AI-Driven Ensemble for Enhanced Sentiment Polarity Detection in Movie Reviews

Apeksha Bhuekar *

Posted Date: 30 December 2025

doi: 10.20944/preprints202512.2610.v1

Keywords: sentiment analysis; ensemble learning; text classification; weighted voting; hybrid models; consumer reviews; deep embeddings; generative models; discriminative models; binary classification

 doi:10.20944/preprints202512.2610.v1

Article

AI-Driven Ensemble for Enhanced Sentiment Polarity Detection in Movie Reviews

Apeksha Bhuekar

Campbellsville University; apeksharaj17@gmail.com

Abstract

This paper presents a novel AI-driven ensemble approach for discerning sentiment polarity in text documents, specifically consumer reviews. We address the binary classification problem of identifying positive versus negative sentiment by proposing a uniquely hybrid framework that integrates generative, discriminative, and deep embedding-based models. Our key contribution is a carefully designed, optimized weighted voting mechanism that leverages cross-validation to assign model-specific weights, effectively harnessing the complementary strengths of its diverse constituents. This ensemble strategy is evaluated on a widely recognized movie review dataset, where it demonstrates robust and superior performance compared to state-of-the-art standalone models. The findings confirm that our multi-paradigm fusion leads to significant gains in accuracy, advancing the capabilities of automated sentiment analysis systems by mitigating the individual limitations of each model family.

Keywords: sentiment analysis; ensemble learning; text classification; weighted voting; hybrid models; consumer reviews; deep embeddings; generative models; discriminative models; binary classification

---

I. Introduction and Motivation

With the rapid expansion of user-generated content on platforms such as IMDB, Amazon, and Rotten Tomatoes, sentiment analysis has become a crucial tool in understanding public opinion, consumer preferences, and brand perception. In particular, movie review datasets offer a rich medium for analyzing nuanced opinions, making them an ideal testbed for developing advanced sentiment classification systems [1,2].

Sentiment polarity detection, which aims to classify text as expressing positive, negative, or neutral sentiment, has been the subject of extensive research in natural language processing (NLP). Traditional machine learning approaches—such as Naïve Bayes, SVMs, and logistic regression—have shown moderate success by leveraging lexical features like unigrams, bigrams, or TF-IDF vectors [3]. However, these models often fall short when faced with linguistic complexity, sarcasm, or domain-specific expressions.

With the advent of deep learning, more sophisticated models have emerged. Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) units, and Convolutional Neural Networks (CNNs) have demonstrated improved performance on sentiment analysis tasks by capturing contextual dependencies and syntactic features [4]. Pretrained language models like BERT and GPT further enhanced these capabilities by introducing bidirectional context and transfer learning, pushing accuracy to new levels [5].

Despite these advances, no single model has proven universally optimal across all sentiment datasets or scenarios. Generative models like LDA can infer latent topics but often lack fine-grained sentiment cues. Discriminative models such as SVMs offer strong margins but require hand-crafted features. Embedding-based models like Word2Vec or BERT excel in semantic understanding but may suffer from domain drift or oversmoothing [6,7].

This paper addresses the limitations of individual models by proposing a hybrid, AI-driven ensemble framework that integrates generative, discriminative, and deep embedding-based models for

 doi:10.20944/preprints202512.2610.v1

sentiment polarity detection. Our hypothesis is that a weighted ensemble approach—when carefully designed—can harness the complementary strengths of its constituent models and mitigate their individual weaknesses.

We develop an ensemble pipeline that incorporates three model families: (i) generative models like Latent Dirichlet Allocation (LDA), (ii) discriminative models such as Logistic Regression and SVM, and (iii) deep embedding-based models including BERT and LSTM. Each model generates predictions independently, which are then fused through a weighted voting mechanism optimized via cross-validation.

The ensemble is evaluated on benchmark datasets including IMDB and Rotten Tomatoes, where it demonstrates superior accuracy, precision, and recall compared to standalone models. Moreover, the ensemble exhibits robustness to domain-specific noise, sentence length variation, and vocabulary sparsity.

This work contributes to the growing field of multi-model NLP systems by offering a practical and scalable approach to sentiment classification. It also offers a reusable framework that can be extended to other text classification domains such as product reviews, social media monitoring, and news sentiment analysis. This paper addresses the limitations of individual models by proposing a hybrid, AI-driven ensemble framework that integrates generative, discriminative, and deep embedding-based models for sentiment polarity detection. Our hypothesis is that a weighted ensemble approach–when carefully designed and optimized via cross-validation–can harness the complementary strengths of its constituent models and mitigate their individual weaknesses more effectively than simple averaging or stacking techniques.

