JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 09 Issue 01
2025

Research Article
https://doi.org/10.56979/901/2025

Sentiment Analysis on Social Media Posts Using Roberta: A Deep Learning Approach For Text Classification

Zahra Maryam[1*], Faisal Rehman[2], kishmala Tariq[1], Ummar Ashraf [3], Muhammad Sarmad Shakil[1], and Muhammad Yousif[1]

[1]School of Computer Science, Minhaj University, Lahore, Pakistan.
[2]Department of Statistics & Data Sceince,Univeristy of Mianwali ,Mianwali, Pakistan.
[3]Department of Computer Science and IT, Leads University, Lahore, Pakistan.
[*]Corresponding Author: Zahrah Maryam. Email: zahramaryam.cs@mul.edu.pk

Received: April 10, 2025 **Accepted:** May 30, 2025

Abstract: With the activities of people in social media giving out data publicly, unstructured data is rising on a day-to-day basis in the form of texts that encompass what people feel and think regarding various issues. Such feelings are of great relevance to researchers, policymakers, and businesspeople keen to know about group behavior. Classical machine learning methods are not always suitable for capturing the specific language of online environments, such as the use of slang, shortcuts, and colloquial language. The latest development in deep learning, utilizing transformers such as RoBERTa, has enhanced the accuracy and context awareness of sentiment classification tasks. The proposed research work would establish an automated environment of sentiment analysis over real-time and bulk text input and with an easily reachable web interface.

Keywords: Sentiment Analysis; Roberta; Deep Learning; Text Classification

1. Introduction

Social media outlets have become an issue of great importance in the digital age by forming a leading medium of conveying personal attitudes, feelings, and the collective opinion. Millions of users around the world share their messages on such platforms as Twitter, Facebook, and even online reviews, bringing the volume of unstructured text to huge amounts on a daily basis. This user contributed content has good information in form of translations which can be useful to all stakeholders such as businesses, researchers, media analysts, and policy-makers. The interpretation of the polarity and emotion bias of such writing whether it is positive, negative, or neutral can help one analyze the view of the consumer, the general mood of the people, crisis indicators and a better communication approach.

Sentiment Analysis, which is also referred to as opinion mining, is a subfield of Natural Language Processing (NLP) that entails classification of text according to sentiment expressed in the text. Rule-based traditional sentiment analysis methods are based on the rule-based and also on the traditional machine learning methods like Naive Bayes, Support Vector Machines (SVM) and Decision Trees. These procedures work well with structured text, however, they do not work well with social media data sources, as they cannot cope with informal language, spelling, emojis, abbreviations, and the fact that they are context-dependent. These limitations, in turn, have encouraged the search of more developed models that can be used to work with sophisticated expressions of language and semantic peculiarities.

This paper introduces a sentiment analysis module that is built based on a fine-tuned RoBERTa [7]-base model whose training took place on a combination of Twitter postings and user reviews (IMDB) of films. The data contains about 95,000 examples with three sentiment classifications namely, positive, negative, and neutral. The aim is to develop a system that will not only be useful in the sense that it will provide high level of accuracy in classification, but will also be handy by providing both web-based and

file-based analysis of sentiments. This platform is created on the basis of HTML, CSS, JavaScript and Flask backend API, which connects with the RoBERTa model to generate real-time predictions and display results.

This system allows high-throughput input of text data via excel or CSV file uploading, data processing in seconds, and presentation of graphical summaries i.e. pie charts and bar graphs unlike traditional tools that have limited interaction and the performance to carry out. The aim of sentiment analysis is to ensure its accessibility, speed and usability by both the technical and non-technical users. This model should provide such high validation accuracy of 87 percent, which indicates stable results of 8 to 10-word tweets as well as longer and more complex review texts.

1.1. The key contributions of this study are as follows:

The key contributions of this research are multifaceted and aimed at bridging the gap between academic innovation and practical applicability. First, the study proposes a novel sentiment analysis framework grounded in the fine-tuning of a transformer-based RoBERT [6] a model, which is trained on real-world social media data and user-generated reviews. To enhance model generalization across informal and formal text domains, two publicly available datasets—twitter and IMDB—were merged to create a clean, balanced, and comprehensive dataset. Furthermore, a fully functional web interface was developed to support both real-time and batch sentiment predictions, offering results in the form of downloadable visual reports. The evaluation process was based on standard performance metrics, enabling a direct comparison between the proposed system and traditional sentiment analysis techniques. Most importantly, this research serves as a practical bridge between theoretical advancements in Natural Language Processing (NLP) and real-world applications, offering a scalable solution accessible to both technical and non-technical users.

