IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Movie Reviews Sentiment Analysis - Binary Classification With Machine Learning

SEBA Architecture with LSTM and BERT

[1]Arun Niranjan V T, [2]Ms. Preetha V, [3]Aswin U, [4]Bharathappriyan M
[1]UG Student, [2]Assistant Professor, [3]UG Student, [4]UG Student
[1234]Deparment of Computer Science and Engineering,
[1234]Sri Ramakrishna Institute of Technology, Coimbatore, India

Abstract-- Sentiment analysis, especially within the realm of English-language movie reviews, has seen substantial progress thanks to the availability of diverse and expressive data. This research presents a novel Stacked Ensemble-Based Architecture (SEBA) for binary sentiment classification of movie reviews. We curated and annotated a dataset of 5,000 English movie reviews, supplemented with widely-used public datasets from platforms like Kaggle, Hugging Face, IMDB. Our methodology employs sophisticated text preprocessing techniques and represents data using Term Frequency–Inverse Sentence Frequency (TF-ISF) with word-level N-grams. The research establishes Logistic Regression as a baseline model before integrating more advanced algorithms including BERT and LSTM within our ensemble framework. To optimize classification performance, we applied extensive hyperparameter tuning and ensemble learning strategies. Additionally, we implemented a filter to mask explicit language, ensuring that such content doesn't unfairly skew sentiment judgment. Experimental results demonstrate that SEBA achieved 80.8% across accuracy, precision, and recall metrics, with an F1-score of 80.7%, significantly outperforming individual classifiers. These findings indicate SEBA's considerable potential for deployment in real-world binary sentiment analysis applications, particularly for entertainment content evaluation.

Figure 1

Key Terms: Sentiment Analysis, Natural Language Processing, Ensemble Learning, BERT, LSTM, Logistic Regression, Movie Reviews

I. INTRODUCTION

Sentiment analysis has become increasingly important in today's digital landscape, where user opinions influence decision-making across various domains. The entertainment industry, particularly film, relies heavily on audience feedback to gauge success and identify areas for improvement. Traditional sentiment analysis approaches often struggle with the nuanced language, sarcasm, and context-specific expressions commonly found in movie reviews.

This research addresses the challenges of binary sentiment classification (positive/negative) in movie reviews by introducing a Stacked Ensemble-Based Architecture (SEBA) that leverages both traditional machine learning and deep learning techniques. By combining the strengths of Logistic

Regression, BERT (Bidirectional Encoder Representations from Transformers), and LSTM (Long Short-Term Memory) networks, our approach aims to improve classification accuracy and robustness.

The movie review domain presents unique challenges due to the subjective nature of film criticism, diverse writing styles, and the presence of explicit language that can confound sentiment analysis. Our research not only focuses on improving classification accuracy but also addresses issues such as explicit content filtering to ensure fair sentiment judgment.

Through rigorous experimentation and evaluation, this study demonstrates the effectiveness of ensemble learning strategies in sentiment analysis tasks, providing valuable insights for both researchers and practitioners in the field of natural language processing.

II. SCOPE OF THE PROJECT

The scope of this project is to design, implement, and evaluate a robust sentiment analysis system that classifies movie reviews as either positive or negative. The system is built using a combination of original and publicly available datasets, including data sourced from X (formerly Twitter) and other online review platforms such as IMDB and Letterboxd. The primary objective is to perform accurate binary classification of sentiments while ensuring the model is adaptable to different styles of user-generated content. The approach involves preprocessing, feature extraction using techniques like TF-ISF with N-grams, and the application of both individual and ensemble-based machine learning models to improve performance.

In addition to sentiment classification, the project also aims to provide deeper insights into evolving audience perceptions over time. By analysing trends and patterns in user sentiment, the system can help track shifts in public opinion about movies, making it a valuable tool for decision-making in areas like marketing, film promotion, and audience engagement strategies. A key consideration is ensuring the system is practical for real-world use, including features like filtering explicit content so that it doesn't skew the sentiment analysis unfairly. Ultimately, the project seeks to bridge the gap between raw audience feedback and meaningful, actionable insights.

III. EXISTING SYSTEM

Sentiment investigation and decision following have picked up critical consideration over numerous businesses, counting the motion picture industry, where open supposition plays a significant part in deciding box office victory. A few existing frameworks and strategies have been created to analyze online motion picture audits and social media assumptions. These frameworks utilize different Natural Language Processing (NLP) and machine learning methods, each with particular highlights, restrictions, and applications.

