LAUTECH Journal of Engineering and Technology 19 (3) 2025: 155-160
10.36108/laujet/5202.91.0341

Sentiment Analysis of Movie Reviews using Word Embeddings and Machine Learning Techniques

Ipadeola I. O., Ojo J. A., and Adebayo I. G.

Department of Electronic and Electrical Engineering, Ladoke Akintola University of Technology, Ogbomoso.

Article Info

Article history:

Received: June 12, 2025
Revised: July 30, 2025
Accepted: Aug. 4, 2025

Keywords:

Sentiment analysis,
Word embedding,
Machine learning

Corresponding Author:
igadebayo@lautech.edu.ng

ABSTRACT

In this study, sentiment analysis of movie reviews was carried out using word embeddings and machine learning techniques. Sentiment analysis, as an opinion mining technique, involves using feature extraction methods to understand the opinions and emotions expressed in text—particularly in domains such as movie reviews, where public sentiment plays a strong role in shaping consumer decisions. For sentiment analysis to be effective, text must be converted into a form that a computer can process. This involves transforming words or documents into vectors using word embedding techniques. Common techniques include Bag of Words, TF-IDF, and Word2Vec. In this study, TF-IDF and Bidirectional Encoder Representations from Transformers (BERT) were selected to compare their effectiveness in analyzing sentiment in movie reviews. The research used the IMDb dataset, which is widely recognized and commonly used in text mining tasks. Various machine learning models were applied, including Support Vector Machine (SVM), XGBoost, and Long Short-Term Memory (LSTM). Results showed that the combination of TF-IDF and SVM produced the highest accuracy, outperforming more complex models such as BERT with LSTM. The findings suggest that simpler word embedding techniques, when paired with effective classifiers, can give strong performance in sentiment analysis.

INTRODUCTION

The widespread adoption of social media and digital platforms has profoundly transformed how individuals communicate, share, and evaluate information. Platforms such as Twitter, Facebook, and Reddit have become central to public discourse, allowing users to voice opinions on topics ranging from products and services to politics and entertainment. For businesses and media industries, this user-generated content (UGC) has emerged as a crucial resource for gauging public sentiment and informing decision-making processes (Gibson *et al*., 2025; Rathor *et al*., 2024; Xu, 2024; Yang, 2024).

Sentiment analysis, a subfield of natural language processing (NLP), focuses on the computational identification and classification of opinions, emotions, and attitudes expressed in text (Zhou and Liu, 2023). It is extensively applied in customer feedback systems, product and service reviews, political sentiment tracking, and entertainment media. In the film industry, sentiment analysis of movie reviews can help stakeholders better understand audience reactions and predict commercial success (Nair *et al*., 2024; Asha *et al*., 2023; Nkhata *et al*., 2025).

A key factor influencing the accuracy of sentiment classification models is the method used for textual feature representation. Traditional approaches, such as Term Frequency–Inverse Document Frequency (TF-IDF) provide a statistical measure of word relevance across documents but do not account for semantic or contextual meaning. Despite this limitation, TF-IDF remains popular due to its simplicity and compatibility with classical machine learning algorithms like Support Vector Machines (SVM) and XGBoost (González-Carvajal and Garrido-Merchán, 2020; Kumar and Bansal, 2023;

Subramaniyaswamy *et al*., 2024). In recent years, the field of NLP has seen significant advancements with the development of contextual embeddings such as Bidirectional Encoder Representations from Transformers (BERT). These models generate dynamic word representations by considering the full sentence context and have demonstrated superior performance across a range of NLP tasks, particularly when combined with deep learning architectures like Long Short-Term Memory (LSTM) and convolutional neural networks (CNNs) (Gibson *et al*., 2025; Zhou and Liu, 2023; Nurul and Roziati, 2023; Nkhata *et al*., 2025).

