Performance Evaluation of Sentiment Analysis on Reddit Comments: Insights and Improvement Opportunities for Naive Bayes, SVM, and BERT Models

Niket Birannavar[1], Priya Prasad[2], Gayatri Dandgal[3], Mandar Ganiger[4],Prathamesh Redekar[5], Veena Badiger[6]
Department of Computer Science and Engineering
KLE Technological University Campus, Belagavi, Karnataka, India.
02fe22bcs058@kletech.ac.in[1], 02fe22bcs073@kletech.ac.in[2], 02fe22bcs037@kletech.acc.in[3],
02fe22bcs051@kletech.ac.in[4], prathameshredekar.mss@kletech.ac.in[5] and
veena.badiger92@gmail.com[6]

***Abstract*—Sentiment analysis plays a key role in Natural Language Processing (NLP). It aims to pull out sentiments, opinions, and emotions from text. Social platforms like Reddit show how important this is. People express many different emotions there. This makes sentiment analysis useful for market analysis, consumer feedback, and social behavior research. This study tackles problems in emotion classification. These include not having enough emotion databases, making things too simple with binary classification, and the high computing costs of processing big datasets like GoEmotions. We use the GoEmotions dataset to test machine learning and transformer-based models. This dataset has 58,000 comments labeled with 27 emotions. We look at Support Vector Machines (SVM), Naive Bayes, and BERT. Our findings show that BERT does better than the others. It's more accurate and captures small differences in sentiment well. BERT's strong performance shows it's a good fit for complex sentiment analysis tasks. This paper stresses the need for accuracy and efficient computing. It sets a standard to help push fine-grained sentiment analysis research forward.**

I. INTRODUCTION

Sentiment analysis is crucial in areas like brand analysis, user profiling and market forecasting where understanding user sentiment on platforms like social media, news websites and online reviews is key. But many existing datasets simplify emotions into basic binary categories which doesn't capture the rich emotional spectrum especially on dynamic platforms like Reddit.

This research aims to improve sentiment analysis capabilities in multilabel emotion categorization and tackle the challenges of interpreting complex emotional nuances that traditional databases miss. BERT, a deep transformer model has shown to be effective in various NLP tasks due to its ability to learn contextual relationships in text. But it struggles to articulate complex emotional states because it tends to oversimplify. Conventional machine learning models like SVM and Naive Bayes are efficient in computation but fall short in intricate sentiment analysis. The limited number of datasets used in previous studies has restricted model performance in real world scenarios especially in more advanced emotion classification beyond just positive or negative sentiments.

This study combines advanced NLP techniques with traditional machine learning models to evaluate the performance of different classifiers in multilabel emotion classification using the GoEmotions dataset. It focuses on transformer-based models like BERT to understand complex sentiments and compare them with standard models like SVM and Naive Bayes. By expanding the dataset and evaluating models based on computational efficiency, accuracy, recall and F1 scores, the paper aims to set new benchmarks for multilabel sentiment analysis and investigate the trade-off between computational cost and model performance.

The paper is organized as follows:Section I provides the challenges of sentiment analysis in Reddit comments, Section II discusses review of related work, SectionIII analyses the GoEmotions dataset and , Section IV describes the models used and Section V analyses the performance of the different models.

II. LITERATURE WORK / BACKGROUND

Sentiment analysis, the art of reading emotions, opinions and attitudes from text, is more important than ever across many industries. From social media monitoring to healthcare and stock market analysis, we need to decode human sentiment from vast amounts of text data more than ever. As this need grew so did the number of ways to do sentiment analysis. Early methods were simple, lexicon based or basic machine learning models. But as we got to see the complexity of human

emotions, we realized those methods were too simplistic to capture the subtleties in text.

Mayur Wankhade *et. al* [1],has written about the evolution of sentiment analysis over the years and the strengths and weaknesses of the methods. The problem was clear: how to detect sarcasm, manage domain specific language and handle multilingual data. Those challenges persisted as traditional methods struggled to understand text in all its forms. But the potential of sentiment analysis to drive insights across industries kept the research going to find better ways to decode sentiment from text.

Over time more advanced methods started to emerge. Birjali *et. al* [2], documented this shift, how hybrid approaches combining lexicon based and machine learning techniques improved sentiment classification. The need for models that could adapt to different domains like healthcare and finance became critical and new technologies started to emerge to meet those demands.