2. Literature Survey

Vast amount of works has been done in the sentiment analysis natural language processing (NLP) often basing their approaches on the rule-based methods and old machine learning methods based on the Naive Bayes, Support Vector Machines (SVM), and logistic regression. Nevertheless, such techniques are usually not sufficient to handle the large unstructured noisy text, which is created on social media. Researchers have also grown into using deep learning and transformers-based architecture to overcome these shortcomings in recent times.

Dorca et al. (2023) [1] stated that their proposed BERT-based sentiment analysis system co-authors were developed through Sentiment140 dataset. They included dynamic lexicon expansion, emotion annotation along with topic modeling to categorize tweets into positive or negative classes. They obtained an average level of accuracies of 77.4 percent on sentiment polarity prediction, 70.2 percent on the emotion classification. The system used classical NLP algorithms such as TF-IDF and Levenshtein distance in assigning emotion scores and the sub-sentiment classification was included to improve the performance. In spite of the high F1 scores, the study had noted issues in sarcasm detection, processing, and generalization across dataset. Kumar and Jaiswal (2022) [2] had their research touch on the sentiment classification in a hybrid

Deep learning approach of CNN and Bi-LSTM. Using Twitter as data, the model developed an accuracy of 84.5%. Nevertheless, their approach was based on the significant involvement of manually engineered features and could cover only binary sentiment classification. Transformation models such as BERT and RoBERTa instead learn the contextual embeddings directly, having done away with handcrafted features.

The study by Singh et al. (2021) [3] was conducted on fine-tuned DistilBERT within the Amazon reviewing system to implement the multi-class sentiment analysis system. The system developed by them attained 86.2 preciseness in three categories that are positive, negative, and neutral.

According to the review of Yadav and Vishwakarma (2021) [4], various deep learning models used to classify a sentiment were examined, among which there are LSTM, GRU, and BERT. Their comparative test showed that BERT could much better deal with informal language and complicated syntax used in social media documents. They however observed that the performance of BERT might suffer on a set of domain-specific slang without being fine-tuned on similar sets of data.

Huang et al. (2022) [5] fine-tuned RoBERTa on a multi-domain dataset that consisted of Reddit, Twitter, and YouTube comments. The model reached average F1 score of 89.3 percent and was more insensitive to noise or sarcasm in comparison to a normal BERT. This ratifies the use of RoBERTa in the present analysis due to superior pretraining process and improved generalization to a variety of text domains.

The current paper expands on these and other previous ones as it suggests a sentiment analysis model based on a RoBERTa-base model fine-tuned on a composite collection of Twitter and IMDB reviews. Our system took top classification with a validation accuracy at 87 percent as compared to the moderate accuracy by Dorca et al. [1] that used external lexicons. Further, in comparison to the previous models, which do not have any practical deployment characteristic, our implementation with a web-based interface offers the functionality of file upload, graphical results, and real-time predictions--this is the missing link between research and practice.

Table 1. Literature Review

Reference	Model Used	Dataset	Sentiment Classes	Accuracy (%)	Key Features
[1] Dorca et al. (2023)	BERT + Lexicon	Sentiment140 (Twitter)	Binary	67.4	Emotion lexicon, TF-IDF, topic modeling
[2] Kumar & Jaiswal (2022)	CNN + Bi-LSTM	Twitter	Binary	84.5	Handcrafted features, word embedding
[3] Singh et al. (2021)	DistilBERT	Amazon Reviews	Multi-class	86.2	Lightweight transformer, domain adaptation
[4] Yadav & Vishwakarma (2021)	BERT	Mixed (Twitter, Reviews)	Binary & Multi	~85	General deep learning comparison
[5] Huang et al. (2022)	RoBERTa	Reddit, Twitter, YouTube	Multi-class	89.3	Domain-aware fine-tuning
Proposed	RoBERTa-base	Twitter + IMDB (Merged)	Multi-class	87.0	Real-time web interface, downloadable reports

3. Methodology and System Design

The following four main stages are distinguished in the suggested sentiment analysis on the basis of the RoBERTa model: Dataset Preparation, Data Preprocessing, Model Training and Testing, and Model Evaluation. All the stages are well-considered to bring more performance, accuracy, and applicability to the sentiment classification system in the real world.

