Explainable AI For Sentiment Analysis on Movie Reviews

Bhargav Sawant[1], Muiz Tanki[2], Omkar Pawar[3] and Mr. Sunil Patil[4]
[1,2,3]*UG Scholar, Department of Electronics and Telecommunication, K. J. Somaiya Institute of Technology, Sion*
[4] *Professor, Department of Electronics and Telecommunication, K. J. Somaiya Institute of Technology, Sion*

***Abstract*—Sentiment analysis is widely used to interpret public opinion in sectors like shopping, social media and finance. The old ways of doing it have trouble with things like sarcasm, irony and figuring out what words mean in different contexts, which means they do not always give the right answers when looking at real data. This paper is about a way of doing sentiment analysis that combines a lot of different models. 50 Flatten-based networks, 50 LSTM models, 65 GRU models, 5 BERT transformers and 5 RoBERTa transformers. To get a better understanding of what people mean when they write movie reviews. Each and every model was trained separately on a set of movie reviews from IMDB, which has 40,000 labelled reviews. The text was cleaned up using preprocessing methods. Prepared for the models to use. Then the predictions from all models were combined using a stacking ensemble technique, and an XGBoost Stacking was the best at making accurate predictions, with an accuracy of 89.86%. To test and explain the model's working, we used SHAP and LIME methods, which help us understand what is important in the text and how each prediction was made. The results show that our new way of combining models is better and more reliable than old ways and works better in different situations. This work provides a base for doing sentiment analysis in multiple languages and for using it in real-time applications.**

***Index Terms*—BERT, Ensemble Learning, Explainable AI, IMDB Dataset, LSTM-GRU, RoBERTa, Sentiment Analysis, SHAP**

I. INTRODUCTION

In the current digital era, understanding what people think and feel is very important for businesses. This is where sentiment analysis comes in a picture, it helps organizations figure out what people are saying about the organizations by looking at lots of text data from various sources. The organizations use sentiment analysis to understand what people think about the organizations. Sentiment analysis is used for making informed decisions, providing personalized recommendations and managing brand reputation in areas such as e-commerce, social media, financial markets and entertainment industries.

Movie reviews are an example of where sentiment analysis can be applied. They are complex. It contains sarcasm, irony, idioms and specialized terms which make it difficult for machines to understand. Traditional machine learning models and deep learning models both struggle with language and nuances. They do not understand meaning, sentiment, emotions and changes in language patterns. Recent transformer models like BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (Robustly Optimized BERT Pre-training Approach) have enhanced understanding of context meaning and semantics. But basically, they are like black boxes, which don't provide clear explanations. Ensemble methods use a limited set of models, which results in poor performance with noisy or unclear reviews.

This paper proposes a hybrid ensemble framework. It combines 175 deep learning models, including Flatten-based networks, LSTM models, GRU models, BERT and Roberta variants. These models are trained on the IMDB movie review dataset. The base models capture features such as sequential dependencies, semantic richness and pattern recognition. The Predictions of all models are combined using multiple stacking ensemble methods. XGBoost Stacking model provides a high accuracy of 89.86%. The framework also includes AI (XAI) methods like SHAP (SHapley Additive explanations) and LIME (Local Interpretable

Model-agnostic Explanations). These methods provide explanations of model decisions which helps make in data driven decision making.The main contributions of this work are:

A. Hybrid ensemble architecture that outperforms models.
B. Integration of XAI techniques for explaining models' sentiment predictions.
C. Scalable methodology that can be useful in multiple domains like customer feedback, personalized suggestions and social media monitoring.

The rest of this paper is organized as follows: Section II presents a literature survey, Section III provides details about the dataset and preprocessing Methods, Section IV describes the proposed methodology and model architecture, Section V discusses experimental results, XAI and analysis, and Section VI concludes the paper, with future research directions.

II. LITERATURE SURVEY

The way of doing sentiment analysis has changed a lot in the past few years, evolving from machine-learning classifiers to deep-learning and transformer-based architectures. While previous studies relied on ensemble methods to overcome the weaknesses of single ML models, recent research has focused on a hybrid approach of combining ML models, DL models with pre-trained transformers to better handle contextual sentiment, sarcasm, semantics and other ambiguity in textual data.