Despite the impressive performance of contextual embeddings, they often require substantial computational resources and large amounts of training data. Interestingly, recent studies indicate that traditional models such as TF-IDF with SVM or XGBoost can achieve competitive results with significantly lower computational cost, especially in smaller or structured datasets (Rathor *et al*., 2024; Asha *et al*., 2023; González-Carvajal and Garrido-Merchán, 2020; Xu, 2024). Therefore, it is essential to evaluate embedding techniques not only based on their complexity or novelty but also on their practical compatibility with different classifiers and their effectiveness in specific applications. This study aims to develop a sentiment classification framework by comparing TF-IDF and BERT embeddings applied to the IMDb movie reviews dataset using SVM, LSTM, and XGBoost classifiers. The goal is to identify which model-embedding combinations yield optimal performance across multiple evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC.

MATERIALS METHODS

Dataset Description

This study utilizes the IMDb movie reviews dataset, a widely adopted benchmark for sentiment classification tasks. The dataset consists of 50,000 reviews, evenly divided into 25,000 positive and 25,000 negative samples, then split into training and test sets, comprising 70% (35,000 samples) and 30% (15,000 samples) of the data, respectively. Each review is labelled based on its sentiment polarity. The dataset is publicly available and commonly used in text mining and natural language processing research due to its balanced distribution and domain specificity. The full experimental procedure is presented in Figure 1.

Data Preprocessing

Before applying machine learning models, the textual data underwent a series of preprocessing steps to enhance model performance. The text was converted to lowercase, and punctuation, special characters, and HTML tags were removed. Stopwords were filtered out to reduce noise in the data. Tokenization was performed to split the reviews into individual words. For deep learning models such as LSTM, sequences were padded to a uniform length to ensure consistency in input dimensions.

Word Embedding Techniques

Two word embedding techniques were employed in this study: Term Frequency–Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT).

- TF-IDF is a statistical method that reflects how important a word is to a document within a collection. It was used to convert the textual data into sparse feature vectors suitable for traditional machine learning models.
- BERT, a pre-trained contextual language model, generates dense and dynamic word representations by considering the full sentence context. The Hugging Face Transformers library was used to extract sentence embeddings for classification.

