Enhancement of Support Vector Machine utilizing RoBERTa applied to Sentiment Analysis of Facebook Data

EUNNA JAZREL M. ARCILLA[1], PATRICIA MAE A. SAMSON[2], RAYMUND M. DIOSES[3], FLORENCIO V. CONTRERAS JR.[4], RICHARD C. REGALA[5], JONATHAN C. MORANO[6], LEISYL M. MAHUSAY[7], JAMILLAH S. GUIALIL[8]

[1, 2, 3, 4, 5, 6, 7, 8] *College of Information System and Technology Management, Pamantasan ng Lungsod ng Maynila, Intramuros, Manila, Philippines*

Abstract- ***Many people are using social media sites like Facebook to express their opinions, experiences, or whatever they want to post online. Understanding user sentiment has become crucial for various applications, ranging from marketing to public opinion analysis. Researchers use natural language processing (NLP) and machine learning algorithms to evaluate textual information from Facebook posts and classify sentiments as positive, negative, or neutral. This study delves into sentiment analysis of Facebook data to better understand how users express their emotions. Additionally, the method addresses the limitations of sentiment analysis on social media due to informal language, slang, and context-dependent phrases. The study aims to develop an enhanced Support Vector Machine algorithm for sentiment analysis of Facebook data by utilizing the RoBERTa (A Robustly optimized BERT) model. To enhance sentiment accuracy, thus the performance of the traditional SVM algorithm, the proposed approach uses VADER to predict initial sentiment labels, loads a pre-trained RoBERTa model as preprocessing techniques, fine-tunes the RoBERTa model and extracts RoBERTa embeddings to optimize the SVM algorithm. This improves the model's capacity to handle imbalanced datasets and efficiently manage larger datasets while filtering out noisy or irrelevant characteristics. To analyze the performance of the proposed technique, results are compared with the result of existing algorithms. The enhanced SVM algorithm significantly outperforms the existing approach in terms of accuracy, precision, recall, and F1-score, with a 4% to 8% improvement in accuracy over the previous algorithm. This research highlights the potential of integrating RoBERTa techniques with SVM for enhanced sentiment analysis.***

Indexed Terms- ***Sentiment Analysis, Support Vector Machine, RoBERTa, Facebook***

I. INTRODUCTION

In today's data-driven world, sentiment analysis, also known as opinion mining, is an essential method that makes it possible to automatically identify and categorize subjective information in text. This data may be utilized to determine consumer preferences, assess public opinion on social problems, and ascertain customer emotion. Sentiment analysis is becoming an essential tool for researchers, corporations, and governments due to the growth of social media and online platforms. We may learn a lot about customer preferences, brand perception, and public opinion by examining sentiment in online reviews, social media posts, and other text-based communication.

The Support Vector Machine (SVM) is a well-known machine learning technology that has grown in popularity due to its ability to address challenging categorization problems in a wide range of applications. SVM models are frequently used in text classification issues because of one of its advantages, which is their capacity to handle huge quantities of features [1]. Support Vector Machines have shown promise in sentiment analysis, a discipline that integrates linguistics and computer science to determine text sentiment automatically. SVMs have

shown excellent performance in prior sentiment analysis studies [2].

Despite the superior performance of Support Vector Machines (SVM) in sentiment analysis, yielding higher and more accurate predictions than other algorithms, they exhibit several limitations. These include challenges in handling large and imbalanced datasets and noise [3][4]. To address the challenges associated with the Support Vector Machine Algorithm, Facebook AI Research developed the cutting-edge deep learning model RoBERTa (Robustly Optimized BERT Pre Training Approach) for applications related to natural language processing (NLP).. Using RoBERTa, a huge transformer-based language model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture, is pre-trained on a large corpus of text data. Nonetheless, RoBERTa modifies and enhances the BERT architecture in several ways, improving performance across a range of NLP applications. Because of its large and diverse training corpus, RoBERTa can learn more dependable and generalizable representations of text data, which can improve performance on subsequent NLP tasks like sentiment analysis and produce state-of-the-art results on sentiment analysis tasks[5].

In this research, we developed an enhanced support vector machine algorithm integrating the roberta model to solve challenges and improve the algorithm's accuracy and performance.Identify the constructs of a Journal – Essentially a journal consists of five major sections. The number of pages may vary depending upon the topic of research work but generally comprises up to 5 to 7 pages. These are:

II. LITERATURE REVIEW

A. Sentiment Analysis

Sentiment analysis is a crucial aspect of Natural Language Processing (NLP), which classifies texts based on the sentiment orientation of the thoughts they include. It identifies a text's contextual polarity and determines whether it is positive, negative, or neutral. This process is also known as opinion mining, as it derives the speaker's opinion or attitude. Sentiment analysis is particularly useful in merchants, stock traders, and election works. Social networks provide a platform for users to express their thoughts on a regular basis, and research is ongoing due to their importance in marketing rivalry and evolving consumer demands. Sentiment analysis requires a training set for optimal performance, and semantic analysis of sentences improves the meaning and accuracy of the results [6].