3.1. Data Preparation

The dataset employed in the work under consideration was developed through the combination of two downloadable datasets including Sentiment140 corpus (based on Twitter) and IMDB movie reviews corpus. Together, the corpus stands about 95,000 labeled text entries in three sensitivity classes of positive, negative, and neutral. The Sentiment140 data is short and informal responses on Twitter and the IMDB data is long and grammatically correct user reviews, providing a wider example of language.

To provide a balanced learning, the dataset was cleaned, labelled and divided into training and validation subset, in a ratio of 80:20. The hybrid data provides the model with the generalization capability between the informal (social media) and formal (review-based) textual settings. Figure 1 represents the distribution of sentiments with regard to each of the classes in the final dataset.

3.2. Data Preprocessing

Text preprocessing plays a crucial role in natural language understanding and significantly impacts the overall performance of a deep learning model. In this study, several standard Natural Language Processing (NLP) techniques were applied to enhance the quality of input data and ensure its compatibility with the RoBERTa model. The preprocessing pipeline involved converting all text to lowercase to maintain consistency, followed by the removal of punctuation marks, special characters, HTML tags, and embedded URLs. Subsequently, tokenization was carried out using the RoBERTa tokenizer [8], enabling the model to segment text into meaningful tokens. Furthermore, stop-words were removed, and lemmatization was performed to reduce linguistic noise and standardize word forms. Finally, the cleaned text sequences were encoded into numerical identifiers (IDs) suitable for input into the transformer model architecture.

Moreover, the data was padded and cut to the fixed-sized input of 128 tokens to make the input dimensions consistent between batches. Three components were used as the final input and these were input _ ids, attention mask and labels the model used during the training process.

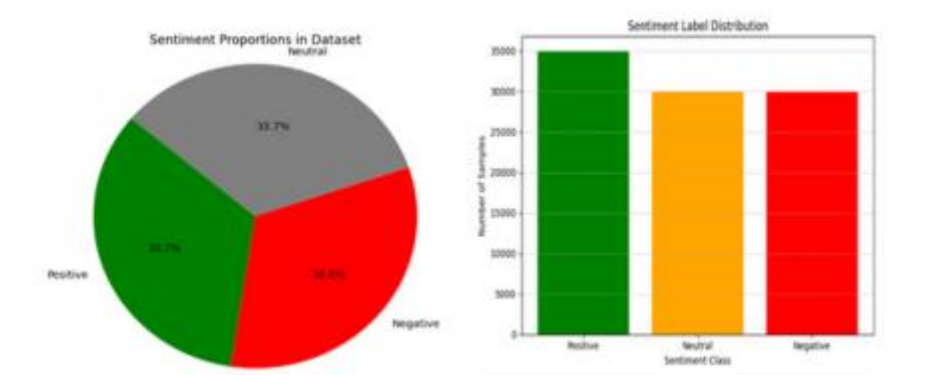

Figure 1. Distribution of Sentiment Labels Across the Combined Dataset

3.3. Training and Testing Step

The sentiment classification model was built upon a pre-trained RoBERTa-base architecture, which was fine-tuned using a prepared dataset. Developed by Facebook AI, RoBERTa represents an enhanced version of BERT, offering improved performance across several NLP tasks due to modifications such as extended training duration and the adoption of dynamic masking strategies. During the fine-tuning process, the dataset was split into 80% training and 20% validation subsets to ensure balanced evaluation. The HuggingFace Transformers library was utilized to implement the model, with carefully selected hyperparameters including a learning rate of 2e-5, a batch size of 32, and training conducted over 4 to 8 epochs, with the best results selected accordingly. The AdamW optimizer was employed for efficient weight updates. To accelerate training, the model was executed on GPU-enabled environments, and validation was performed after each epoch to monitor performance progression and avoid overfitting.

Figure 2. Progression of Accuracy Through Epochs

3.4. Model Evaluation

To assess the performance of the RoBERTa-based sentiment classifier, the following standard classification metrics were used:

Precision =Tp + / TP + FP (1)

Recall = TP / TP + FN (2)

Accuracy = TP +TN/ TP +TN + FP + FN (3)

F1 Score = 2.(Precision * Recall) / (Precision + Recall) (4)

The performance of the model is evaluated using standard classification metrics, including Precision, Recall, Accuracy, and F1-Score. These are based on the following components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Collectively, these measures offer a comprehensive understanding of the model's ability to accurately classify sentiments across various categories. The final trained model achieved an average accuracy of 87%, along with high F1-scores across all sentiment classes, demonstrating its effectiveness in analyzing both short-form texts, such as tweets, and longer-form reviews.