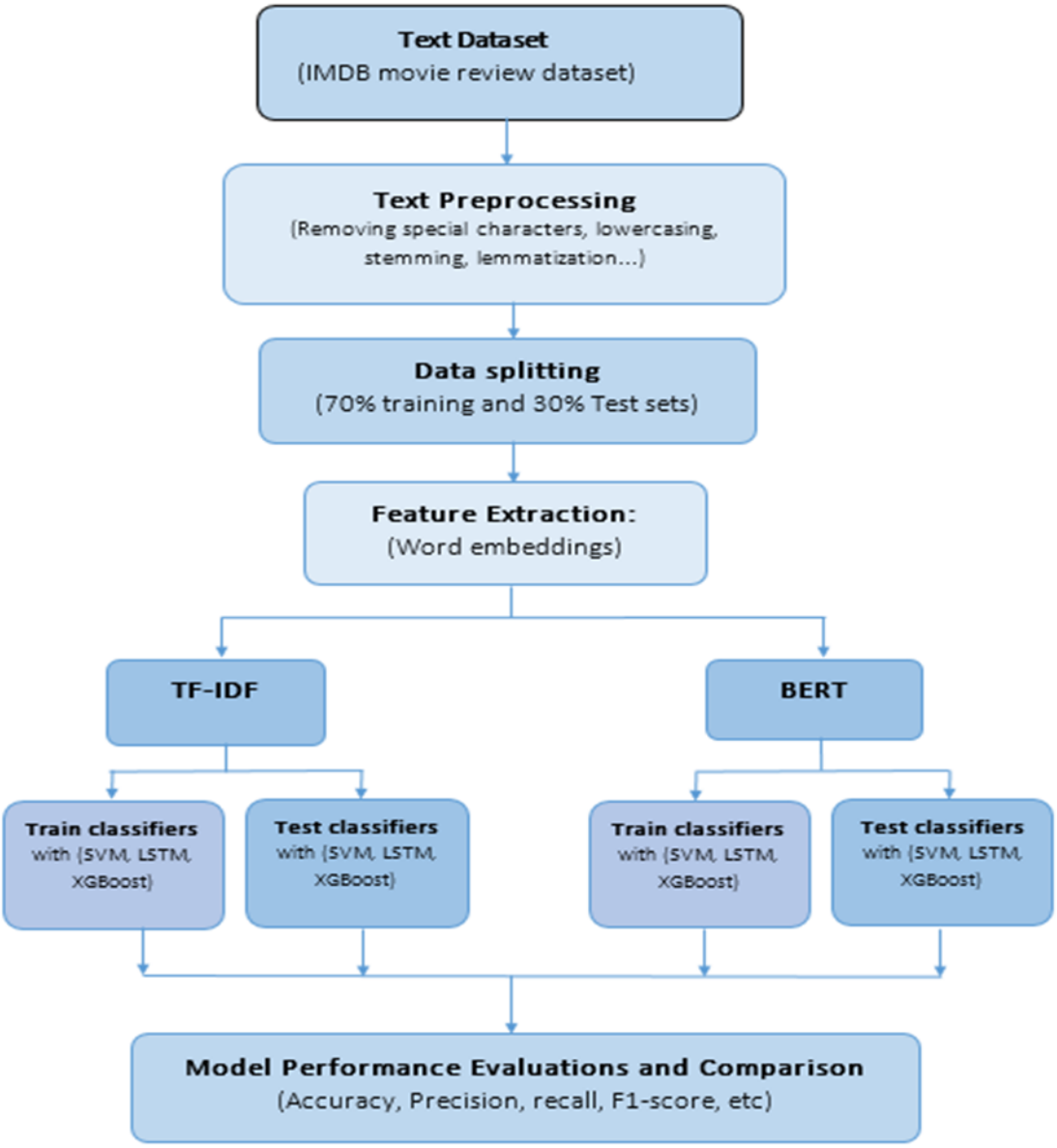

Fig. 1: Experimental process.

These two techniques were selected to enable a comparative analysis of static versus contextual embeddings in sentiment analysis.

Classification Models

Three classifiers were used to assess the performance of the embedding techniques:

- Support Vector Machine (SVM): A supervised learning algorithm effective for high-dimensional data. A linear kernel was used for text classification tasks with TF-IDF features.
- Extreme Gradient Boosting (XGBoost): A decision-tree-based ensemble method known for its speed and accuracy in classification problems.
- Long Short-Term Memory (LSTM): A type of recurrent neural network capable of capturing long-term dependencies in sequential data. LSTM was paired with BERT embeddings to evaluate its performance with contextual input.

Hyperparameters for each model were tuned using grid search and validation data to achieve optimal performance.

Evaluation Metrics

Model performance was evaluated using the following metrics: accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics were chosen to provide a comprehensive assessment of classification effectiveness, especially considering the balanced nature of the dataset.

Implementation Tools

All experiments were conducted using Python 3. Libraries used include:

- Scikit-learn for data preprocessing and implementation of SVM and XGBoost;
- TensorFlow and Keras for LSTM model training;
- Transformers (Hugging Face) for BERT-based embeddings.

The models were trained and evaluated on a standard computing environment with access to a GPU to accelerate deep learning operations.

EXPERIMENTAL RESULTS

In this study, we experimented with SVM, LSTM, and XGboost, as comparative models. The performance is compared based on the accuracy measurement according to the word embedding method of TF-IDF, and BERT. Table 1 and 2 shows results using TF-IDF, and BERT, respectively. In the word embedding of both models, SVM showed the highest accuracy for the test dataset. SVM is 89.85% in TF-IDF and 88.05% in BERT. The lowest accuracy is XGBoost (86.57%) for TF-IDF and LSTM (82.62%) for BERT. Overall, the average accuracy was 88.26% for TF-ID and 85.72% for BERT.