B. Support Vector Machine

Support Vector Machine was used in the study's sentiment analysis of Bangladesh cricket. To train the model, the authors used both a larger ABSA dataset and a smaller subset of their own data. The model's accuracy on the ABSA dataset was 73.49%, according to the results, which showed that it worked well. The study illustrated the effective use of SVM in sentiment analysis of Bangladesh cricket, but the authors also highlighted the requirement for additional data and preprocessing techniques to enhance the model's performance [8].

The accurate classification of minority class objects in imbalanced data sets is a difficult task. Support vector machines, a traditional classification technique, struggle to produce the ideal separation hyperplane for an SVM trained on skewed data, hence they do not perform well for these imbalance data sets[9].

Support Vector Machine is one of the most powerful and robust classification and regression algorithms in multiple fields of application. However, despite having many advantages, SVM has some weaknesses. SVMs are not suitable to classify large datasets due to their high training complexity, which is highly dependent on the input size.[9].

Support Vector Machine in noisy data can be prone to inaccurate results due to their inability to handle outliers, missing values, or incorrect labels. These noise points can significantly impact SVMs' performance, leading to poor generalization and inaccurate predictions. Missing values can cause biased or incomplete models, while incorrect labels can introduce noise and affect SVMs' performance [10].

C. Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa (Robustly Optimized BERT method), which was introduced by Facebook, is one of its well-known variations. In essence, it is a more powerful and capable version of BERT that can handle more data. It has been demonstrated that RoBERTa has a higher prediction power than BERT. Balakrishnan V. et al. (2022)

In [12], the study presents a novel approach to sentiment analysis and key entity detection in social media, utilizing a pre-train model for sentiment analysis and a Machine Reading Comprehension task for key entity detection. The RoBERTa fine-tuning model is used as a pre-training model for fine-tuning, and various fine-tuning methods are employed to implement sentiment analysis and key entity detection. Experimental results show that the RoBERTa fine-tuning model outperforms traditional models in sentiment analysis. The authors consider sentiment analysis as a classification problem, extracting negative emotion information, and key entity detection as a sentence matching or Machine Reading Comprehension task. This unified approach to sentiment analysis and key entity detection is a promising solution for enhancing social media sentiment mining and public opinion analysis.

III. METHODOLOGY

The research conducted in this study is based on Facebook data. The datasets used in this study are collected using Apify. The researchers use this tool to scrape or gather Facebook posts from Facebook pages which consist of 20,000 Facebook posts. The dataset was divided into a training and test dataset with a ratio of 80:20. This dataset was made publicly available with the goal of enabling future study in areas such as present research by providing researchers with a resource that allows them to explore different aspects of Facebook data. However, due to hardware limitations, the researchers only used 1000, 2000, 3000 and 4,000 datasets for different testing to evaluate the performance metrics between the existing algorithm and the proposed algorithm.

The simulation starts with Data Collection to compile a comprehensive dataset of English-language text relevant to Facebook data. Following this is to make the SVM classifier more robust in handling an imbalanced dataset, thus enhancing the performance by Fine-tuning the RoBERTa model. Subsequently, to increase the scalability of SVM classifiers to large datasets by loading pre-trained RoBERTa models This enabled handling larger volumes of data efficiently, improving the classifier's performance and reducing computational overhead. lastly, by generating RoBERTa embeddings to enhance the SVM algorithm and make it more robust to noise in the dataset. This involved leveraging RoBERTa's embedding generation capabilities to capture semantic information from the data, thereby enhancing the classifier's resilience to noisy or irrelevant features.

A. Data Preprocessing using RoBERTa Tokenizer

The researchers used the RoBERTa model as preprocessing techniques; the RoBERTa model employs the Byte-Pair (BPE) algorithm to learn tokens [13]. It learns individual characters from a training corpus and examines the most common adjacent characters using a frequency count. The first character remains the same, while the others are denoted with a G-symbol (e.g., ’yes’ ˙ → ’y’, ’Ge’, ’ ˙ Gs’). The highest frequency adjacent pair is added to the vocabulary, resulting in a vocabulary with all characters, including mergers.[14]

B. Fine Tuning of RoBERTa Model

It sets up a tokenizer and a classification model, prepares training and test data, and configures an AdamW optimizer with parameters such as batch size[32], epochs[3], and learning rate[5e-5]. A scheduler is put up to improve the learning rate during training. Sequences are padded for classification, and DataLoader is used to efficiently handle training data. Class weights are determined to account for imbalance, and a loss function is created. The training loop runs for the set number of epochs, computes the loss, calculates the gradients, and uses backpropagation to update the model parameters. Evaluation findings are saved and repeated for a set number of epochs. Fine tuning of RoBERTa is used to improve the performance of the model, thus handling an imbalanced dataset.