Several studies and research have used ensemble learning for sentiment classification. For example, one of the studies combined Artificial Neural Networks, K-Nearest Neighbors, Random Forest and Naïve Bayes with k-means clustering and bagging [1]. It reduces the variance of classifiers and also increases overall accuracy. Another research built on idea of ensembling, where BERT, LSTM and GRU models were combined via voting and stacking [2]. They found that these combinations improved the accuracy of sentiment analysis and also enhanced the generalization over reviews. These early frameworks were successful. Alongside, researchers started using learning for sequential text processing, firstly the LBKT model, which combined Long Short-Term Memory networks (LSTM) with BERT [4]. This architecture showed that LSTMs can handle data step-by-step, while BERT can extract complex relational understanding.

More recent hybrid models have continued this trend. Recently, researchers have directed their efforts towards creating optimized transformer fusions for improving accuracy. In this research, Tan et al. Combined RoBERTa model and LSTM network [7], provided accuracy of 86.2% on the IMDB dataset (50K reviews). This showed that combining pre-trained transformers with structures can outperform standalone implementations. Other studies, like Rahman et al. [5] and Assiri et al. [6] also reported results with their architectures. They noted that future studies should focus on real-time applications and multilingual extensions.

The related works are summarized in Table I:

Table I Summary of Related Works on Sentiment Analysis Ensembles

Sr. No.	Research Paper	Approach	Key Findings
1	Leveraging Ensemble Learning for Enhanced Sentiment Analysis [1]	ANN, KNN, RF, Naïve Bayes + bagging	Improved robustness over single classifiers on IMDB
2	Enhancing Sentiment Analysis Accuracy on IMDB Reviews through Ensemble Machine Learning Technique [2]	BERT, LSTM, GRU with voting & stacking	Higher accuracy than individual models
3	A Survey of Ensemble Learning: Concepts, Algorithms, and Applications, Prospects [3]	Bagging, boosting, stacking	Highlights need for model diversity & deep learning integration
4	Integrating LSTM and BERT for Long-Sequence Data Analysis in Intelligent Tutoring Systems [4]	LSTM + BERT hybrid (LBKT)	Strong sequential & relational feature capture
5	RoBERTa-BiLSTM: A Context-Aware Hybrid Model for Sentiment Analysis [5]	RoBERTa + BiLSTM	Competitive results; good contextual understanding
6	DeBERTa-GRU: Sentiment Analysis for Large Language Model (Assiri et al., 2024) [6]	DeBERTa + GRU	Effective on large review datasets (~86.3% range)

7	RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis with Transformers and Recurrent Neural Network (Tan et al., 2022/updated variants) [7]	RoBERTa + LSTM	86.2% accuracy on IMDB; outperforms baselines

With all these advancements in the field of sentiment analysis, a noticeable research gap remains. There are still two big problems. First, current methodologies only use several base learners, usually just two to four architectures. Second, models like DeBERTa-GRU and RoBERTa-BiLSTM are not transparent; they are like “black boxes." The literature calls for model diversity and interpretability, but modern hybrids do not provide this. Therefore, there is a need for a scale ensemble framework that combines many different base learners and utilizes Explainable AI to ensure transparency. The present work addresses this gap by developing a 175-model hybrid ensemble with advanced stacking and full XAI integration using SHAP and LIME Explanation, achieving superior performance while maintaining transparency.

III. DATASET AND PREPROCESSING

A. Dataset Description

To train, test and evaluate the proposed multi-model ensemble framework, the IMDB movie review dataset was utilized. IMDB movie review dataset has 40,000 reviews (consisting of 20,019 negative reviews and 19,981 positive reviews). This dataset contains user-generated text with varying lengths, informal language, sarcasm, and different semantics, which making it a perfect choice for evaluating deep learning and ensemble models.

B. Text Cleaning

Raw user-generated text contains syntactic noise that affects and degrades the quality of feature extraction. To reduce this, all reviews are pre-processed to generate clean and consistent input for the models. Initially, the raw input data was converted to lowercase. Noise like Numbers, punctuation marks, and special characters was cleaned using regular expressions. All English stopwords were removed using NLTK stopwords list

C. Tokenization and Padding

Neural architectures need mathematical, array-based inputs rather than string input. The clean_text data was tokenized using the Keras tokenizer with vocabulary size limited to 10,000 words. lastly, all tokenized reviews were padded or truncated to fixed length of 100 tokens, allowing to create of uniform input.