Margarita Rodriguez *et. al* [3],realized the importance of capturing not just the sentiment but the temporal changes in sentiment especially for applications like tracking public opinion during political events or predicting stock market trends. His study "A Review on Sentiment Analysis from Social Media Platforms" investigates the application of sentiment analysis to assess public opinion in fields such as finance, healthcare, politics, and marketing. It highlights the relevance of temporal dynamics and causality, particularly in tracking sentiment changes over time and forecasting outcomes such as stock market patterns or political events. While classic methods like lexicon-based approaches and machine learning are widely used, sophisticated models such as BERT and GPT-3 are underutilized. The study finds a substantial commercial interest, with over 8,000 patents registered, but also identifies a disconnect between academic research and actual corporate uses..

As the field progressed researchers realized that sentiment analysis wasn't just about labeling text as positive or negative. Human emotions were more complex than that. Jingfeng Cui *et. al* [4], explored the combination of machine learning and deep learning techniques to classify multi label emotions. This led to the rise of transformer based models like BERT which can understand the context within a sentence and capture sentiments much more accurately than ever before. But sentiment analysis was not complete. The next frontier was multimodal sentiment analysis which combined text, audio and visual data. The paper provides a detailed analysis of research hotspots, including technological directions and constraints, as well as future research prospects.

Ringki Das *et. al* [5],and Linan Zhu *et. al* [6],have explored this new direction where multiple data streams were used to improve sentiment detection. By combining these modalities the models could detect emotions more accurately especially when context or tone of voice played a key role in understanding the sentiment. Ringki Das *et. al* [5], study demonstrates how complementing data streams from multimodal techniques can improve sentiment analysis accuracy. It highlights trends, difficulties, and potential paths in its discussion of deep learning methods such as transformer-based models and convolutional neural networks (CNNs). The study points up areas that require more investigation in order to develop more reliable sentiment analysis systems by merging several modalities.

And Linan Zhu *et. al* [6], "Multimodal Sentiment Analysis Based on Fusion Methods: A Survey," provides a thorough overview of multimodal sentiment analysis. In order to improve sentiment recognition, the authors emphasize the shortcomings of conventional text-based sentiment analysis and concentrate on integrating text, audio, and video data. Several fusion approaches (early, late, tensor-based, word-level, translation-based), datasets, feature extraction strategies, and difficulties encountered in the field are among the main topics covered. By offering insights into various model frameworks and their uses in fields like public opinion analysis and product reviews, the study seeks to direct future research. Table I summarizes the literature review.

III. DATASET DESCRIPTION

One popular resource for emotion classification tasks is the GoEmotions dataset. It includes 58,000 carefully selected Reddit comments in English that are labeled with a neutral label in addition to a wide range of 27 emotion categories. admiration, amusement, anger, confusion, joy, sadness and others. Comments that don't convey any certain sentiment are represented by the neutral label. Usually, the dataset is divided into three sections: Machine learning models, validation set, test Set Nature of multiple labels: It is a multi-label classification challenge since each statement could be linked to several different emotions. References: Reference is been taken from Reddit platform Here are the 27 emotions in the dataset: Admiration, Amusement, Anger, Annoyance, Approval, Caring, Confusion, Curiosity, Desire, Disappointment, Disapproval, Disgust, Embarrassment, Excitement, Fear, Gratitude, Grief, Joy, Love, Nervousness, Optimism, Pride, Realization, Relief.

Neutral serves as an additional label for emotionally neutral comments. Dataset Structure: Each row includes: Text: A Reddit comment. Labels: One or more associated emotion categories. Applications: Emotion Detection: Understanding emotional expressions in text. Mental Health Monitoring: Detecting signs of depression or stress. Feedback Analysis: where the analysis of customer sentiments has been reviewed. User interface: amplifying the strength of virtual assistants

This GoEmotions data set made a great case for consideration and applies the model to classifications of the emotions in the text.