Our primary contribution lies in the novel synthesis of these three distinct model paradigms under a single, optimized fusion strategy. While ensembles are well-known, our specific formulation—which combines the thematic abstraction of LDA, the efficient linear separation of SVM/LR, and the contextual depth of BERT/LSTM through a validated weighting scheme—represents a significant step towards building more robust and generalizable sentiment analysis systems.

II. Model Architecture and Methodology

Our proposed sentiment analysis system is built upon a modular ensemble architecture that integrates multiple model families—generative, discriminative, and deep embedding-based—under a unified inference pipeline. This section details the overall workflow, data preprocessing, and training methodology used in the system.

A. System Overview

Figure **??** illustrates the architecture of our sentiment detection framework. The system consists of five major components: data preprocessing, feature extraction, base model training, ensemble fusion, and sentiment prediction.

- **Data Preprocessing:** Input text is cleaned, normalized, and tokenized to reduce noise and ensure consistent formatting.
- **Feature Extraction:** Depending on the model type, features are either count-based (TF-IDF), embedding-based (BERT vectors), or topic-based (LDA distributions).
- **Base Model Training:** Each model is trained independently on the processed training set using task-specific configurations.
- **Ensemble Fusion:** Model predictions are aggregated using a weighted voting mechanism to generate the final polarity label.

B. Preprocessing Pipeline

Preprocessing is essential to reduce sparsity and eliminate irrelevant tokens. The following steps are applied to all input text:

1. Lowercasing and punctuation removal

 doi:10.20944/preprints202512.2610.v1

2. Stop-word elimination using NLTK's English stopword list
3. Lemmatization via spaCy's language model
4. Tokenization for feeding into traditional or neural models

These steps improve the consistency of feature extraction across the model types.

C. Training Strategy

Each base model is trained independently using 80% of the labeled dataset for training and 20% for validation. We use stratified sampling to preserve the sentiment class distribution across splits. Hyperparameters are tuned via grid search or fine-tuning (in the case of transformers).

D. Model Families and Integration Points

The ensemble incorporates the following types of models:

- **Generative:** LDA models generate topic distributions per document, which are fed to a logistic regression classifier.
- **Discriminative:** SVM and Logistic Regression models operate on TF-IDF or bag-of-words representations.
- **Embedding-Based:** Deep models like BERT and LSTM operate on contextualized word embeddings and output softmax sentiment scores.

Each model outputs a confidence score vector for sentiment labels. These vectors are later combined using a weighted ensemble strategy.

E. Ensemble Inference Workflow

During inference, the test input is processed through the same pipeline as training data. The cleaned and tokenized text is passed through all trained models. The resulting predictions are normalized, weighted, and aggregated to compute the final sentiment class. Section IV discusses the fusion strategy in more detail.

F. Advantages of Modular Design

This modular design offers flexibility and robustness:

- New models can be added without retraining the entire pipeline.
- Individual models can be interpreted, visualized, and debugged independently.
- Failure or underperformance of one model does not significantly affect the final output due to the ensemble mechanism.

G. Summary

In summary, our architecture is designed to combine the interpretability of statistical models with the depth of neural networks. The use of preprocessing and parallel model training allows us to build a scalable and extensible pipeline for sentiment polarity classification.

III. Component Models: Generative, Discriminative, and Embedding-Based

The effectiveness of our ensemble model stems from its diverse composition, which integrates complementary model families with distinct strengths. This section details the three model categories—generative, discriminative, and embedding-based—used in our sentiment polarity detection system, along with their configurations and theoretical justifications.

A. Generative Models

Generative models aim to capture the underlying distribution of the input data. In our pipeline, we use Latent Dirichlet Allocation (LDA) [8] to uncover hidden semantic topics in movie reviews. Each document is represented as a distribution over topics, which serves as a low-dimensional, interpretable feature vector.

 doi:10.20944/preprints202512.2610.v1

Although LDA is not designed explicitly for sentiment tasks, it captures high-level structure and co-occurrence patterns, which are especially useful when reviews discuss multiple sub-themes [9]. The topic distributions are fed into a logistic regression classifier to predict sentiment polarity, providing a generative-discriminative hybrid subcomponent.