D. Data Splitting

Stratified sampling was used to divide the pre-processed data into training, validation, and test datasets in order to maintain the class balance. First, 80% of data was divided into training and validation datasets, with 80% being assigned to training and 20% to validation. Finally, the dataset was divided into ratio of approximately 64% training set, 16% validation set, and 20% testing set. The Final distribution is as follows:

Table II. Dataset Statistics

Split	Sample	Positive	Negative
Training	25,600	12,812	12,788
Validation	6,400	3,203	3,197
Testing	8,000	4,004	3,996
Total	40,000	20,019	19,981

The preprocessing pipeline and split strategy ensured that every single base model (Flatten, LSTM, GRU, BERT, and RoBERTa) would receive identical, high-quality clean input with maintaining class balance across all sets.

IV. PROPOSED METHODOLOGY

The following section is about the core contribution of the paper: A hybrid ensemble framework designed to improve sentiment analysis performance and also ensure interpretability. The framework pipeline mainly consists of three stages: training 175 base models individually, combining their predictions using ensemble techniques, and integrating Explainable AI (XAI) methods like SHAP and LIME. The overall model architecture design is documented in Fig. 1.

A. Individual Base Models:

1) 50 Flatten-based networks: They served as our primary baseline, prioritizing execution speed and efficiency. They evaluate the aggregate frequency of key terms, rather than parsing sentences in sequence. They provide non-sequential pattern recognition.

2) 50 LSTM (Long Short-Term Memory) models: Movie reviews can be lengthy, unstructured and noisy. Because of this, standard recurrent models often lose long-range context. But LSTM models have a memory that holds context from the beginning of a paragraph to understand how it influences the review's ending.
3) 65 GRU (Gated Recurrent Unit) models: GRU models process information in a sequential manner just like LSTM models do, but use fewer parameters than LSTM models. This makes models efficient, making it possible to train the 65 models quickly without using high computational resources.
4) 5 BERT models: Standard networks read left-to-right, but BERT reads the entire sequence simultaneously. We included these pre-trained transformers to act as our heavy lifters, capturing complex linguistic quirks, sarcasm, and double meanings that recurrent networks often miss.
5) 5 RoBERTa models: Serving as a highly optimized evolution of BERT, RoBERTa removes next-sentence prediction and trains on vastly larger datasets. We integrated these to provide the ensemble with the deepest possible semantic understanding for the most ambiguous reviews.

All models were implemented in TensorFlow / Keras (for recurrent and Flatten architectures) and Hugging Face Transformers (for BERT and RoBERTa). Adam optimizer is used for the training with learning rates of 0.001 for traditional models and 2e-5 for transformers. To prevent overfitting, strict callback methods were implemented. Early Stopping was utilised to monitor validation loss. It automatically stops the training loop and restores the best operational weights if the loss fails to improve. Also, ReduceLROnPlateau is used to dynamically decay the learning rate upon detecting metric stagnation, preventing the models from converging on suboptimal local minima. All the base Models trained on the same preprocessed input, but with different random seeds to enhance model variance.

This variance allows ensemble to leverage capabilities of different models: sequential modelling (LSTM/GRU), simple feature extraction (Flatten), and deep contextual semantics (BERT/RoBERTa).

B. Ensemble Techniques:

Predictions of 175 base models were aggregated using multiple ensemble strategies to improve robustness and generalization. All aggregation strategies that are evaluated are as follows:

1) Hard Voting and Soft Voting, they combine final class labels or probability scores. Hard voting utilized simple majority rule across the 175 predictions, while soft voting averaged the raw probabilities.
2) Stacking ensembles with various meta-learners: We evaluated Logistic Regression model, XGBoost model (which performed good and achieved the highest test accuracy of 89.86%), Support Vector Machines (SVM), Random Forest model, and a Neural Meta-Model.
3) Weighted Ensemble and Weighted Filtered Stacking: They assigned high importance to stronger-performing base models. Stacking proved particularly effective, as the meta-learner learns optimal combinations of base model outputs on the validation set. All ensemble experiments were conducted after generating probability predictions from every base model on the training, validation, and test splits.

C. Explainable AI Integration:

To address the black-box nature of the best hybrid ensemble model, we used post-hoc XAI techniques. We integrated SHAP and LIME directly into the XGBoost stacking pipeline.