TABLE I: EVALUATION OF REVIEWED RESEARCH PUBLICATIONS FOR TECHNIQUE

Author	Year	Approach	Advantages	Disadvantages
Mayur Wankhade,A.C.S. Rao, C.Kulkarni *et al.* [1]	2008	Machine learning and Lexicon-based	Provides information for both researchers and beginners	Issues with detecting sarcasm and domain-specialty.
Marouane Birjali, Mohammed Kasri, A. Beni-Hssane*et al.* [2]	2023	Machine learning and lexicon based	Highlights uses in several industries, including as healthcare and corporate intelligence	Difficulties in managing ambiguous text, sarcasm, and multilingual data
Margarita Rodríguez-Ibánez *et al.* [3]	2023	Lexicon-based, SVM, Bert,GPT-3	Limited use of advanced AI technique like transformers.	
Jingfeng Cui et al*et al.* [4]	2023	Keyword occurrence, machine learning and natural language processing	Gives a wide historical review of research area in sentiment analysis	Limited investigation into future integration between developing AI models such as transformers and older approaches
Rinki Das *et al.* [5]	2023	CNN, Transformers-based models	Emphasizes multi-modal over uni-modal techniques	Lacks substantial industry adoption
Linan Zhu *et al.* [6]	2023	Fusion-based models	Enhanced sentiment classification accuracy	Complexity in data alignment and fusion

IV. METHODOLOGY

This paper evaluates sentiment analysis using three methodologies: BERT, SVM, and Naive Bayes. We start with data collection and preprocessing, followed by splitting the data into training and testing sets. Each model is trained and evaluated using metrics like accuracy, precision, recall, and F1-score to compare their performance and highlight their strengths and weaknesses.

A. Naive Bayes and SVM

This paper compares the power of Naive Bayes with SVM for [specify task, e.g, text categorization] using Machine Learning. The dataset was partitioned into three sets namely training, test and validation to ensure fair class representation. The data preprocessing began with the removal of punctuation, followed by the deletion of stop words and the conversion of text to lowercase, thus reducing the noise and enhancing some major features that are good for effective categorization. This has greatly helped track the quality and relevance of the data to machine learning by normalizing it.

TF-IDF vectorization where text representation is basically transformed into numeric formatting by measuring the weight of expressions as stated by their applicability. This technique emphasized distinctive and significant phrases while removing the more generic ones, leading to valuable input for both models. Naive Bayes, using scikit-learn's MultinomialNB, was selected for its efficiency and probabilistic nature. SVM, constructed with the SVC class and a linear kernel, showed strong performance with high-dimensional data. A grid search was conducted to optimize hyperparameters, enhancing the performance of both models.

The models were evaluated based on many benchmarks such as precision, recall, F1-score, and accuracy. Most of the metrics were chosen to compare and determine the performance of models, taking into account correct and wrong classifications. The Naive Bayes classifier registered an accuracy of 29.0 percent and 29.3 percent accuracy in the validation and test sets, respectively. The SVM classifier reached a test accuracy of 15.9 percent and a validation accuracy of 15.8 percent, respectively. It was discovered that the accuracy of Naive Bayes exceeded the SVM in this task that shows its simplicity and efficiency in handling textual data. The relatively low accuracy of both models point to potential issues that could arise with feature representation and multi-label classification.

The test results suggest that the Naive Bayes algorithm is effective for the text classification task at the baseline level, but SVM's performance indicates that further work on a feature-engineering issue or hyperparameter tuning may be necessary. Forthcoming research is likely to engage in more advanced techniques, like deep learning or ensemble methods, which aim to make the classification problem less error-prone. The outcomes of the present research can be taken as the initial step to know in what way machine learning methods can be transformed into text data analysis, and consequently a good ground for the conduction of future research on the subject, will allow a myriad of exciting results to be brought to light.

B. Bert

This study applies several machine learning algorithms for text classification such as: Naive Bayes, SVM and BERT. The dataset was cleaned by performing text normalisation, which included lower casing and stop word removal. The models on other hand were trained by employing vectors extracted using TF-IDF for Naive Bayes and SVM, while BERT utilized a pre-trained embedding for the improvement of text representation.

The accuracy level obtained on test set for the Naive-Bayes model was 29.3 percent, while the SVM model achieved only 15.9 percent. This shows the limitations of the standard techniques on this dataset. On the other hand, however, the BERT model managed to outperform the other two traditional techniques with an accuracy of 74.1 percent, proving that it is better able to make sense in the context of language semantics.