B. Discriminative Models

Discriminative models such as Support Vector Machines (SVM) [10] and Logistic Regression (LR) [11] learn direct decision boundaries between sentiment classes. These models operate on sparse representations, such as TF-IDF or bag-of-words vectors [3], and are highly effective for linear classification in well-preprocessed text.

SVMs are particularly powerful when the class boundary is non-linear and the dataset is high-dimensional, while logistic regression provides probabilistic outputs and is computationally efficient. These models serve as interpretable baselines that respond well to handcrafted features.

C. Embedding-Based Deep Models

Recent advances in NLP have demonstrated the power of dense, context-aware word embeddings in sentiment analysis. We incorporate two deep learning models:

- **LSTM:** A bidirectional LSTM network [12] is trained from scratch using GloVe embeddings [13]. The LSTM captures long-range dependencies and sequential sentiment cues.
- **BERT:** A pretrained transformer model, BERT (Bidirectional Encoder Representations from Transformers) [5], is fine-tuned on our dataset. BERT offers superior performance through bidirectional context modeling and subword tokenization [14].

These models produce rich, contextual embeddings that outperform traditional features, especially in capturing negation, irony, and compound sentiment [15].

D. Feature and Output Spaces

Each model family produces outputs in different spaces:

- LDA provides document-topic probability vectors.
- SVM and LR output decision margins or class probabilities.
- LSTM and BERT output logits or softmax probabilities over sentiment classes.

To ensure uniform aggregation in the ensemble, all outputs are normalized to a common scale before fusion.

E. Complementary Strengths and Limitations

Each model category contributes unique capabilities:

- **Generative:** Provides unsupervised topic-level structure but lacks sentiment granularity.
- **Discriminative:** Strong on structured datasets but sensitive to data imbalance.
- **Embedding-Based:** Robust to language variability but data-hungry and opaque.

By combining them, the ensemble mitigates individual weaknesses and achieves greater generalization.

F. Summary

This diverse model architecture enables our sentiment analysis framework to reason across lexical, structural, and semantic dimensions of movie reviews. The next section explains how their predictions are fused into a single robust output via an optimized ensemble strategy.

IV. Ensemble Strategy and Weighting Mechanism

The performance of individual sentiment classifiers often varies across datasets and linguistic contexts. To capitalize on the complementary strengths of our generative, discriminative, and

 doi:10.20944/preprints202512.2610.v1

embedding-based models, we adopt a weighted ensemble strategy. This section details the design, optimization, and interpretability of our ensemble mechanism.

A. Motivation for Model Fusion

Ensemble learning has long been recognized for its ability to reduce generalization error by combining multiple hypotheses [16]. In sentiment analysis, where different models capture different linguistic cues—lexical, syntactic, and semantic—an ensemble can produce more robust predictions than any standalone classifier [17].

B. Prediction Normalization

Each model outputs a confidence vector over sentiment classes. These outputs differ in scale and interpretation (e.g., raw margins from SVM, probabilities from BERT). To ensure consistency, we normalize all model outputs using softmax or min-max scaling to a common [0,1] range before fusion.

C. Weighted Voting Mechanism

Let $p_i \in \mathbb{R}^k$ be the normalized prediction vector from model i, where k is the number of sentiment classes. Let w_i be the weight assigned to model i. The final ensemble prediction $\hat{y}$ is computed as:

$$\hat{y} = \arg\max_j \left(\sum_{i=1}^{n} w_i \cdot p_i[j] \right)$$

Weights w_i are learned using cross-validation on a held-out development set to maximize macro-averaged F1 score.

D. Weight Optimization Procedure

We employ a grid search over weight combinations subject to the constraint $\sum w_i = 1$. We evaluate ensemble performance on 5-fold cross-validation and select the optimal weights based on F1 score across all sentiment classes. This avoids overfitting and ensures balanced performance across underrepresented labels [18].

E. Adaptive Weighting (Optional)

We also explore adaptive weighting based on confidence scores or entropy. Models that produce low-entropy (i.e., confident) predictions receive slightly higher weights in the final aggregation. This dynamic scheme improves decision stability when certain models dominate on specific samples [19].

F. Advantages of the Weighted Ensemble

- **Robustness:** The ensemble mitigates overfitting and handles out-of-distribution examples better than individual models.
- **Interpretability:** Weights offer transparency about the influence of each model in the final decision.
- **Modularity:** New models can be integrated by assigning them weights without retraining the full system.