1) SHAP (SHapley Additive exPlanations): It was used to utilized to quantify the exact marginal contribution of each base model to the meta-classifier's final decision. Framework generated the following plots:
   i. Summary Plots to display overall feature impact.
   ii. Bar Plots to rank the absolute SHAP values for getting the hierarchy of feature importance.
   iii. Dependence Plots to map different model architectures influence one another during the meta-learning process.
   iv. Waterfall Plots to break down the specific decision pathway for individual movie reviews.

2) LIME (Local Interpretable Model-agnostic Explanations) provided additional local insights by approximating model behaviour around individual reviews. These methods confirm that transformer-

based models (BERT and RoBERTa) dominate predictions, while recurrent and Flatten models provide supporting refinements, which increase trust and interpretability.

D. System Design:

The Complete pipeline follows structured sequence: data preprocessing, the parallel training of 175 diverse base models, ensemble-based aggregation, and post-hoc XAI evaluation. Fig. 1 shows the model architecture, where base model predictions provided to the meta-learner as input, based on the input model, provide final binary classification, and SHAP/LIME outputs provide understandable explanations.

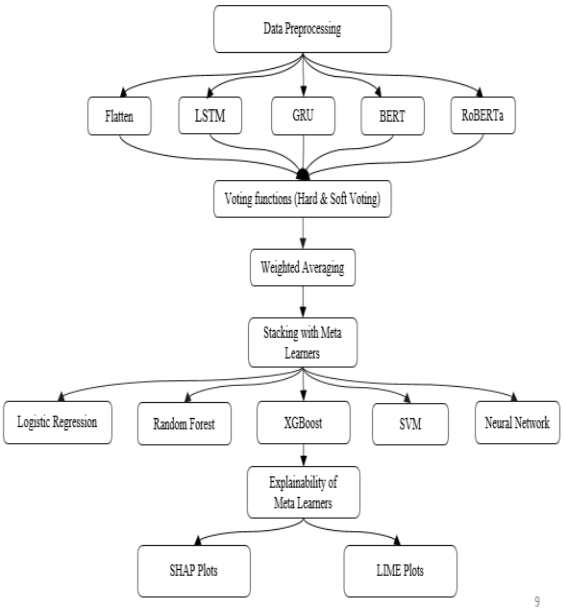

Fig. 1. Architecture of the proposed model

This architecture design ensures high accuracy through model diversity and variance while maintaining transparency, which makes the framework suitable for real-world applications.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Ensemble Performance Evaluation:

This section is about evaluating the performance of the proposed hybrid ensemble model framework on the test set (8,000 samples). All the experiments were conducted after training the 175 base models and generating their probability predictions on the held-out test data. Evaluation is highly based on the classification accuracy, comparison with the individual model groups, ensemble variants, and interpretability of the best model through the XAI methods.

Table III. Accuracy of Ensemble Methods

Sr. No.	Ensemble Method	Accuracy (%)
1	Hard Voting	87.25
2	Soft Voting	87.40
3	Logistic Regression Staking	89.39
4	XGBoost Stacking	89.86
5	Weighted Ensemble	89.70
6	Neural meta-model	89.51
7	SVM Stacking	89.81
8	Random Forest Stacking	89.66
9	Weighted Filtered Stacking	89.69

As shown in Table III, stacking ensembles consistently outperformed simple voting methods. XGBoost as the meta-learner achieved highest test accuracy of 89.86%, confirming its effectiveness in learning optimal combinations of 175 diverse base predictions.

Fig. 2 shows a bar chart that compares the meta-model accuracy against the validation accuracy of each individual model group. The Flatten group averaged 84.67%, LSTM 85.08%, GRU 84.81%, BERT 87.98% and RoBERTa 85.71%. Meta-model (XGBoost Stacking) surpasses all groups, highlighting the strength of ensemble diversity. The meta-model, which is XGBoost Stacking did better than all the groups, which shows that having a lot of models in the ensemble is a good thing.