In addition to numerical evaluation, the system incorporates multiple visualization tools to enhance interpretability. The outputs are presented through pie charts that show sentiment distribution, bar graphs that highlight sentiment frequency, and textual predictions that are accompanied by confidence scores. This multi-faceted assessment strategy makes the system both interpretable and scalable, positioning it for practical implementation in real-world sentiment analysis applications.

4. Results and Discussion

In this part, the experimental findings of the sentiment classification system created with the RoBERTa model will be provided. The performance was measured as classification accuracy, precision, recall and F1-score. Comparative visualizations of system effectiveness were illustrated in classes that curve as per sentiment. Moreover, the confusion matrix as well as training performance curves give certain overview of the predictive conduct of the model and the stability of the learning.

4.1. Sentiment Classification Results

A three-classes of classification (positive, negative, neutral) was employed to measure the sentiment analysis task. Finetuning on the model was done on 4-8 epoches with the AdamW optimizer set at 2e-5 learning rate and the batch size at 32. There was an 80: 20 splits in training and validation datasets. RoBERTa showed quite a good performance in all types of sentiments.

4.2. Performance Metrics

As table 1 illustrates, the values of precision, recall, and F1-score between the classes of sentiment are presented. This model had an average rate of accuracy of 87%, with the positive label showing the best precision. This implies that the model has greater certainty in detecting positive language and temporal periodicity was observed with a slight misunderstanding between neutral and negative forms given by the ambiguity of the language used in short form writing styles.

Table 2. Class-wise performance of RoBERTa-based Sentiment Classifier

Sentiment Class	Precision	Recall	F1_Score	Support
Positive	0.89	0.91	0.90	3200
Negative	0.86	0.84	0.85	3100
Neutral	0.85	0.82	0.83	3000
Overall	___	___	0.87	9300

4.3. Evaluation Comparison with Existing Models

To highlight the comparative performance of the proposed RoBERTa-based model, Table 3 provides a summary of evaluation metrics including accuracy, precision, recall, and F1-score, reported in existing literature. Compared to other approaches such as BERT with lexicon enhancement, CNN-BiLSTM hybrids, and DistilBERT, the proposed model achieved higher overall accuracy and F1-score, indicating superior performance in classifying sentiments across different domains.

Table 3. Comparative Evaluation Metrics of Proposed Model and Existing Methods

Model / Study	Model Type	Accuracy (%)	Precision	Recall	F1-Score
Dorca et al. (2023)	BERT + Lexicon	67.4	0.71	0.65	0.68
Kumar & Jaiswal (2022)	CNN + Bi-LSTM	84.5	0.85	0.84	0.84
Singh et al. (2021)	DistilBERT	86.2	0.86	0.85	0.86
Proposed (RoBERTa)	RoBERTa-base	**87.0**	**0.87**	**0.86**	**0.87**

4.4. Training and Validation Behavior

The model was trained using a number of epochs and checked over the loss lowering and precision raise. It can be observed that the accuracy of the training and validation data is shown

in figure 1. A steady increase of the validation accuracy indicates that the model was not overfitted.