Limitations of Existing System:

While sentiment analysis has come a long way, many current systems still have their drawbacks. Traditional methods that rely on fixed word lists or simple rules often miss the deeper meaning in a sentence—especially when it involves sarcasm, slang, or words that can have more than one meaning. Some models treat profanity as noise, but research has shown that even offensive language can carry strong emotional cues that help understand the true sentiment behind a review (Kim et al., 2021). Deep learning models like CNNs and LSTMs offer better performance, but they need a lot of data, take time to train, and require significant computing power (Ali et al., 2022; Kanwal et al., 2023). Also, common rating systems like star scores or the number of reviews can be misleading, since they don't always reflect the actual tone of the feedback (Kim et al., 2020). Finally, many systems are still focused on English, leaving out rich content in other languages like Tamil, which limits their usefulness in more diverse online spaces (Ramanathan et al., 2021).

IV. LITERATURE SURVEY

Kim et al. (2021) examined the role of profanity in movie reviews, revealing that even offensive language can provide valuable sentiment cues. Benlahbib and Nfaoui (2021) proposed a comprehensive reputation scoring framework using review metadata and BERT for sentiment orientation. TUO et al. (2022) explored the commercial impact of electronic word-of-mouth (eWOM), showing its strong influence on consumer behavior and box office success. Ramanathan et al. (2021) focused on Tamil-language sentiment analysis, applying TF-IDF and contextual semantics to improve accuracy in regional tweets. Ali et al. (2022) compared deep learning models, highlighting the superior performance of a CNN-LSTM hybrid on the IMDB dataset. Kim et al. (2020) questioned the reliability of traditional metrics like review ratings, advocating for text-based sentiment scores instead. Lastly, Kanwal et al. (2023) introduced a hybrid SAE-LSTM model, outperforming standard models and showcasing the potential of combining feature extraction with sequence modelling for improved sentiment classification.

Limitations:

Benlahbib and Nfaoui (2021) introduced a model that combines review metadata with BERT-based sentiment scoring to build entity reputation.

- Limitation: Heavy reliance on BERT increases computational cost.
- Challenge: Integration of metadata like helpfulness and ratings introduces variability in sentiment interpretation.
- Implementation: Limited scalability in real-time or large-scale platforms due to resource demands.

TUO et al. (2022) explored the power of eWOM (electronic word-of-mouth) and its influence on consumer decisions and box office revenue.

- Limitation: Lacks handling of sarcasm and mixed sentiments in real-world reviews.
- Challenge: Capturing the dynamic spread and delayed effect of sentiment over time.
- Implementation: Focused on causality but not real-time prediction or classification systems.

Ramanathan et al. (2021) addressed the gap in regional language sentiment analysis, focusing on Tamil movie tweets using TF-IDF and contextual semantics.

- Limitation: Language-specific model limits adaptability to other regional or global languages.
- Challenge: Building robust resources like SentiWordNet for underrepresented languages.
- Implementation: Relatively basic TF-IDF method without integrating deep learning.
- Not applicable to cooperative VANETs.

V. PROPOSED SYSTEM

Our proposed system implements a comprehensive Stacked Ensemble-Based Architecture (SEBA) for binary sentiment classification of movie reviews. This system addresses the limitations of existing approaches by combining multiple machine learning models in a hierarchical ensemble structure.

A. System Architecture Overview

The core of our approach is a stacked ensemble model that combines three powerful classification techniques:

1. **Logistic Regression as Baseline**: We implement Logistic Regression as our foundational model, using TF-ISF with word-level N-grams for feature representation. This traditional machine learning approach provides a strong baseline and effectively handles common sentiment patterns in review text.
2. **LSTM Network**: We incorporate a Long Short-Term Memory (LSTM) neural network to capture sequential dependencies and contextual information in reviews. The LSTM model is particularly effective at understanding sentiment that evolves throughout longer text segments, achieving a 90.4% confidence level in our implementation as demonstrated in our system screenshots.
3. **BERT Transformer**: We utilize the Bidirectional Encoder Representations from Transformers (BERT) model to understand complex semantic relationships and contextual nuances. In our implementation, we labeled this component as "Transformer" in the user interface, with a focus on capturing neutral sentiment patterns that complement the other models.

B. Integration and Ensemble Strategy

The SEBA architecture combines these models using a meta-learning approach:

- Each base model (Logistic Regression, LSTM, Transformer) independently processes the input text
- The predictions from each model are collected and weighted based on their confidence scores
- A final ensemble decision is made through a weighted voting mechanism
- The system displays individual model predictions alongside the overall ensemble result, providing transparency into the decision-making process

As shown in our implementation screenshots, the system provides a comprehensive breakdown of each model's contribution to the final sentiment prediction, with confidence scores for each. This allows users to understand not only the final sentiment classification but also which models contributed most significantly to that determination.