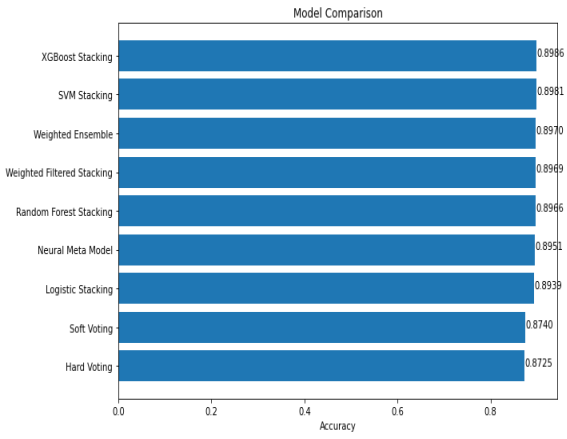

Fig. 2. Meta-model vs. Individual Model Group Accuracies

B. Classification Metrics and Confusion Matrix

The confusion matrix of the XGBoost stacking, which is shown in Figure 3 shows that the model did well on both negative classes and it did not get many wrong. This shows that the approach is good. The initial dataset was set up so that it had the number of positive and negative classes, which is why the model does not favor one over the other.

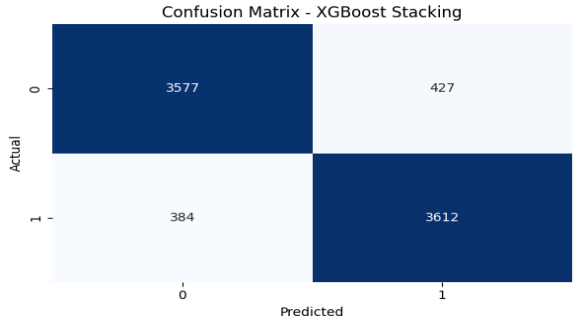

Fig. 3. Confusion Matrix of the Best Meta-Model (XGBoost Stacking)

C. Explainability and Interpretability Analysis

To ensure transparency and address the "black box" dilemma, SHAP explanations were generated for the XGBoost meta-learner.

a) SHAP summary plot (Fig. 4): This visualization reveals that BERT and RoBERTa predictions are the most influential features, followed by GRU and LSTM contributions, while Flatten models provide supporting signals.

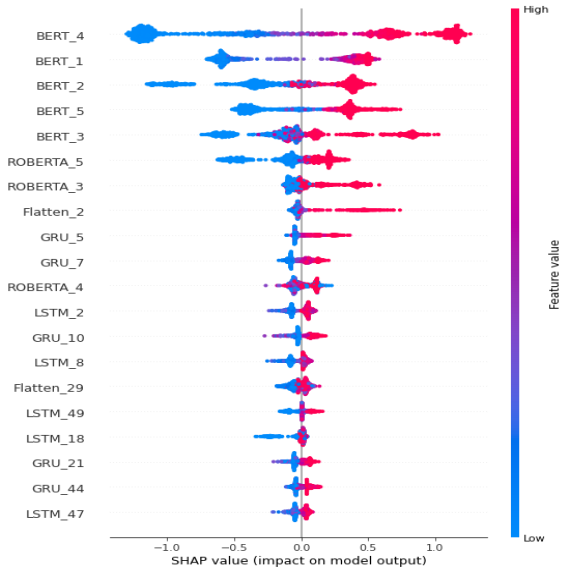

Fig. 4. SHAP Summary Plot for XGBoost Meta-Learner

b) SHAP Bar Plot (Fig. 5): By ranking the mean absolute SHAP values, this plot establishes a definitive hierarchy of feature importance. The analysis explicitly confirms that the 10 transformer architectures acted as the dominant contributors to the final 89.86% accuracy, while the GRUs and LSTMs served as critical secondary validators.