C. RoBERTa embeddings

RoBERTa is a prominent NLP model that uses Deep Learning techniques for natural language processing. The RoBERTa model employs embedding techniques like token embedding and position embedding. It uses optimum MLM pre-training approaches and eliminates Next Sentence Prediction (NSP). The model's enhanced natural language processing is obtained through longer training times and larger datasets. Furthermore, RoBERTa incorporates advanced data augmentation methods such as sentence order modification and token randomization to improve its comprehension of context and sentence links.[15]

EXISTING SUPPORT VECTOR MACHINE			
Dataset	Precision	Recall	F1-score
1,000	68.06	63.35	58.40
2,000	63.51	61.05	61.59
3,000	77.29	64.45	67.82
4,000	82.35	72.36	75.33

PROPOSED SUPPORT VECTOR MACHINE WITH ROBERTA			
Dataset	Precision	Recall	F1-score
1,000	66.93	66.61	66.50

Testing the dataset in various ways not only strengthens model training but also facilitates comprehensive evaluation of performance metrics between the existing and proposed algorithms, thereby enhancing the credibility of the results. Essentially, this dataset serves as a crucial component in furthering our understanding of sentiment analysis on Facebook data.

III. RESULTS

EXISTING SUPPORT VECTOR MACHINE	PROPOSED SUPPORT VECTOR MACHINE WITH ROBERTA

Dataset	Accuracy	Dataset	Accuracy
1,000	63.35%	1,000	71.49%
2,000	67.75%	2,000	72.50%
3,000	75.29%	3,000	79.28%
4,000	79.63%	4,000	84.55%

Table 1 Accuracy Results of Enhanced Support Vector Machine with RoBERTa

Results have shown that the proposed algorithm consistently outperforms the existing algorithm in all 4 types of data sets. It is observed that the results of the existing support vector machine are lower due to the variability in the datasets that affect the accuracy. This method not only showed the algorithm's robustness with imbalanced datasets, but it also proved that the accuracy increases were consistent and dependable under a variety of situations and data sizes. Table 1 clearly demonstrates that the proposed system consistently exceeds the performance of the existing algorithm in all testing scenarios. The proposed algorithm achieved 4% to 8% improvement in the accuracy, the integration of RoBERTa models contributes significantly to this enhancement, offering more comprehensive handling of linguistic nuances and context than the traditional SVM technique. The findings demonstrate the effectiveness of embedding new deep learning algorithms into classic machine learning frameworks to better handle complicated, real-world data in sentiment analysis.

2,000	65.14	65.13	64.97
3,000	77.65	74.30	75.51
4,000	82.59	82.79	82.68

Table 2 Combined Evaluation Metrics of Existing and Enhanced Support Vector Machine Algorithm

Table 2 illustrates the combined evaluation metrics for both the existing and enhanced algorithms. Utilizing pre-trained RoBERTa models with SVM classifiers showcased notable enhancements in scalability, as indicated by performance metrics. The models' performance underwent evaluation using diverse metrics, offering quantitative insights into their effectiveness in aspects like precision, recall,

and F1-score. These metrics are pivotal for gauging how adeptly a model addresses the particular tasks or complexities of the dataset. Integrating RoBERTa models into the SVM framework not only enhanced the classifiers' scalability but also bolstered their overall effectiveness, as evidenced by the comprehensive evaluation metrics.

The implementation of RoBERTa embeddings to the enhanced method increased its capacity to resist noise in datasets. By capturing semantic subtleties, these embeddings strengthened the classifier and improved sentiment analysis accuracy. Furthermore, a significant improvement was the addition of a Neutral label alongside existing sentiment labels, effectively addressing inconsistencies and inaccuracies in the prior algorithm. This enhancement enabled a more detailed analysis of sentiment.

Overall, the improved algorithm represents a significant advancement in sentiment analysis, overcoming the limitations of previous systems. The researchers established a more robust and precise system capable of managing large datasets while offering nuanced sentiment analysis by combining new approaches with rigorous assessment.