All these models were constructed with the help of python using scikit-learn and hugging face transformers. These highlights the importance of sophisticated techniques such as BERT in improving accuracy for complex text classification tasks.

V. RESULT AND ANALYSIS

The results and analysis section compares the performances of BERT, SVM, and Naive Bayes models for sentiment analysis. This section provides detailed metrics including accuracy, precision, recall, and F1-score for each model. Through such comparisons, we identify the best approach for the task of sentiment analysis and comment on the implications of our results.

A. SVM

This analysis's goal was to use a Support Vector Machine (SVM) model to categorize text input into pre-established emotional groups. The dataset was divided into training (147,857 samples), validation (31,684 samples), and test (31,684 samples) sets. It contained 28 emotion labels. On the test set, the SVM model's accuracy was 15.91%, while on the validation set, it was 15.78%. The test set showed comparable results to the validation set, which had average precision, recall, and F1-scores of 56%, 17%, and 23%, respectively.

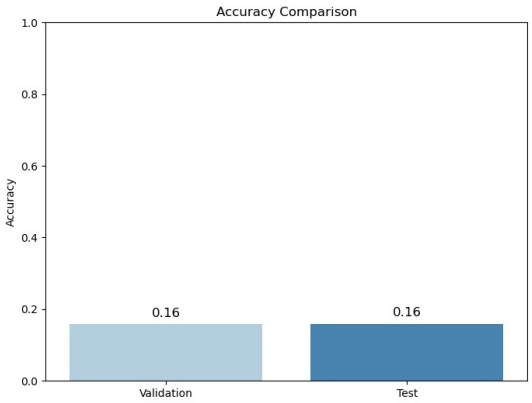

Fig. 1: Validation and Test Accuracy comparision

Significant differences existed in performance across labels, with dominating categories like "neutral" having higher precision (Precision = 89%, Recall = 72%, F1 = 79%). The unbalanced distribution of labels may have contributed to the low recall of the majority of labels, which showed that it was difficult to capture all real cases. Given the intricacy of the multi-label classification job, the SVM model's overall performance was moderate. Addressing label imbalance, experimenting with kernelized SVMs, and fine-tuning hyperparameters could all lead to improvements. Comparisons with more sophisticated models, including ensemble approaches or neural networks, may also be a part of future research.

B. Navie Bayes

The purpose of this investigation was to assess the Naive Bayes model's ability to categorize text input into pre-defined emotional groups. The dataset was separated into three sets: training (147,857 samples), validation (31,684 samples), and testing (31,684 samples), each with 28 emotion labels. On the test set, the Naive Bayes model achieved an accuracy of 29.3%, while on the validation set, it was 29.0%.

These results indicated that the model was stable across all the data sets as well.The Naive Bayes model had the average precision, recall and F1 scores of 60%, 35% and 44% respectively. Performance was much higher for the dominating categories such as 'neutral' (precision = 82%, recall = 75%, F1 = 78%). However, the model had a hard time predicting the less frequent categories which resulted in low recall values for all the minor categories. This is due to the fact that Naive Bayes is a very weak model, which makes it a good baseline for text categorization, but due to its independence assumptions, it is not capable of capturing more complex patterns in the data. There are certain enhancement which can be made in the future to avoid these flaws such as using feature engineering techniques, ensemble methods or using some other advanced classifiers to overcome the limitations of the model. The changes made in this way could help in enhancing the overall accuracy and ensure that no label is left unclassified in between.

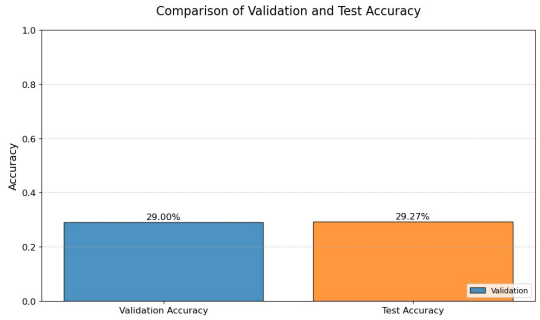

Fig. 2: Validation vs Test Accuracy

C. BERT

The goal of this work was to assess how well a BERT based model can classify text into 28 predefined emotional categories. The dataset was divided into three subsets: validation (31,684 samples), testing (31,684 samples), training (147,857 samples).