Table 1: Experimental result of TF-IDF (%)

Model	Sentiment	Accuracy	Precision	Recall	F1-Score	AUC-ROC
SVM	Negative (0)	89.85	91.00	89.00	90.00	90.00
	Positive (1)		89.00	91.00	90.00	
LSTM	Negative (0)	88.36	88.00	88.00	88.00	88.00
	Positive (1)		88.00	89.00	88.00	
XGBoost	Negative (0)	86.57	88.00	85.00	86.00	87.00
	Positive (1)		86.00	88.00	87.00	
Average		88.26	88.33	88.33	88.17	88.33

DISCUSSION

This study aimed to evaluate the effectiveness of different word embedding techniques, specifically TF-IDF and BERT, in sentiment analysis tasks involving movie reviews. The results indicate that the TF-IDF embedding, when paired with a Support Vector Machine (SVM) classifier, achieved the highest accuracy of 90%, marginally outperforming the BERT embedding combined with a Long Short-Term Memory (LSTM) network, which attained an accuracy of 88%. These findings suggest that traditional, sparse embedding methods like TF-IDF can be highly effective, particularly when used with linear classifiers such as SVM.

Table 2: Experimental result of BERT (%)

Model	Sentiment	Accuracy	Precision	Recall	F1-Score	AUC-ROC
SVM	Negative (0)	88.05	87.00	89.00	88.00	88.00
	Positive (1)		89.00	87.00	88.00	
LSTM	Negative (0)	82.62	85.00	79.00	82.00	83.00
	Positive (1)		81.00	86.00	83.00	
XGBoost	Negative (0)	86.50	86.00	87.00	86.00	87.00
	Positive (1)		87.00	86.00	86.00	
Average		85.72	85.83	85.67	85.50	86.00

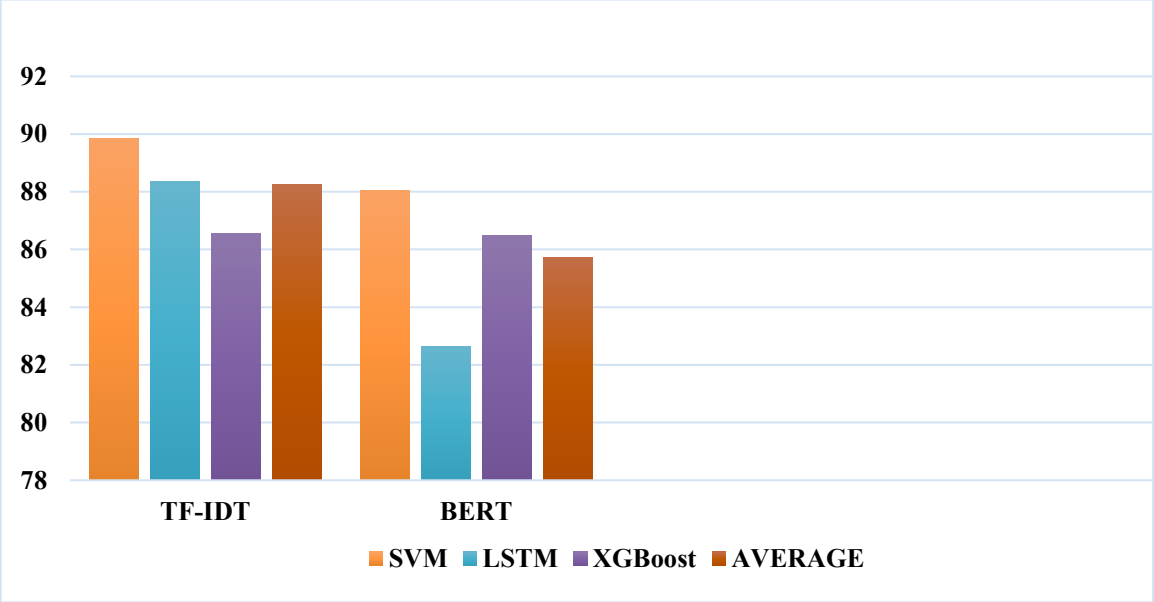

Figure 2: Shows the bar chart of the performances for each model.