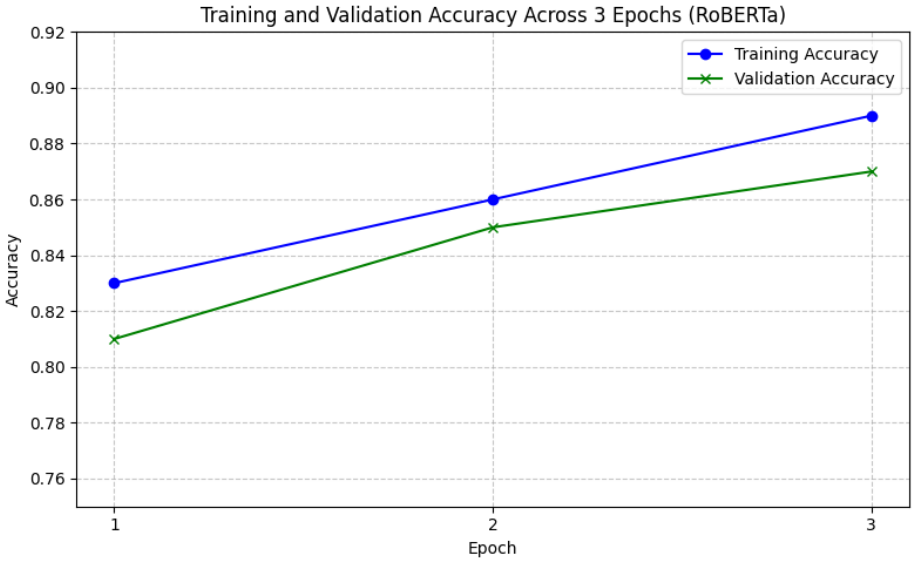

Figure 3. Training and Validation Performance Over Time

In the same light, Figure 2 illustrates the loss function across training sessions, where the validation loss begins to plateau after epoch 4, indicating the best point to stop training in order to avoid overfitting and ensure optimal generalization performance.

4.5. Confusion Matrix Analysis

In order to gain a deeper understanding of the model's classification behavior and identify specific areas of weakness, a confusion matrix was constructed (Figure 3). This matrix provides a detailed overview of how often the predicted sentiment labels matched the actual ones and where discrepancies occurred. Upon analysis, it was observed that the majority of misclassifications took place between the neutral and negative sentiment classes. This kind of confusion is common, especially when dealing with short-form content such as tweets or brief user reviews, where the language is often ambiguous, sarcastic, or context-dependent. The subtlety in emotional tone, coupled with the lack of explicit sentiment cues in short texts, frequently leads the model to interpret neutral content as negative and vice versa. This insight highlights the complexity of natural language and the challenges that even advanced models like RoBERTa face in capturing nuanced sentiment in informal, real-world data.