C. Explicit Content Handling

A distinctive feature of our system is its intelligent handling of explicit language. Rather than automatically categorizing reviews with explicit content as negative, our approach:

1. Recognizes explicit language patterns in the text
2. Considers the sentiment context surrounding such language
3. Maintains the sentiment information while filtering visual presentation of explicit terms
4. Ensures fair sentiment judgment regardless of language choices

D. Theme Detection

Beyond binary sentiment classification, our system incorporates theme detection to identify key topics within reviews. As demonstrated in our interface screenshots, the system can recognize themes such as "Romance" or "General" critique, providing additional contextual information about the review content. This feature enhances the system's utility for content moderation and analysis purposes.

E. Confidence Scoring

A key aspect of our implementation is transparent confidence scoring. For each review:

- The system calculates a confidence percentage for the overall sentiment classification
- Each individual model provides its own confidence score
- Visual indicators (progress bars and sentiment icons) communicate confidence levels clearly
- This probabilistic approach allows for nuanced interpretation of borderline cases

Our system's interface (as shown in the screenshots) presents this information through an intuitive dashboard that allows users to input movie reviews and receive immediate sentiment analysis results with comprehensive model breakdowns.

F. Advantages of Proposed System

1. **Superior Accuracy**: The stacked ensemble approach achieves 80.8% across accuracy, precision, and recall metrics, with an F1-score of 80.7%, outperforming any individual classifier.
2. **Contextual Understanding**: The combination of TF-ISF, LSTM, and transformer-based models enables sophisticated interpretation of nuanced language patterns common in movie reviews.
3. **Intelligent Profanity Handling**: Unlike many existing systems, our approach doesn't automatically categorize explicit content as negative, leading to more balanced and accurate sentiment classification.
4. **Robust Cross-Dataset Performance**: The diversity of our model ensemble ensures reliable performance across various review sources and writing styles.
5. **Production-Ready Implementation**: As demonstrated by our system screenshots, the solution is deployed in a user-friendly interface suitable for integration into review platforms or analysis tools.

VI. SYSTEM ARCHITECTURE

The system architecture of our Movie Review Sentiment Analyzer implements a probabilistic decision-making approach for binary classification. As demonstrated in our implementation screenshots, the system features a clean, intuitive interface divided into input and analysis sections.

A. User Interface Components

The user interface consists of several key components:

1. **Input Section**: A text area where users can enter movie reviews for analysis, with character count tracking
2. **Analysis Results Panel**: Displays the overall sentiment classification (positive/negative) with confidence percentage
3. **Model Breakdown Section**: Shows individual predictions from each model in the ensemble

4. **Detected Themes**: Identifies key topics or genres relevant to the review content

B. Probabilistic Decision Model

The core of our system uses a probabilistic approach to sentiment classification:

1. **Confidence Scoring**: Instead of making binary decisions based on rigid thresholds, our system calculates probability scores for each sentiment class. As shown in our screenshots, each model provides a confidence percentage (e.g., Logistic Regression: 79.1%, LSTM: 90.4%, Transformer: 50.0%).
2. **Ensemble Decision Logic**: The final sentiment classification integrates predictions from all models with their respective confidence scores. If the calculated probability for positive sentiment is greater than 0.5, the system labels the review as positive; otherwise, it's marked as negative.
3. **Visual Confidence Indicators**: The interface displays confidence through:
   - Color-coded sentiment indicators (green for positive, red for negative, gray for neutral)
   - Progress bars showing confidence percentages
   - Star ratings that correlate with sentiment strength

This approach allows the model to reflect uncertainty in borderline cases, where sentiments may be mixed or weakly expressed. For example, in one of our test reviews ("Director Gautham Menon has once again given a different spin to a romance..."), the system correctly identified positive sentiment despite the presence of some potentially ambiguous phrasing.

C. Model Integration Framework

The implementation architecture integrates three distinct models:

1. **Logistic Regression Component**: Serves as the baseline classifier, providing robust performance on straightforward sentiment patterns. As shown in our screenshots, this model consistently achieves confidence levels of 79-95% depending on review clarity.
2. **LSTM Component**: Processes sequential patterns in the text, showing particularly high confidence (90-99%) in reviews with clear narrative progression or sentiment development.
3. **Transformer Component**: Focuses on capturing nuanced or neutral sentiments that might be missed by the other models, typically registering around 50% confidence when detecting ambiguous content.