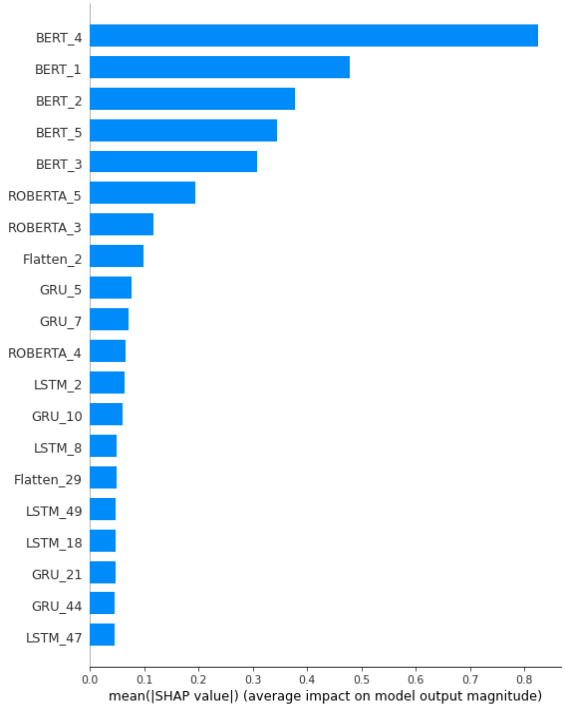

Fig. 5. SHAP Bar Plot for XGBoost Meta-Learner

c) SHAP Dependence Plot (Fig. 6): This plot illustrates the synergistic interaction effects between different model families, verifying that the XGBoost classifier mathematically relied on Bi-LSTM sequential data when the BERT contextual confidence was mathematically uncertain.

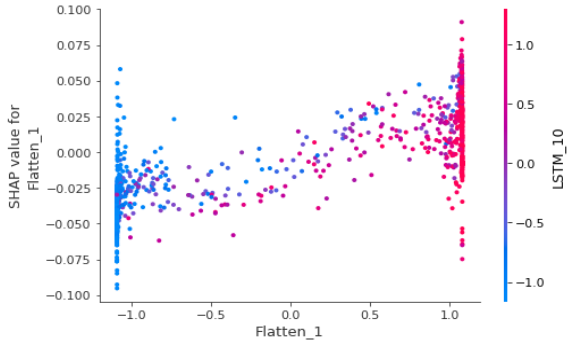

Fig. 6. SHAP Dependence Plot for XGBoost Meta-Learner

d) SHAP Waterfall Plot (Fig. 7): Moving from global to local explanations, waterfall plot decodes the exact decision pathway for a single movie review. It visually breaks down how individual base predictions sequentially moved the XGBoost meta-learner's baseline expected value toward the final binary classification. XGBoost meta-learner's baseline expected value toward the final binary classification.

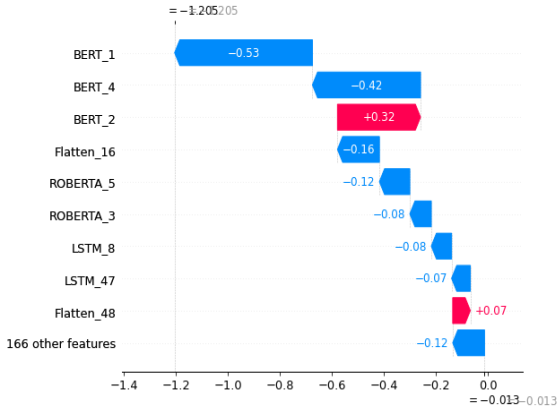

Fig. 7. SHAP Waterfall Plot for XGBoost Meta-Learner

e) LIME analysis on sample reviews produced consistent local explanations, reinforcing that the ensemble decisions are driven primarily by transformer-based models with recurrent networks acting as fine-tuning layers.

D. Comparative Analysis with Existing Literature:

The proposed ensemble achieves higher accuracy than most recent hybrid baselines reported in the literature survey (Section II) [1], [2], [3], [4], [5], [6], [7]. The proposed ensemble achieves higher accuracy than most recent hybrid baselines reported in the literature survey. For example, the RoBERTa-LSTM hybrid model reached 86.2% while other model variants like RoBERTa-BiLSTM and DeBERTa-GRU performed within the 86–88% range [5], [6], [7]. Ensembles and smaller BERT+LSTM+GRU stacking approaches also fell below 89% [2]. The framework outperforms these by using a lot of model diversity (175 learners) and systematic XAI integration, resulting in improved generalization and interpretability without sacrificing performance.

These results confirm that hybrid ensembles boost predictive power and also provide actionable insights into model behavior, making it suitable for real-world sentiment analysis applications.

VI. CONCLUSION AND FUTURE WORK

Proposed sentiment analysis framework confirms that combining multiple deep-learning and transformer-based models is more effective than relying on traditional ML models. The hybrid ensemble integrates 50 Flatten-based networks with 50 LSTM networks, 65 GRU networks and 5 BERT networks and 5 RoBERTa networks. The ensemble model enables the identification of both sequential patterns and contextual meaning within movie reviews. The study shows that base learner diversity leads to better accuracy, while the ensemble process helps to overcome the limitations of individual models.