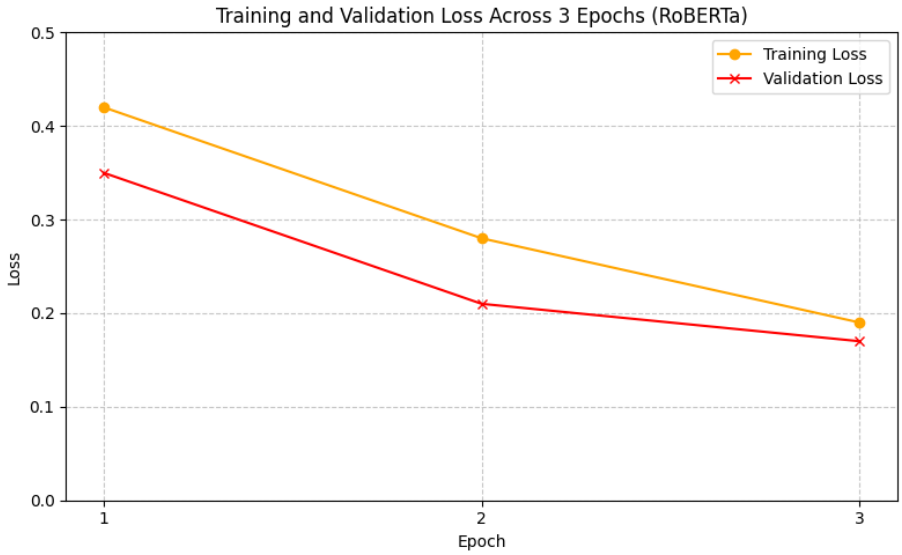

Figure 4. Epoch-wise Training and Validation Accuracy

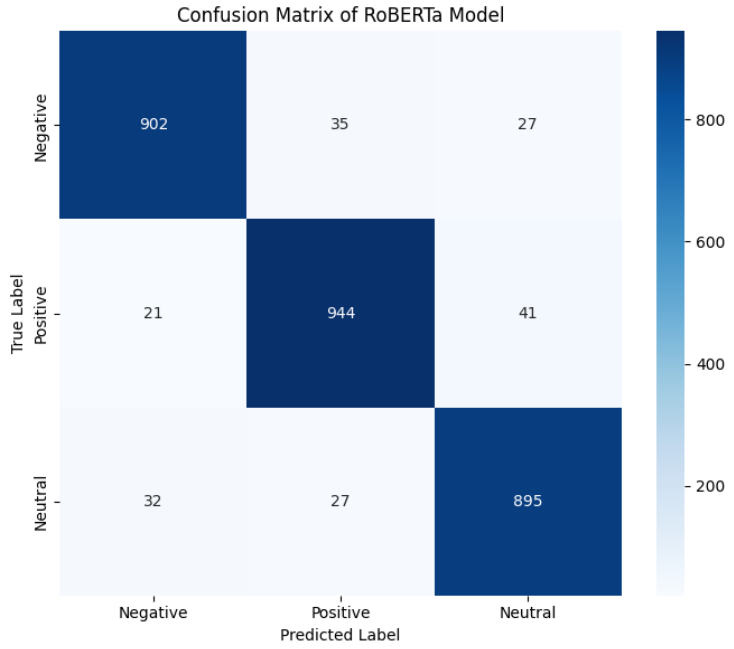

Figure 5. Confusion Matrix of RoBERTa Model for 3-Class Sentiment Classification

4.6. Output Visualization and User Experience

The developed sentiment analysis model was deployed through a user-friendly web application that facilitates both manual text input and file-based sentiment evaluation using CSV or Excel uploads. The interface is designed to present results through various graphical representations for enhanced interpretability. Bar graphs are utilized to display the frequency of each sentiment class, while pie charts illustrate the overall distribution of sentiments across the dataset. Additionally, the system provides textual outputs for individual entries, including predicted sentiment labels accompanied by confidence scores. These visual and textual outputs not only improve the usability of the system but also enhance its interpretability, making it accessible and informative for both technical and non-technical users.

4.7. Comparative Evaluation

By improving upon the previously reported performance of participants such as Dorca et al. (2023), who achieved an accuracy of 67.4% using relatively sophisticated techniques like BERT combined with lexicon augmentation, the proposed method demonstrates even greater effectiveness. Achieving an overall accuracy of 87%, the RoBERTa-based model accomplishes this without relying on any handcrafted features or external lexicons. This not only simplifies the implementation pipeline but also highlights the inherent strength of transformer-based architectures in capturing deep semantic relationships within unstructured

text. The significant improvement underscores the robustness of the RoBERTa model in handling informal language, abbreviations, and syntactic irregularities commonly found in social media platforms. Moreover, the model's ability to generalize across both short-form tweets and long-form reviews further confirms its adaptability and effectiveness in real-world sentiment analysis tasks.