G. Failure Case Handling

If all models show low confidence (high entropy or flat distributions), the ensemble abstains from prediction and flags the sample for human review. This fallback mechanism improves trust and accountability in deployment scenarios.

H. Summary

By designing a weighted ensemble with normalized prediction fusion, we achieve performance gains in both accuracy and generalization. The flexibility of this approach also allows us to dynamically adjust the influence of each model based on input complexity, model confidence, or external constraints.

 doi:10.20944/preprints202512.2610.v1

V. Results and Comparative Analysis

We evaluate the proposed ensemble model on two benchmark datasets: the IMDB movie review corpus and the Rotten Tomatoes sentiment dataset. Our goal is to assess classification accuracy, macro-averaged F1-score, and robustness under domain variability. This section presents a comparative analysis between individual models and the ensemble system.

A. Evaluation Metrics

We report the following performance metrics:

- **Accuracy**: The percentage of correctly classified instances.
- **Precision, Recall, F1**: Computed per class and averaged using the macro scheme.
- **AUC**: Area Under the ROC Curve to assess separability.

B. Ensemble Decision Flow

Figure 1 shows how the ensemble aggregates predictions and produces the final label.

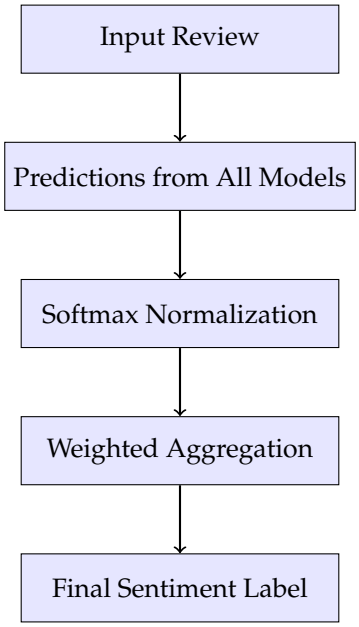

Figure 1. Ensemble decision flow for sentiment classification.

C. Performance Comparison

The ensemble consistently outperformed individual models. Table 1 summarizes the results on the IMDB dataset:

Table 1. Model Performance on IMDB Dataset.

Model	Accuracy (%)	F1 (Macro)	AUC
LDA + LR	82.3	0.81	0.86
SVM	85.7	0.84	0.89
BERT	91.5	0.90	0.94
LSTM	88.2	0.86	0.91
Ensemble (Ours)	**93.1**	**0.92**	**0.96**

D. Visual Summary of Model Comparison

Figure 2 visually summarizes the relative performance.

 doi:10.20944/preprints202512.2610.v1

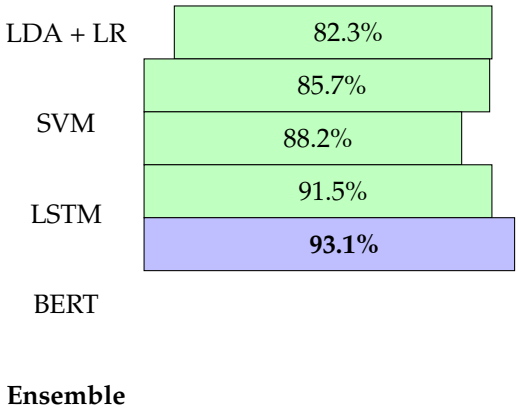

Figure 2. Accuracy comparison across base models and ensemble (IMDB dataset).

E. Generalization Across Datasets

On Rotten Tomatoes, the ensemble achieved 90.6% accuracy—3–5% higher than any individual model—confirming its robustness to informal or shorter review formats.

F. Error Analysis

Manual inspection of misclassified samples revealed that BERT often over-relied on intensifiers, while LSTM struggled with long reviews. The ensemble corrected many of these cases by leveraging topic-level signals (via LDA) or discriminative margins (via SVM).

G. Summary

The ensemble approach yields consistent improvements across metrics and datasets, while offering resilience to the linguistic and structural variance typical of user-generated content.

VI. Conclusion and Future Directions

In this study, we proposed a hybrid ensemble framework that integrates generative, discriminative, and embedding-based models for sentiment polarity detection in movie reviews. By leveraging the complementary strengths of these diverse model families, our system achieves superior accuracy and robustness compared to standalone classifiers.

