Journal of Environmental Research and Engineering

JERE, 2026, 2 (1): 1-6

Sentiment Analysis on the IMDb Dataset: A Comparative Study of Classical, Deep Learning, and Transformer-Based Models for

Osman Toplu *
*Computer Engineering, Engineering and Natural Science Faculty, Iskenderun Technical University, Iskenderun, Türkiye

osmantoplu.lee25@iste.edu.tr

ABSTRACT

Text data analysis, particularly sentiment analysis, has become a critical research area due to the rapid growth of user-generated content on digital platforms. Advances in natural language processing (NLP) have enabled the automatic analysis of opinions expressed in movie reviews, social media posts, and customer feedback, making sentiment classification an essential task in both academic research and industrial applications. This study aims to conduct a comparative analysis of classical machine learning, deep learning, and transformer-based models for sentiment analysis using the IMDb 50K dataset.

ARTICLE INFO

Received 2026-05-27,
Accepted 2026-06-29,
Publication Date 2026-06-30

Keywords: Sentiment Analysis, IMDb Dataset, Machine Learning, Deep Learning

Corresponding Author: Osman Toplu, E-mail: osmantoplu.lee25@iste.edu.tr

ÖZET

Metin veri analizi, özellikle de duygu analizi, dijital platformlarda kullanıcı tarafından oluşturulan içeriklerin hızla büyümesi nedeniyle kritik bir araştırma alanı haline gelmiştir. Doğal dil işlemedeki (NLP) ilerlemeler; film incelemelerinde, sosyal medya paylaşımlarında ve müşteri geri bildirimlerinde ifade edilen görüşlerin otomatik olarak analiz edilmesini sağlayarak duygu sınıflandırmasını hem akademik araştırmalarda hem de endüstriyel uygulamalarda temel bir görev haline getirmiştir. Bu çalışma, IMDb 50K veri kümesini kullanarak duygu analizi için klasik makine öğrenmesi, derin öğrenme ve dönüştürücü tabanlı (transformer) modellerin karşılaştırmalı bir analizini yapmayı amaçlamaktadır.

MAKALE BİLGİSİ

Keywords:
Sentiment Analiz, IMDb Dataset, Makine Ögrenmesi, Derin Ögrenme

INTRODUCTION

Text data analysis, particularly sentiment analysis, has become a critical research area due to the rapid growth of user-generated content on digital platforms. Advances in natural language processing (NLP) have enabled the automatic analysis of opinions expressed in movie reviews, social media posts, and customer feedback, making sentiment classification an essential task in both academic research and industrial applications (Pang & Lee, 2008; Atayolu and Kutlu, 2023; Jurafsky & Martin, 2023; Urban, 2024; Oncu and Kutlu, 2024). Early sentiment analysis studies primarily relied on statistical methods and rule-based approaches; however, these techniques often struggled to capture the semantic and contextual complexity of natural language (Manning et al., 2008).

The primary challenge in sentiment analysis arises from the inherent ambiguity and variability of human language. Similar emotional states can be expressed using different syntactic structures, lexical choices, and contextual cues. As a result, classical machine learning approaches based on handcrafted features often fail to achieve high classification accuracy. In contrast, deep learning-based artificial intelligence models are capable of learning hierarchical and contextual representations from large-scale datasets, allowing them to model semantic relationships more effectively (Hochreiter & Schmidhuber, 1997).

The evolution of sentiment analysis methods is clearly reflected in studies conducted on the IMDb 50K movie review dataset. Maas et al. (2011) introduced this large-scale benchmark dataset, which has since been widely adopted for evaluating sentiment classification models. Subsequent studies have demonstrated the effectiveness of both classical and deep learning approaches on this dataset. For example, Tomal (2024) reported that TF–IDF-based Logistic Regression achieved strong baseline performance among classical methods, while tree-based models such as Random Forest and Decision Tree produced comparable results.

Deep learning-based models have further improved sentiment classification performance by capturing sequential and contextual dependencies within text. Kanse et al. (2025) showed that recurrent neural network architectures, particularly LSTM models, were able to learn long-range dependencies more effectively than traditional methods, achieving accuracy levels above 90%. Convolutional neural networks have also been successfully applied to sentence classification tasks by learning local n-gram patterns through convolutional filters (Kim, 2014).

More recently, transformer-based language models have become the dominant approach in sentiment analysis. Devlin et al. (2019) demonstrated that BERT, which processes text bidirectionally using self-attention mechanisms, significantly outperforms recurrent and convolutional models. Building upon this architecture, Liu et al. (2019) proposed RoBERTa, an optimized variant of BERT that improves performance through enhanced pretraining strategies. Studies such as Alaparthi and Mishra (2020) have shown that fine-tuned transformer-based models achieve state-of-the-art results on the IMDb dataset, often exceeding 94% accuracy.