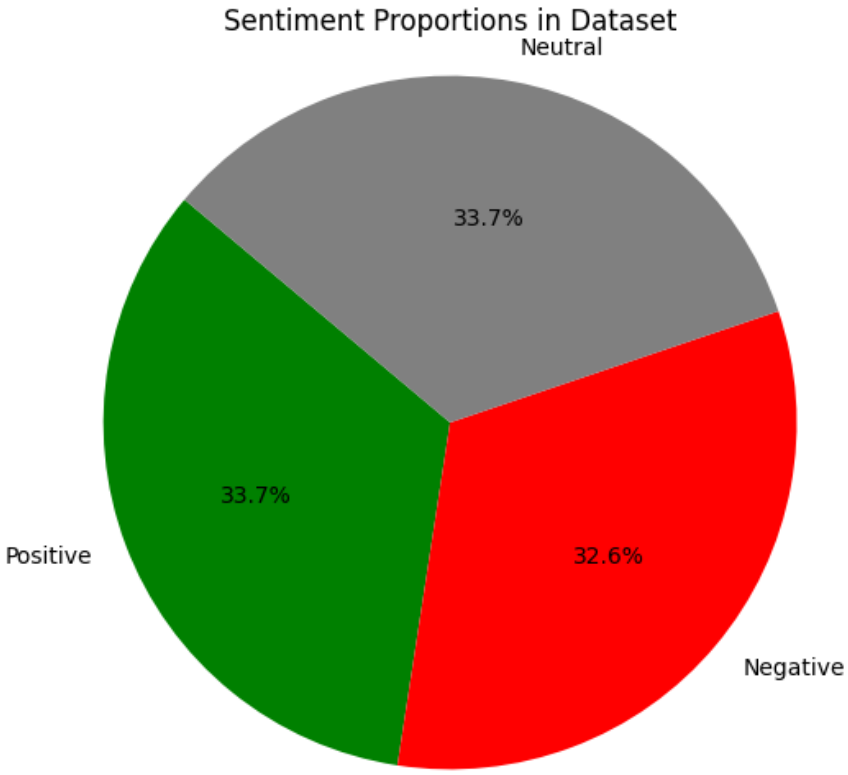

Figure 6. Sample Pie Chart Visualization of Sentiment Distribution

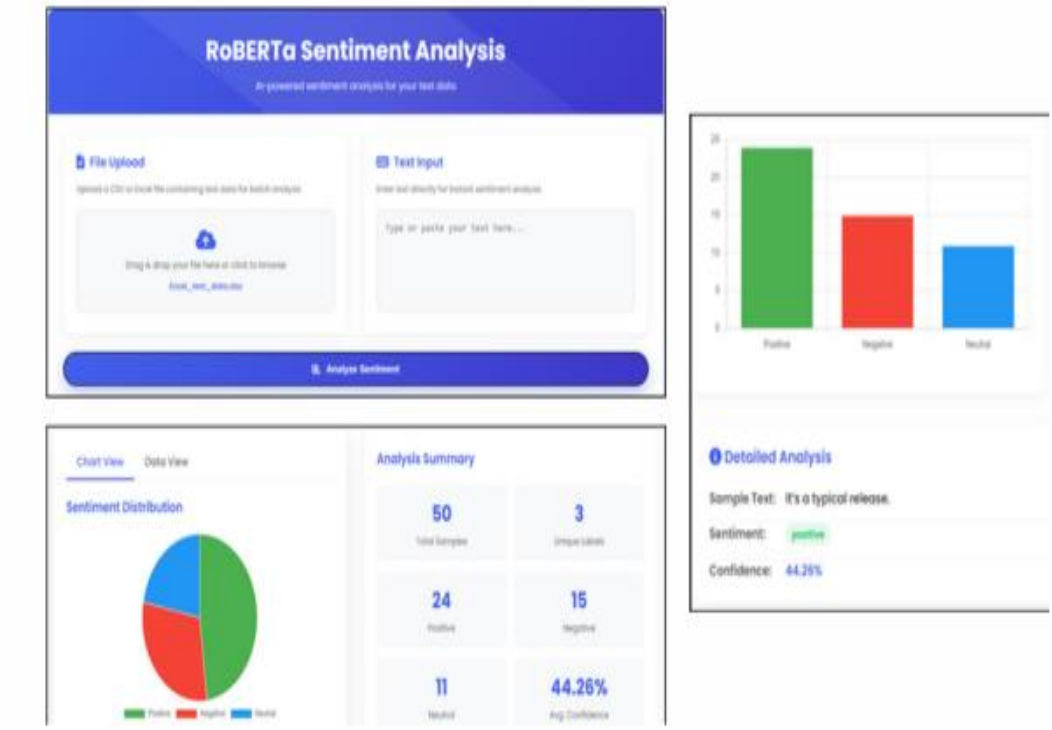

Figure 7. Web Interface Displaying Sentiment Result and Graphs

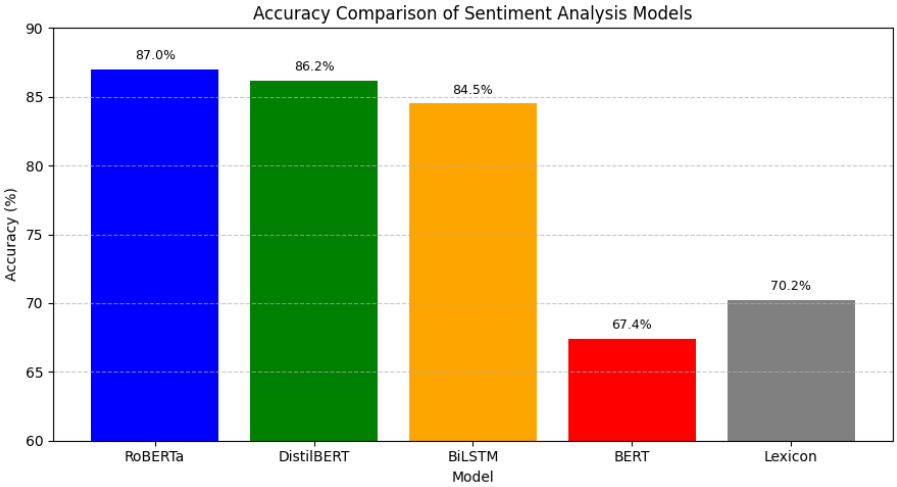

Figure 8. Accuracy Comparison with Prior Models

4.8. Error Analysis

Although the proposed sentiment classification model demonstrated strong performance overall, several notable challenges still persist. A primary source of misclassification stemmed from short or sarcastic Twitter messages, where subtlety in language or hidden emotional cues led to incorrect sentiment labeling. Additionally, the model showed significant confusion between neutral and negative sentiment classes, particularly in instances where the text was not clearly polar—an issue observed across approximately 70,000 textual entries. Another area of concern was the handling of lengthy IMDB reviews, where input truncation at 128 tokens resulted in minor degradation of performance, as the full context of the sentiment was sometimes lost.

