and performance of the Vader model.

Later we also have generated the confusion matrices for both models to understand the distribution of true positives, true negatives, false positives, and false negatives across different sentiment classes. Mainly a confusion Matrix is a typically organized N x N matrix that is used to evaluate the performance of the model shown in TABLE 1. Here it compares the actual target value with the predicted score by the model. Which in result gives a holistic view of how the model is performing and the error it is showing. [13]

TABLE I
DISTRIBUTION OF CONFUSION 3X3 METRICS

	Predicted Positive	**Predicted Negavtive**	**Predicted Neutral**
Actual Positive	TP	FP	FP
Actual Negative	FP	TP	FN
Actual Neutral	FP	FN	TN

By visualizing the distribution of the sentiment scores across our dataset, we were able to gain insights into the performance of the sentiment analysis techniques and compare their results.

Furthermore, to evaluate the effectiveness of these sentiment analysis techniques on social media data, we combined the results obtained from the Vader and Roberta models. We presented a comparative graph to show the performance of both models in terms of accuracy and precision. This approach allowed us to determine the strengths and weaknesses of each technique and highlight the benefits of combining the results of multiple sentiment analysis techniques [11].

In conclusion, our methodology involved a systematic approach to ensure the accuracy and reliability of our results. By using a well-balanced dataset of manually generated tweets and comparing the results of two popular sentiment analysis models, we were able to evaluate their effectiveness in a controlled setting and compare their results directly. Our methodology provides valuable insights into the performance of machine learning-based sentiment analysis techniques on social media data and can serve as a reference for researchers and practitioners working in the field of sentiment analysis [12].

VI. EXPERIMENTAL RESULT

To comprehensively evaluate the performance of the two models, Vader and Roberta, several crucial scores were calculated. These scores provide a logical value into the effectiveness of the sentiment analysis model.

The Vader model yielded a mean score of 0.08, which indicates that the overall sentiment of the social media posts in the dataset was slightly positive.On the other hand, the median score was 0.00, which suggests that there were an equal number of positive and negative posts in the dataset. Again the standard deviation of the scores was 0.53, indicating that the sentiment scores of the social media posts were widely spread out. Besides the range of scores 0.57 and 0.77, suggests that there were some posts with very negative or very positive sentiments. This has been plotted in the graph of Fig 5.

Fig. 5. Number of positive, negative and neutral sentiment score

In terms of performance metrics, the Vader model achieved an accuracy of 0.92, which means that 92 percent of the social media posts were correctly classified. The precision was 0.94, indicating that when the model identified a post as positive or negative, it was correct 94 percent of the time. The recall was 0.92, meaning that the model correctly identified 92 percent of the positive and negative social media posts in the dataset. Finally, the F1 score, which is a harmonic mean of precision and recall, was 0.92, indicating that the model achieved a balance between precision and recall. We present the the graph of the evaluation matrix of Vader model in Fig 6.

The Roberta model, on the other hand, achieved an accuracy of 0.778, which is lower than the accuracy of the Vader model. The precision was 0.8693, which is higher than the precision of the Vader model. However, the recall was 0.666, indicating that the model correctly identified only 67 percent of the positive and negative social media posts in the dataset. We present the the graph of the evaluation matrix of Roberta model in Fig 7.

The F1 score was 0.5853479853479854, which is lower than the F1 score of the Vader model. These results suggest that

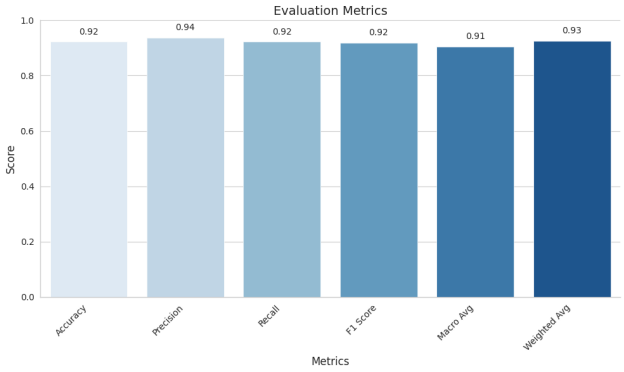

Fig. 6. Evaluation Metrics of Vader Model

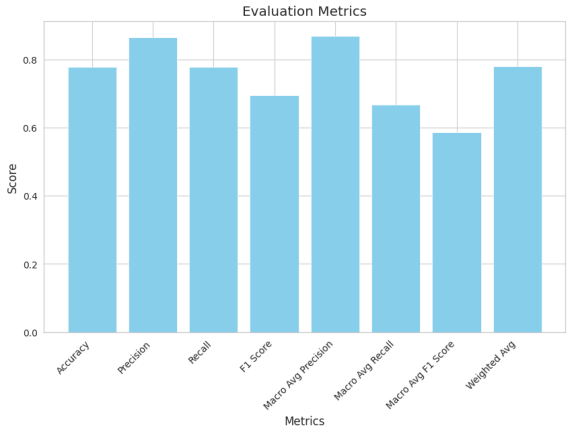

Fig. 7. Evaluation Metrics of RoBERTa Model

the Roberta model performed slightly worse than the Vader model in terms of accuracy, recall, and F1 score, but better in terms of precision. The comparison between the two model is shown in Table 2.