Based on these findings, this study aims to conduct a comparative analysis of classical machine learning, deep learning, and transformer-based models for sentiment analysis using the IMDb 50K dataset. The dataset and methodological framework adopted in this study are described in the following section.

MATERIALS AND METHODS

Dataset

The dataset used in this study is the IMDb 50K movie reviews dataset, originally introduced by Maas et al. (2011). It consists of 50,000 movie reviews evenly distributed between positive and negative sentiment classes. The dataset has been preprocessed to remove HTML tags and other noisy elements and is commonly used as a benchmark for sentiment analysis research.

Machine Learning Models

Several machine learning and deep learning models are considered in this study.

Logistic Regression is a linear classification method applied to TF–IDF-based text representations. Due to its simplicity and computational efficiency, it serves as a strong baseline model for sentiment classification tasks (Pang & Lee, 2008).

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network designed to capture long-range dependencies in sequential data. By maintaining internal memory states, LSTM models are particularly effective in modeling contextual relationships in text (Hochreiter & Schmidhuber, 1997).

Convolutional Neural Networks (CNNs) perform sentence classification by extracting local textual patterns using convolutional filters. CNN-based models have been shown to be highly effective in capturing n-gram features and local semantic information (Kim, 2014).

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based language model that learns bidirectional contextual representations through self-attention mechanisms. BERT-based models have consistently achieved high performance in sentiment analysis tasks (Devlin et al., 2019).

RoBERTa with Deep Neural Networks (RoBERTa + DNN) is an optimized extension of BERT that improves performance through enhanced pretraining procedures and fine-tuning strategies, leading to superior classification results (Liu et al., 2019).

Performance Metrics

The performance of the models is evaluated using standard classification metrics. Accuracy measures the proportion of correctly classified instances, while **precision** reflects the correctness of positive predictions**. Recall** indicates the proportion of actual positive instances that are correctly identified, and the **F1-score** represents the harmonic mean of precision and recall. These metrics are widely used in machine learning evaluations and are implemented using established libraries such as Scikit-learn (Pedregosa et al., 2011).

RESULTS AND DISCUSSION

This study presents a literature-based experimental scenario comparing Logistic Regression, CNN, LSTM, BERT, and RoBERTa models on the IMDb 50K dataset. Previous research consistently indicates that deep learning-based approaches outperform classical machine learning methods in sentiment classification tasks. Logistic Regression provides a strong baseline; however, its performance remains limited due to its reliance on sparse feature representations.

CNN models improve classification performance by effectively capturing local textual patterns, while LSTM models further enhance results by modeling long-range contextual dependencies. Transformer-based models, particularly BERT and RoBERTa, achieve the highest performance by leveraging bidirectional contextual representations and large-scale pretraining. Among all evaluated approaches, RoBERTa-based models are reported to deliver the most consistent and accurate results in the literature (Devlin et al., 2019; Liu et al., 2019).

Overall, findings from prior studies demonstrate that transformer-based architectures significantly outperform classical and recurrent neural network-based approaches in sentiment analysis tasks.

ACKNOWLEDGEMENTS

The author acknowledges the use of artificial intelligence–based language tools for linguistic refinement, grammar correction, and structural suggestions during the preparation of this manuscript. The scientific content, methodological design, literature selection, and interpretation of results were conducted entirely by the author.

REFERENCES

Alaparthi, S., & Mishra, M. (2020). BERT: A sentiment analysis odyssey. International Journal of Engineering and Advanced Technology, 9(4), 263–268.

Atayolu, Y., Kutlu, Y. (2023). Effective Use of Content as a Feature in IMDB Dataset Analysis. Journal of Artificial Intelligence with Applications, 4(1), 1-8, doi : 10.5281/zenodo.14587323

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT) (pp. 4171–4186). Association for Computational Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Jurafsky, D., & Martin, J. H. (2023). Speech and language processing (3rd ed., draft). Stanford University.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1746–1751). Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (pp. 142–150). Association for Computational Linguistics.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Oncu, M., Kutlu, Y. (2024). Performance Comparison of Known AI Translation Tools for Turkish Language. Tethys Environmental Science, 1(3), 117-126, doi : 10.5281/zenodo.13269233

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1–2), 1–135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Urban, D. (2024). Integrating Textual and Visual Data For Social Media Analysis: Insights Into Sentiment, Behavior, and Trends. Journal of Artificial Intelligence with Applications, 5(1), 18-21, doi : 10.5281/zenodo.14644001