CONCLUSION

The integration of RoBERTa with Support Vector Machine showed promising outcomes, including increased accuracy and classification performance over the previous approach. By using a fine-tuned RoBERTa model, the algorithm's issues with imbalanced datasets have been resolved. This allows for better handling of subtle or complex class distributions, improving prediction accuracy and reliability. Adding a pre-trained RoBERTa model to the existing algorithm, which previously struggled with large datasets due to its complexity being tied to the size of the input, has also shown improvements. This adjustment has helped the classifier work more efficiently with larger datasets, enhancing its overall performance. The implementation of generated RoBERTa embeddings has played a significant role in filtering out noisy or irrelevant features, greatly contributing to the improvement of SVM’s ability to derive meaningful representations and boosting its overall performance.

This study has demonstrated the effectiveness of using the RoBERTa model in enhancing the Support Vector Machine. Researchers recommend trying out different algorithms to thoroughly grasp the capabilities of the enhanced algorithm. To further improve the reliability and validity of the results, future researchers should use larger datasets to give a more thorough and robust foundation for their studies, resulting in more reliability and accuracy in their findings. Additionally, It is recommended to future researchers to have a laptop that has good specifications to ensure that it can efficiently handle larger datasets, enabling smoother data processing and analysis. These recommendations are intended to guide future research into enhancing the capabilities of the enhanced Support Vector Machine algorithm across a wide range of supervised learning methods.

ACKNOWLEDGMENT

The researchers would like to express their profound gratitude to the Lord Almighty for His divine guidance and continuous presence during this research journey.

Our heartfelt thanks goes to our loving family who have given us unwavering love and support which have been a source of strength and encouragement throughout this journey. We also want to acknowledge our friends, whose unwavering companionship and encouragement lifted our spirits and fueled our determination to pursue this research with diligence and passion.

Special thanks to our thesis adviser, panel members, and coordinators for their helpful guidance and valuable insights that supported us throughout our research journey. Their advice really made a difference and helped us improve our work, enabling us to successfully finish this research.

REFERENCES

[1] Zainuddin, N., & Selamat, A. (2014, September). Sentiment analysis using support vector machine. In 2014 international conference on computer, communications, and control technology (I4CT) (pp. 333-337). IEEE.

[2] Mullen, T., & Collier, N. (2004, July). Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 412-418).

[3] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215.

[4] Li, H. X., Yang, J. L., Zhang, G., & Fan, B. (2013). Probabilistic support vector machines for classification of noise affected data. Information Sciences, 221, 60-71.

[5] Kumar, B. P., & Sadanandam, M. (2023). A Fusion Architecture of BERT and RoBERTa for Enhanced Performance of Sentiment Analysis of Social Media Platforms.

[6] Devika, M., Sunitha, C., & Ganesh, A. (2016, January 1). Sentiment Analysis: A Comparative Study on Different Approaches. Procedia Computer Science. https://doi.org/10.1016/j.procs.2016.05.124

[7] Anyim, J. A. (2014). A comparative evaluation of sentiment analysis techniques on Facebook data using three machine learning algorithms: Naïve Bayes, maximum entropy and support vector machines (Doctoral dissertation).

[8] Mahtab, S. A., Islam, N., & Rahaman, M. M. (2018, September). Sentiment analysis on bangladesh cricket with support vector machine. In 2018 international conference on Bangla speech and language processing (ICBSLP) (pp. 1-4). IEEE

[9] Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189-215.

[10] Li, H. X., Yang, J. L., Zhang, G., & Fan, B. (2013). Probabilistic support vector machines for classification of noise affected data. Information Sciences, 221, 60-71.

[11] Balakrishnan, V., Shi, Z., Law, C. L., Lim, R., Teh, L. L., & Fan, Y. (2022). A deep learning approach in predicting products' sentiment ratings: a comparative analysis. The Journal of Supercomputing, 78(5), 7206-7226.

[12] Zhao, L., Li, L., Zheng, X., & Zhang, J. (2021, May). A BERT based sentiment analysis and key entity detection approach for online financial texts. In 2021 IEEE 24th International conference on computer supported cooperative work in design (CSCWD) (pp. 1233-1238). IEEE.

[13] Sennrich, R., Haddow, B., & Birch, A. (2016, June 9). Edinburgh Neural Machine Translation Systems for WMT 16. arXiv.org. https://arxiv.org/abs/1606.02891

[14] Van Os, R., (2022) Lexical Substitution with Transformers-based Language Models (Doctoral dissertation, tilburg university).

[15] Azizah, S. F. N., Cahyono, H. D., Sihwi, S. W., & Widiarto, W. (2023, August 9). Performance Analysis of Transformer Based Models (BERT, ALBERT and RoBERTa) in Fake News Detection. arXiv.org. https://arxiv.org/abs/2308.04950