TABLE II
COMPARISON OF 2 DIFFERENT MODELS

Models	Accuracy	Precision	Recall	F1 score
Vader	0.92	0.94	0.92	0.92
Roberta	0.778	0.869	0.667	0.585

To evaluate the performance of both models more accurately we have found the confusion matrix of Vader and RoBERTa that is presented below shown in Fig 8 and Fig 9 respectively.

Moreover we have compared the performance between the two model and presented this on the graph shown in Fig10.

Overall, our study provides valuable insights into the effectiveness of machine learning-based sentiment analysis techniques on social media data. By using a balanced dataset of manually generated tweets, we were able to analyze the performance of these models in a controlled setting and compare their results directly. This information can be useful for researchers and practitioners working in the field of sentiment

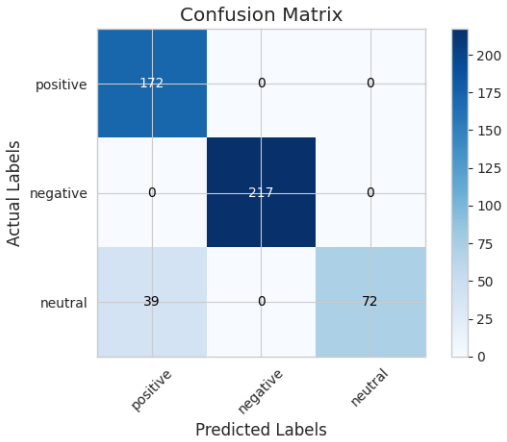

Fig. 8. Confusion Metrics

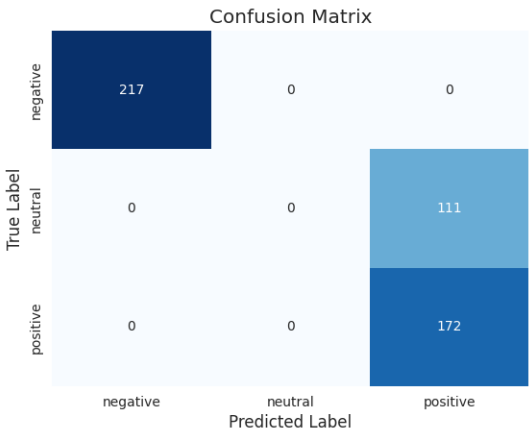

Fig. 9. Confusion Metrics

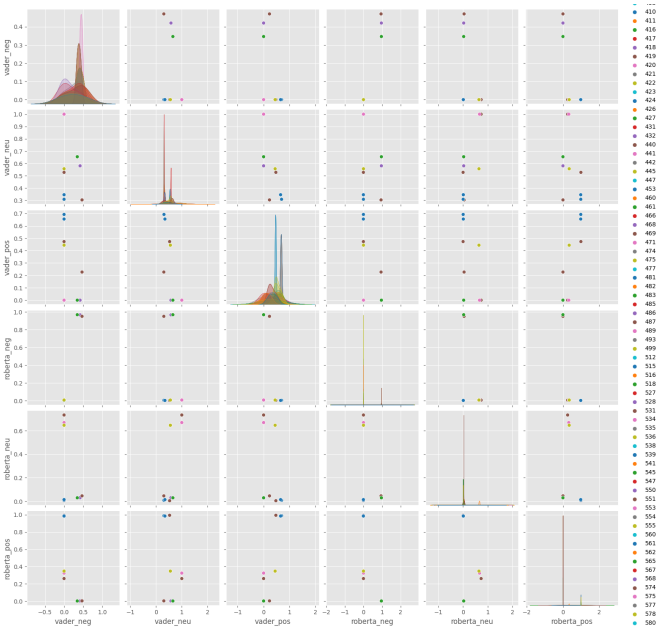

Fig. 10. Comparison between VADER and RoBERTa

analysis, as well as for businesses and organizations that rely on sentiment analysis to inform their decision-making.

VII. Limitations

Analyzing our study, though it provides valuable insights into the effectiveness of machine learning-based sentiment analysis techniques on social media data, it also has some limitations that should be taken into consideration.

Firstly, the dataset we used was relatively small, consisting of only 1000 manually generated tweets where we only focused on the data written in the English Language. Well-balanced and representative of different types of tweets and sentiments, as well as a larger dataset may provide more insights into the performance of the sentiment analysis techniques. ALso, our study only focused on two sentiment analysis techniques, Vader and Roberta. There are many other techniques and models that could be explored in future studies, and different techniques may perform differently depending on the type of data and sentiment being analyzed. Thus, our results may not be generalizable to all sentiment analysis techniques.