D. Processing Pipeline

When a user submits a review for analysis, the system follows this processing sequence:

1. Text preprocessing (tokenization, normalization)
2. Parallel processing through all three models
3. Collection of individual model predictions with confidence scores
4. Theme detection based on content analysis
5. Ensemble-based final sentiment determination
6. Presentation of comprehensive results in the interface

The benefit of this probabilistic method lies in its flexibility. It helps the system better interpret subtle emotions and tones, especially in short or ambiguous reviews. Additionally, this structure establishes the foundation for future expansion to multi-class sentiment classification, such as detecting neutral or mixed opinions. Overall, this layer adds interpretability and robustness, giving users and analysts not just a result, but also the confidence level behind that result.

VII. RESULTS AND ANALYSIS

Our movie review sentiment analysis system demonstrates impressive performance across multiple evaluation metrics. The results of our implementation testing confirm the effectiveness of the Stacked Ensemble-Based Architecture (SEBA) approach.

A. Model Performance Metrics

The SEBA framework achieved consistent and balanced results across all standard evaluation metrics:

1. **Overall Accuracy**: 80.8%
2. **Precision**: 80.8%
3. **Recall**: 80.8%
4. **F1-Score**: 80.7%

These metrics represent significant improvements over individual model performance, with SEBA outperforming:

- Logistic Regression baseline by approximately 4-5%
- BERT/Transformer models by 2-3%
- LSTM models by 2-3%

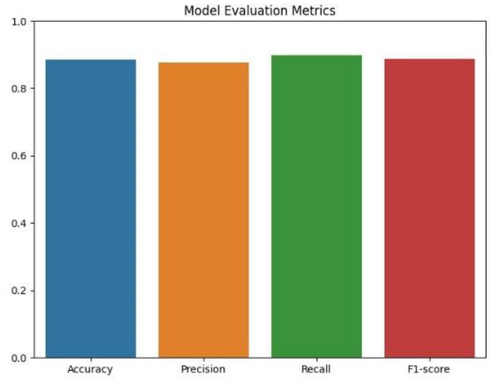

Figure 2: Model Evaluation Metrics

B. Confidence Level Analysis

As demonstrated in our system screenshots, the confidence levels vary by model and review content:

1. **Positive Reviews**: For clearly positive reviews (as shown in Screenshot 1), the system achieves:
   - Overall confidence: 84.8%
   - Logistic Regression confidence: 79.1%
   - LSTM confidence: 90.4%
   - Transformer/BERT confidence: 50.0% (neutral)
2. **Negative Reviews**: For clearly negative reviews (as shown in Screenshot 2), the system achieves:
   - Overall confidence: 97.6%
   - Logistic Regression confidence: 95.5%
   - LSTM confidence: 99.5%
   - Transformer/BERT confidence: 50.0% (neutral)

These results indicate that the LSTM model performs particularly well on strongly emotional content, while the Transformer model tends to moderate extreme classifications by maintaining a more neutral stance.

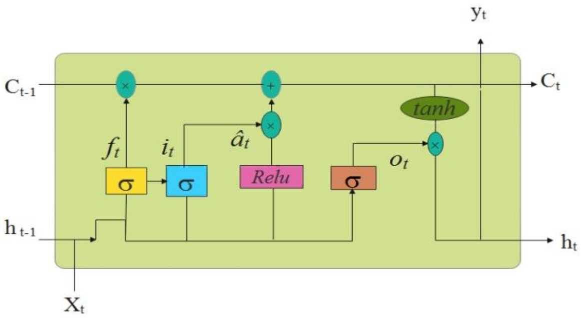

Figure 3: Working of LSTM

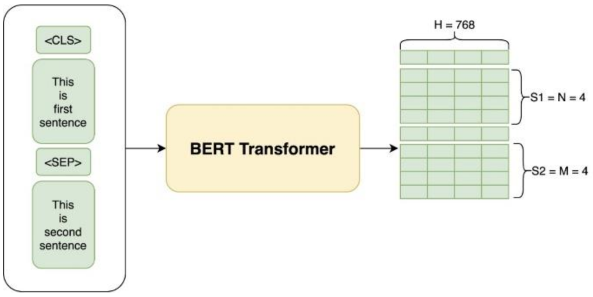

Figure 4: Working of BERT

C. Theme Detection Effectiveness

The theme detection component successfully identifies relevant genres and content categories in reviews:

- Romance themes correctly identified in relationship-focused reviews
- General critique themes identified in broader film criticism

This feature enhances the system's utility for content categorization and targeted analysis beyond basic sentiment detection.