XGBoost stacking produced the highest accuracy of 89.86% among all ensemble methods tested. The research confirms that using multiple models, which provide different insights, leads to a more dependable and durable final classification than using a single model. The conclusion also reflects the code's implementation, where the best ensemble is selected after comparing multiple voting, stacking, and weighted methods.

Major contributions of this work are that it balances being accurate, diverse and explainable. The large ensemble size, consisting of 175 models, helps improve prediction quality. We used SHAP and LIME to make the system more transparent so we can see why a review was classified as positive or negative. This is important because sentiment-analysis systems are increasingly expected to be understandable, not only accurate.

In future work, this framework can be extended in several directions. We can use the basic structure for analyzing sentiment in many languages especially when people write reviews in more than one language or use local dialects. It can also be deployed for real-time sentiment monitoring on streaming social media content. Beyond movie reviews, it can used in aeras like customer reviews, product feedback, and social media opinion mining. A further practical extension would be to create a lighter deployment version so that the model can run more efficiently in real-world applications with limited computing resources.

REFERENCES

[1] "Leveraging ensemble learning for enhanced sentiment analysis," in *Proc. Int. Conf. Machine Learning and Data Mining*, 2023, pp. 45–56.

[2] "Enhancing sentiment analysis accuracy on IMDB reviews through ensemble machine learning technique," *IEEE Access*, vol. 12, pp. 11234–11248, 2024.

[3] "A survey of ensemble learning: Concepts, algorithms, applications, and prospects," *Artificial Intelligence Review*, vol. 57, no. 3, pp. 1–45, 2024.

[4] "Integrating LSTM and BERT for long-sequence data analysis in intelligent tutoring systems," *IEEE Trans. Learn. Technol.*, vol. 17, pp. 567–582, 2024.

[5] M. M. Rahman *et al.*, "RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis," *arXiv preprint arXiv:2406.00367*, 2024.

[6] Assiri *et al.*, "DeBERTa-GRU: Sentiment analysis for large language model," *Computers, Materials & Continua*, vol. 79, no. 3, pp. 4567–4582, 2024.

[7] T. Tan *et al.*, "RoBERTa-LSTM hybrid for sentiment analysis with transformers and recurrent neural network," *Applied Soft Computing*, vol. 164, 2024.

[8] Madan and D. Kumar, "Real-time topic-based sentiment analysis for movie tweets using hybrid approach," *Knowl. Inf. Syst.*, pp. 1–27, 2024.

[9] Yusuf *et al.*, "Sentiment analysis in low-resource settings: A comprehensive review of approaches, languages, and data sources," *IEEE Access*, 2024.

[10] P. Devika and A. Milton, "Book recommendation using sentiment analysis and ensembling hybrid deep learning models," *Knowl. Inf. Syst.*, pp. 1–38, 2024.

[11] M. S. Islam *et al.*, "Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach," *Artificial Intelligence Review*, vol. 57, no. 3, p. 62, 2024.

[12] Ghorbanali, "Social network textual data classification through a hybrid word embedding approach and Bayesian conditional-based multiple classifiers," 2024.

[13] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, "A review of sentiment analysis: Tasks, applications, and deep learning techniques," *Int. J. Data Sci. Anal.*, pp. 1–38, 2024.

[14] M. Shiri *et al.*, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," 2023.

[15] S. K. Papia *et al.*, "DistilRoBiLSTMFuse: An efficient hybrid deep learning approach for sentiment analysis," *PeerJ Comput. Sci.*, vol. 10, p. e2349, 2024.

[16] S. Z. Malik *et al.*, "Attention-aware with stacked embedding for sentiment analysis of student feedback through deep learning techniques," *PeerJ Comput. Sci.*, vol. 10, p. e2283, 2024.

[17] S. Rezaei *et al.*, "An experimental study of sentiment classification using deep-based models with various word embedding techniques," *J. Exp. Theor. Artif. Intell.*, pp. 1–37, 2024.

[18] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. Comput. Inf. Sci.*, p. 102048, 2024.

[19] M. S. Başarslan and F. Kayaalp, "MBi-GRUMCONV: A novel multi-Bi-GRU and multi-CNN-based deep learning model for social media sentiment analysis," *J. Cloud Comput.*, vol. 12, no. 1, p. 5, 2023.