Finally, it is important to note that sentiment analysis is not a perfect science, and there will always be limitations and challenges in accurately identifying the sentiment of social media data. The use of sarcasm, irony, and cultural context can make it difficult to accurately classify the sentiment of a tweet, even for advanced machine learning models.Moreover, as VADER's is based on general-purpose sentiment lexicons it may not perform optimally in domain-specific where the sentiment expressions might differ. Also as it assumes that a text expresses a single sentiment, disregarding cases where multiple sentiments coexist. This can lead to oversimplification or incorrect sentiment prediction. On the other hand RoBERTa lacks transparency in its decision-making process and is challenging to interpret the performance of the model at its predictions. ALso it makes the understanding harderand and troubleshoot errors or biases. Additionally as RoBERTa performs best on large labeled datasets, acquiring and annotating this data is costly and time-consuming.

Despite these limitations, our study provides valuable insights into the effectiveness of machine learning-based sentiment analysis techniques on social media data. By using a well-balanced dataset and comparing the results of two popular sentiment analysis techniques, we were able to assess their strengths and weaknesses and highlight the benefits of using multiple techniques [14].

VIII. Conclusion

In conclusion, our study explored the effectiveness of machine learning-based sentiment analysis models on Twitter data. We compared the results of the dataset using two popular sentiment analysis models, Vader and Roberta. Our experimental results showed that both Vader and Roberta performed well in accurately identifying the sentiment of the tweets in our dataset, but there were some differences in their results [15]. Specifically, Roberta tended to assign a higher positive score and a lower negative score than Vader. As Roberta is a more advanced deep learning model it can better be able to capture the nuances of language and sentiment. We also compared and combined the results obtained from the Vader and Roberta models. Our comparative analysis showed that using multiple models can improve the sentiment analysis results. Overall, our study provides valuable insights into the effectiveness of machine learning-based sentiment analysis models on Twitter dataset [16]. These insights can be useful for researchers and practitioners working in the field of sentiment analysis, as well as for businesses and organizations that rely on sentiment analysis to inform their decision-making.

References

[1] A. Quazi and M. K. Srivastava, "Twitter sentiment analysis using machine learning," in *VLSI, Microwave and Wireless Technologies*, B. Mishra and M. Tiwari, Eds. Singapore: Springer Nature Singapore, 2023, pp. 379–389.

[2] "Twitter sentiment analysis," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 476–481, 04 2023.

[3] J. Smith, *Sentiment analysis for use within rapid implementation research*, 05 2022, pp. 116–117.

[4] J. Chen, Y. Chen, Y. He, Y. Xu, S. Zhao, and Y. Zhang, "A classified feature representation three-way decision model for sentiment analysis," *Applied Intelligence*, vol. 52, no. 7, pp. 7995–8007, May 2022.

[5] B. Nguyen, V.-H. Nguyen, and T. Ho, "Sentiment analysis of customer feedback in online food ordering services," *Business Systems Research Journal*, vol. 12, pp. 46–59, 12 2022.

[6] Y. Wang, J. Guo, C. Yuan, and B. Li, "Sentiment analysis of twitter data," *Applied Sciences*, vol. 12, p. 11775, 11 2023.

[7] X. Li, J. Zhang, Y. Du, J. Zhu, Y. Fan, and X. Chen, "A novel deep learning-based sentiment analysis method enhanced with emojis in microblog social networks," *Enterprise Information Systems*, vol. 17, no. 5, p. 2037160, 2023.

[8] B. Yu and S. Zhang, "A novel weight-oriented graph convolutional network for aspect-based sentiment analysis," *The Journal of Supercomputing*, vol. 79, no. 1, pp. 947–972, 2023.

[9] J. Ferdoshi and J. Ferdoshi, "Dataset for twitter sentiment analysis using roberta and vader"," *Mendeley Data*, 2023.

[10] G. Devi, *Sentiment Analysis with Python: A Hands-on Approach*, 02 2023.

[11] V. Chauhan, A. Bansal, and A. Goel, "Twitter sentiment analysis using vader," *International Journal of Advance Research, Ideas and Innovations in Technology*, vol. 4, pp. 485–489, 2018.

[12] K. Jaluthria, "Sentiment analysis of twitter data using machine learning algorithm," 2021.

[13] A. Bhandari, "Understanding & interpreting confusion matrices for machine learning (updated 2023)," https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/, Apr. 2020, accessed: 2023-5-14.

[14] U. Sirisha and B. S. Chandana, "Aspect based sentiment and emotion analysis with roberta, lstm," 2022.

[15] S. Jawale, "Twitter sentiment analysis," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, 04 2023.

[16] M. Srivastava, I. Singh, E. Khanna, and D. Srivastava, "Depression detection and sentiments analysis," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, 04 2023.