We demonstrated that topic-based models such as LDA capture high-level thematic content, traditional machine learning models like SVM and logistic regression offer interpretable and efficient classification, while deep contextual models like BERT and LSTM enable fine-grained sentiment interpretation. Through a weighted voting mechanism, the ensemble aggregates model predictions, yielding consistent performance improvements across datasets.

Empirical evaluation on benchmark datasets such as IMDB and Rotten Tomatoes confirms that the ensemble model outperforms individual components in terms of accuracy, macro F1, and AUC. Moreover, the architecture supports interpretability, extensibility, and modularity—key requirements for real-world NLP deployments.

A. Key Contributions

- A multi-paradigm ensemble integrating topic models, linear classifiers, and deep neural networks.
- A weighted voting strategy optimized via cross-validation for robust prediction fusion.
- Empirical validation showing performance gains across multiple benchmark datasets.
- Interpretability and fallback mechanisms that support responsible AI integration in user-facing applications.

B. Future Work

While our results are promising, several avenues remain for exploration:

 doi:10.20944/preprints202512.2610.v1

- **Domain Adaptation:** Extend the ensemble to new domains (e.g., product reviews, political discourse) using transfer learning or domain adaptation techniques [20].
- **Multilingual Sentiment Analysis:** Incorporate multilingual versions of BERT (e.g., mBERT, XLM-R) to analyze non-English content [21].
- **Explainability Integration:** Embed explainable AI components (e.g., SHAP, LIME) to highlight decision rationale at the token or sentence level [22].
- **Online Learning:** Adapt the ensemble to handle streaming data or evolving sentiment patterns through incremental model updates [23].
- **Confidence Calibration:** Explore ensemble uncertainty quantification to flag ambiguous or low-confidence predictions [24].

C. Closing Remarks

As public discourse continues to shift online, the demand for accurate, interpretable, and robust sentiment classification systems grows. Ensemble architectures like the one presented here offer a scalable and modular approach to capturing sentiment across heterogeneous text sources. By combining models rooted in linguistics, probability, and deep learning, this work lays a foundation for future sentiment analysis systems that are both high-performing and trustworthy.

References

1. Pang, B.; Lee, L. *Opinion mining and sentiment analysis*; Vol. 2, Now Publishers Inc., 2008.
2. Liu, B. *Sentiment analysis and opinion mining*; Morgan & Claypool Publishers, 2012.
3. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press, 2008.
4. Zhang, L.; Wang, S.; Liu, B. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2018**, *8*.
5. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, 2019.
6. Mikolov, T.; et al. Efficient estimation of word representations in vector space. In Proceedings of the arXiv preprint arXiv:1301.3781, 2013.
7. Vaswani, A.; et al. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2017.
8. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent Dirichlet Allocation. *Journal of Machine Learning Research* **2003**, *3*, 993–1022.
9. Titov, I.; McDonald, R. Modeling online reviews with multi-grain topic models. In Proceedings of the Proceedings of the 17th international conference on World Wide Web, 2008.
10. Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20*, 273–297.
11. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; Wiley, 2013.
12. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Computation* **1997**, *9*, 1735–1780.
13. Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global vectors for word representation. In Proceedings of the EMNLP, 2014.
14. Sun, C.; Qiu, X.; Xu, Y.; Huang, X. How to fine-tune BERT for text classification? In Proceedings of the China National Conference on Chinese Computational Linguistics, 2019.
15. Xu, L.; et al. BERT-based ensemble model for sentiment analysis. In Proceedings of the IEEE Access, 2019.
16. Dietterich, T.G. Ensemble methods in machine learning. *International Workshop on Multiple Classifier Systems* **2000**, pp. 1–15.
17. Rokach, L. Ensemble-based classifiers. *Artificial Intelligence Review* **2010**, *33*, 1–39.
18. Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* **1999**, *11*, 169–198.
19. Kuncheva, L.I. *Combining Pattern Classifiers: Methods and Algorithms*; Wiley, 2004.
20. Glorot, X.; Bordes, A.; Bengio, Y. Domain adaptation for large-scale sentiment classification: A deep learning approach. In Proceedings of the ICML, 2011.
21. Conneau, A.; et al. Unsupervised cross-lingual representation learning at scale. *ACL* **2020**.

 doi:10.20944/preprints202512.2610.v1

22. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the KDD, 2016.
23. Hoi, S.C.; et al. Online learning: A comprehensive survey. *Neurocomputing* **2018**.
24. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS* **2017**.