This aligns with previous research indicating that TF-IDF, despite its simplicity, can yield competitive results in text classification tasks. For instance, Mamata *et al*. (2023) demonstrated that TF-IDF features, when used with traditional classifiers, achieved notable accuracy in sentiment analysis tasks on unstructured datasets. Conversely, while BERT embeddings offer contextualized representations and have shown superior performance in various natural language processing tasks, their effectiveness can be influenced by the choice of classifier and the nature of the dataset. González-Carvajal and Garrido-Merchán (2020) observed that BERT's performance in text classification tasks can vary depending on the specific implementation and the characteristics of the data.

The implications of these results are significant for practitioners in the field of sentiment analysis. They suggest that, in certain contexts, traditional embedding methods like TF-IDF, when combined with appropriate classifiers, can match or even surpass the performance of more complex models like BERT. This has practical relevance, especially

in scenarios where computational resources are limited or where model interpretability is a priority.

Future research could explore the integration of TF-IDF and BERT embeddings to leverage the strengths of both approaches. Additionally, examining the performance of these embeddings across different domains and languages could provide further insights into their generalizability and robustness.

REFERENCES

Asha, K.P., Kumar, S.R. and Raman, R. (2023). A BERT-CNN based approach on movie review sentiment analysis. *SHS Web of Conferences*, 140: 04007. https://doi.org/10.1051/shsconf/202314004007

González-Carvajal, L. and Garrido-Merchán, E.C. (2020). Comparing BERT against traditional machine learning text classification. *arXiv preprint*, arXiv:2005.13012. https://doi.org/10.48550/arXiv.2005.13012

Kumar, R. and Bansal, P. (2023). Sentiment analysis on movie reviews dataset using support vector machines and ensemble learning. *International Journal of Scientific Research in Computer Science*, 11(1): 34–41. https://www.researchgate.net/publication/364523233

Mamata, D., Selvakumar, K., & Alphonse, P. J. A. (2023). A Comparative Study on TF-IDF Feature Weighting Method and Its Analysis Using Unstructured Dataset. *arXiv preprint arXiv:2308.04037*. https://arxiv.org/abs/2308.04037

Nair, M., Harshitha, C.H. and Adiga, P. (2024). Movie review sentiment analysis using ensemble models. *Procedia Computer Science*, 229: 441–448. https://doi.org/10.1016/j.procs.2024.03.048

Nkhata, D., Anjum, U. and Zhan, J. (2025). Sentiment analysis using BERT for movie reviews. *International Journal of Advanced Computer Science and Applications*, 16(2): 120–128. . https://doi.org/10.48550/arXiv.2502.18841

Nurul, S.M. and Roziati, Z. (2023). Sentiment analysis with LSTM recurrent neural network approach for movie reviews using deep learning. *International Journal of Advanced Computer Science and Applications*, 14(4): 79–87. https://www.researchgate.net/publication/379926827

Rathor, A., Dhyani, M., Goyal, A. and Prasad, K. (2024). Sentiment analysis on Twitter data using BERT and deep learning techniques. *PeerJ Computer Science*, 10: e1459. https://doi.org/10.7717/peerj-cs.1459

Subramaniyaswamy, V., Mohan, S. and Chandrasekar, A. (2024). Effectiveness of AdaBoost and XGBoost algorithms in sentiment analysis of movie reviews. *Journal of Intelligent & Fuzzy Systems*, 46(2): 2099–2111. https://doi.org/10.3233/JIFS-223524

Xu, M. (2024). Social media and its implications for opinion mining: A modern review. *Journal of Information Science and Engineering*, 40(1): 101–115.

Yang, F. (2024). Harnessing user-generated content for business insights. *Digital Business Review*, 18(2): 55–68.

Zhou, Q. and Liu, Y. (2023). Sentiment analysis of the consumer review text based on BERT-BiLSTM in a social media environment. *International Journal of Information Technology and Web Engineering*, 18(1): 1–17. https://doi.org/10.4018/IJITWE.325618.