For each subset, the model's generalization capacity and it's robustness on different data distributions was evaluated. The BERT model reached an accuracy of 74.07% on the validation set on the task of emotion classification tasks, demonstrating its aptitude to cope with the intricacies of emotion classification tasks.

In terms of evaluation metric, the model achieved F1-score (Micro) : 81.59% and F1-score (Macro) : 74.37%, which

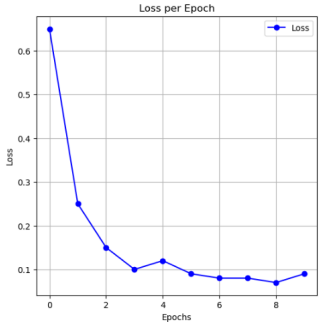

Fig. 3: Loss Reduction over Epochs

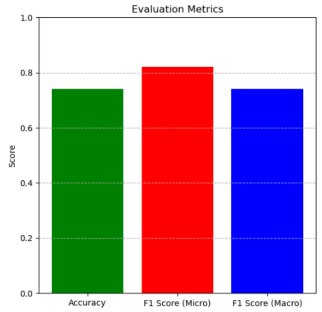

Fig. 4: Performance Metrics

are excellent on the overall performance and balancing the class. We observed particularly good performance on dominant categories, like 'neutral,' in which precision and recall were much higher than for the contrary categories. For the minority classes, the model's recall was lower highlighting the recurring problem of class imbalance in multi-class emotion classification tasks.

To show that pre trained transformer architectures can extract intricate linguistic patterns and semantic nuances within the dataset, the superior performance of the BERT based model is demonstrated. We find BERT to be a substantial improvement in accuracy and F1 scores over traditional models, such as Naive Bayes and SVM and in turn a strong candidate for using in emotion detection and similar tasks.

The success, however, falls short of bridging the performance gap between the dominant category and the one that is minority. Methods like oversampling minority classes, class weighted loss functions or usage of data augmentation methods can solve the imbalance problem and have more balanced performances for all categories. Moreover, it is additionally possible to obtain enhanced accuracy and representation of under represented classes using ensemble approaches that combine BERT with other models. Further future research would include investigating task specific fine tuning and domain adaptation to further extend the boundaries of the model.

VI. CONCLUSION

In this study, we compared three sentiment analysis models—SVM, Naive Bayes, and BERT—using the GoEmotions dataset. While Naive Bayes and SVM are traditional models, BERT represents a more advanced and effective approach in natural language processing. In this comparison, BERT performed significantly better than both SVM and Naive Bayes, offering superior performance overall. This had problems with dealing with large data sets, mainly due to their inability to capture the contextual and nuanced semantic relationships that are inherent in the text. BERT enables the understanding of context and therefore can deal with larger data, which supersedes in tasks of sentiment analysis. These results show that BERT is more appropriate for the complex problems posed by sentiment analysis, particularly when dealing with large volumes of information. However, it still requires research on future endeavors into developing and refining models that can capture more subtle emotions, deal with varying languages, and use multimodal data.

VII. FUTURE SCOPE

Future research should shed further light on the possibility of combining BERT with other models like ensemble techniques to improve performance and handle scalability in a much more better and effective way. In other words, the computational challenges that would arise in the deployment of such advanced models on large-scale datasets are definitely one promising area that must be further explored for sentiment analysis.

REFERENCES

[1] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731–5780, 2022.

[2] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021.

[3] M. Rodríguez-Ibánez, A. Casánez-Ventura, F. Castejón-Mateos, and P.-M. Cuenca-Jiménez, "A review on sentiment analysis from social media platforms," *Expert Systems with Applications*, vol. 223, p. 119862, 2023.

[4] J. Cui, Z. Wang, S.-B. Ho, and E. Cambria, "Survey on sentiment analysis: evolution of research methods and topics," *Artificial Intelligence Review*, vol. 56, no. 8, pp. 8469–8510, 2023.

[5] R. Das and T. D. Singh, "Multimodal sentiment analysis: a survey of methods, trends, and challenges," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–38, 2023.

[6] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Information Fusion*, vol. 95, pp. 306–325, 2023.