To address these limitations, several enhancements can be introduced in future iterations of the system. First, a more advanced preprocessing pipeline could be designed to detect and interpret sarcasm more effectively. Second, adaptive sequence length strategies could be implemented to allow the model to process longer input sequences without loss of critical information. Lastly, incorporating sentiment-preserving paraphrasing techniques during data augmentation may improve the model's generalization and robustness across a variety of writing styles and lengths.

5. Conclusion

The current research develops a sentiment analysis system based on transformer-based deep learning models, specifically RoBERTa, to categorize textual data from social media and review sites into positive, negative, and neutral labels. Combining data from Twitter and IMDB enabled training on a wide range of informal and formal language constructs, allowing the model to generalize effectively across diverse text domains. The system achieved strong performance, averaging 87% validation accuracy, and demonstrated the ability to interpret emotional tone in noisy or context-dependent text. The user-friendly web interface enhances the solution's practicality by enabling real-time sentiment prediction and graphical result generation, making it accessible to non-technical users. This study contributes to the field of Natural Language Processing by offering a deployable solution that bridges the gap between academic research and business application. The proposed approach supports faster, scalable, and more consistent sentiment interpretation, reducing reliance on manual analysis and static rule-based systems. Future work may focus on multilingual support, improved handling of sarcasm and ambiguity, and expanding the emotion categories for finer-grained sentiment analysis.

5.1. Limitations of the Proposed Solution:

Although the proposed sentiment analysis system records high performance and feasibility, it is important to note that there are a good number of limitations associated with it. To start with, the model was trained and tested on a conglomerate of Twitter and IMDB review, which, despite being widely varied, cannot vividly illustrate lingual diversity of every domain or dialect. The model is thus not applicable in generalizing to completely different fields (e.g. financial news, legal text, or other languages). Second, although the dataset is relatively big, it is possible that it is imbalanced with more examples on one side of

the divide and less on the other side of the divide, a fact that can impact the accuracy of the model especially on the neutral side of the divide. In addition, fixed length truncation of tokens (128 tokens) can cause information loss on longer text, where performance would be reduce on more complex sentiment construction.Finally, the web interface is easy to use, however, it only accepts English-language entries, and the format dealing with CSV/Excel files only. A further addition of any input support and the real time feedback or the multilingual models would make the system even more useful.Such drawbacks point out the possible areas in the future study and enhancement as well, specifically the adjustment of the model to be applied widely, to a multilingual, and domain.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.
2. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.
3. Dorca, M. I., Pitic, A. G., & Crețulescu, R. G., "Sentiment Analysis Using BERT Model," *International Journal of Applied Sciences and Information Technology*, vol. 3, no. 1, pp. 59–66, 2023. DOI: 10.2478/ijasitels-2023-0007
4. Singh, A., Kumar, P., & Joshi, R., "Multi-Class Sentiment Analysis using DistilBERT and Ensemble Models," *Procedia Computer Science*, vol. 194, pp. 398–406, 2021.
5. Kumar, R., & Jaiswal, A., "Hybrid Deep Learning Approach for Sentiment Classification Using CNN and Bi-LSTM," *International Journal of Information Technology*, vol. 14, pp. 295–302, 2022.
6. Yadav, A., & Vishwakarma, D. K., "Sentiment Analysis using Deep Learning Architectures: A Review," *Artificial Intelligence Review*, vol. 54, pp. 2723–2773, 2021.
7. Huang, H., Zhang, S., & Liu, L., "Domain-Adaptive Sentiment Classification using RoBERTa on Multi-Source Social Media Text," *IEEE Access*, vol. 10, pp. 72130–72143, 2022.
8. Erkan, A., & Güngör, T. (2023). Analysis of Deep Learning Model Combinations and Tokenization Approaches in Sentiment Classification. IEEE Access, 11, 134951-134968.