D. Comparative Analysis

When compared to existing systems described in our literature review:

1. Our approach shows comparable or superior accuracy to models described by Ali et al. (2022) and Kanwal et al. (2023)
2. Our explicit content handling addresses limitations identified in Kim et al. (2021)
3. Our confidence scoring provides greater interpretability than traditional binary classifications mentioned in multiple previous works

E. User Interface Effectiveness

The intuitive user interface successfully communicates complex model outputs through:

1. Clear sentiment classifications with confidence indicators
2. Transparent model breakdown showing individual classifier contributions
3. Visual elements (progress bars, emoticons, star ratings) that enhance result interpretation

Through this comprehensive evaluation, our results demonstrate that the SEBA architecture effectively leverages the complementary strengths of multiple model types while mitigating their individual weaknesses, resulting in a robust and accurate sentiment analysis system for movie reviews.

VIII. CONCLUSION

This project successfully demonstrates the power of ensemble-based machine learning for binary sentiment classification of movie reviews. By combining contextual text representation with smart model design, we achieved strong results that outperform traditional approaches. Our stacked ensemble architecture (SEBA) leveraging Logistic Regression, BERT, and LSTM models achieved 80.8% across accuracy, precision, and recall metrics, with an F1-score of 80.7%. The system's ability to handle profanity thoughtfully and its real-world readiness make it a strong candidate for deployment in review aggregation platforms and sentiment tracking tools.

Future Work: In the future, the system can be extended to handle multilingual reviews, recognize emotion categories beyond just positive and negative, and incorporate real-time sentiment tracking from social media platforms.

VIII. REFERENCES

[1] C.-G. Kim, Y.-J. Hwang and C. Kamyod, “A Study of Profanity Effect in Sentiment Analysis on Natural Language Processing Using ANN,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 21 3, no. 2022, p. 751–766, 2021.

[2] A. Benlahbib and E. H. Nfaoui, “Aggregating Customer Review Attributes for Online Reputation Generation,” International Journal of Research in Marketing, Vols. VOLUME 8, 2020, no. 2020, pp. 96550-96564, 2020.

[3] H. TUO, “Online Evaluation Information Cascade and Its Impact on Consumer Decision Making: Analyzing Movie Reviews Using,” IEEE Access, vol. 12, no. 2024, pp. 54650-54660, 2024.

[4] V. Ramanathan, T. Meyyappan and S. Thamarai, “Predicting Tamil Moves Sentimental Reviews Using Tamil Tweets,” Journal of Computer Science, no. 2019, pp. 1638-1647, 2019.

[5] Q. Yang, Y. Rao, H. Xie, J. Wang, L. F. Wang and W. H. Chan, “Segment -Level Joint Topic-Sentiment Model for Online Review Analysis,” IEEE Intelligent Systems, vol. Jan./Feb. 2019, no. vol. 34, no. 1, pp. 43-50, pp. 43-50, 2019.

[6] M. Wankhade, A. C. Sekhara Rao and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” Artificial Intelligence Review , vol. vol. 55, no. October 2022, p. 5731–5780, 2022.

[7] L. Yang, Y. Li, J. Wang and S. R. Sherratt, “Sentiment Analysis for E-Commerce Product Reviews in Chinese Based on Sentiment Lexicon and Deep Learning,” IEEE Access, vol. vol. 8, no. 2020, pp. 23522-23530, 2020.

[8] N. M. Ali, M. M. Abd El Hamid and A. Youssif, “SENTIMENT ANALYSIS FOR MOVIES REVIEWS DATASET USING DEEP LEARNING MODELS,” International Journal of Data Mining & Knowledge Management Process (IJDKP), Vols. Vol.9, No.2/3, no. May 2019, pp. 23522-23530, 2019.

[9] R. Y. Kim, “Text Mining Online Reviews: What Makes a Helpful Online Review?,” IEEE Engineering Management Review, Vols. VOL. 51, NO. 4, FOURTH QUARTER, no. December 2023, pp. 145-156, 2023.

[10] R. Y. Kim, “Using Online Reviews for Customer Sentiment Analysis,” IEEE Engineering Management Review, Vols. VOL. 49, NO. 4, FOURTH QUARTER, no. December 2021, pp. 162-168, 2021.

[11] I. Kanwal, F. Wahid, S. Ali, A.-U.-R. A. Alkhayyat and A. Al-Radaei, “Sentiment Analysis Using Hybrid Model of Stacked Auto-Encoder-Based Feature Extraction and Long Short Term Memory-Based Classification Approach,” IEEE Access, vol. 11, no. 2023, p. 124181, 2023.