

A Reproducible Evaluation Framework for TF-IDF-Based Machine Learning in Sentiment Classification

Mr. Bhusare Ganesh Pravin¹ and Prof. J. R. Mahajan²

Student, Department of Artificial Intelligence And Data Science¹

Guide, Department of Artificial Intelligence And Data Science²

Adsul Technical Campus, Chas, Ahilyanagar, India

Abstract: Sentiment analysis is an important Natural Language Processing (NLP) task used to automatically identify the emotional orientation expressed in textual data. Although recent research increasingly focuses on deep learning and transformer-based language models, classical machine-learning methods combined with Term Frequency-Inverse Document Frequency (TF-IDF) remain useful because of their relatively low computational requirements and reproducibility. This paper presents a controlled evaluation framework for comparing classical machine-learning classifiers for binary sentiment classification using the Large Movie Review Dataset (IMDb). The dataset contains 50,000 labeled reviews divided into 25,000 training and 25,000 testing reviews. Logistic Regression, Multinomial Naive Bayes, and Linear Support Vector Machine are evaluated using a common preprocessing and TF-IDF pipeline. On the fixed 25,000-review test set, Linear SVM achieved the highest accuracy of 90.196% and F1-score of 90.153%, followed by Logistic Regression with 89.936% accuracy and 89.980% F1-score. Multinomial Naive Bayes achieved 87.280% accuracy and 86.933% F1-score. The results demonstrate that controlled classical-model baselines can provide strong performance while remaining computationally accessible and reproducible.

Keywords: Sentiment Analysis, Natural Language Processing, TF-IDF, Machine Learning, Logistic Regression, Naive Bayes, Support Vector Machine, Text Classification, IMDb.

I. INTRODUCTION

The rapid growth of digital platforms has generated a large amount of textual information through online reviews, social media, discussion forums, customer feedback, and other digital communication channels. This information contains opinions and emotions that can provide useful insights to organizations and researchers. However, manually analyzing large volumes of textual data is time-consuming and difficult. Sentiment analysis provides a computational approach for automatically identifying the sentiment expressed in text.

Sentiment analysis is a subfield of Natural Language Processing that aims to determine whether a text expresses a positive, negative, or sometimes neutral opinion. Depending on the application, sentiment analysis can be formulated as document-level, sentence-level, or aspect-level classification. Recent literature has examined traditional machine learning, deep learning, large language models, and pretrained models for this task [1], [2].

Classical machine-learning approaches remain useful baselines because they can operate effectively with sparse textual representations while requiring less computational infrastructure than many modern neural approaches. TF-IDF is a commonly used statistical representation for converting textual documents into numerical feature vectors [4].

The objective of this work is to provide a reproducible comparison of three classical classifiers under the same TF-IDF representation and fixed IMDb train-test split. The study evaluates Logistic Regression, Multinomial Naive Bayes, and Linear Support Vector Machine using accuracy, precision, recall, F1-score, and confusion matrices.



II. RELATED WORK

Recent sentiment-analysis research covers traditional machine learning, deep learning, pretrained language models, and large language models. Recent reviews identify challenges such as context, sarcasm, domain adaptation, multilingual text, explainability, and computational cost [1], [2].

Mao, Liu, and Zhang presented a systematic literature review covering sentiment-analysis methods, applications, and challenges. Their review emphasizes that model selection and evaluation should be considered together with the task and application domain [2].

Islam et al. reviewed deep-learning approaches and discussed preprocessing, text representation, model architectures, limitations, and future directions. Their work also demonstrates the continuing interest in stronger neural approaches for sentiment analysis [3]. Bashiri and Naderi further reviewed transformer-based sentiment-analysis models and compared several transformer families across multiple datasets [7].

Recent experimental studies also continue to use classical models as baselines. Brandão et al. evaluated optimized classical machine-learning models for social-media sentiment analysis, while El Majid and Nuari compared classical models with BERT-based approaches across multiple datasets [6], [8]. These studies show that classical models remain useful for establishing accessible baselines. The present study differs by keeping the dataset, preprocessing, TF-IDF configuration, train-test split, and evaluation protocol fixed while changing only the classifier.

Study	Year	Focus	Main relevance to this work
Jim et al. [1]	2024	State-of-the-art review	Identifies current methods, challenges, and research directions.
Mao et al. [2]	2024	Systematic literature review	Reviews methods, applications, challenges, and model selection.
Islam et al. [3]	2024	Deep-learning review	Shows the move toward deep learning while discussing limitations and future scope.
Brandão et al. [6]	2025	Classical ML for social media	Shows that optimized classical models remain useful sentiment baselines.
Bashiri & Naderi [7]	2024	Transformer comparison	Shows the growing use of transformer models across multiple datasets.
El Majid & Nuari [8]	2025	Classical ML vs. BERT	Shows that model performance depends on dataset and language characteristics.
Present study	2026	Controlled TF-IDF + classical ML	Fixes the complete feature pipeline and varies only the classifier on IMDb.

III. TEXT PREPROCESSING

HTML markup is removed from the raw review text, the reviews are converted to lowercase, and consecutive whitespace is normalized. The TF-IDF vectorizer handles tokenization and creates unigram and bigram features. Exactly the same transformed training and testing representations are supplied to all three classifiers to maintain a controlled comparison.

IV. TF-IDF FEATURE REPRESENTATION

Term Frequency-Inverse Document Frequency (TF-IDF) is used to convert reviews into numerical sparse vectors. Term frequency reflects the importance of a term within a document, while inverse document frequency reduces the contribution of terms appearing in many documents.

The TF-IDF representation can be expressed as: $TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$, where t is a term and d is a document. The vectorizer uses unigrams and bigrams, minimum document frequency of 2, maximum document frequency of 0.95, sublinear term frequency, Unicode accent stripping, and a maximum vocabulary size of 100,000 features.



V. DATASET

The Large Movie Review Dataset (IMDb) introduced by Maas et al. is used as the benchmark dataset [5]. It contains 50,000 labeled movie reviews. The supplied dataset contains 25,000 training reviews and 25,000 testing reviews. Each split contains 12,500 positive and 12,500 negative reviews.

VI. CLASSIFIERS

Logistic Regression is a linear classification algorithm that estimates class probabilities from the input feature vector and is suitable for high-dimensional sparse TF-IDF representations.

Multinomial Naive Bayes is a probabilistic classifier commonly used for text classification. It models class-conditional feature probabilities and is computationally efficient for sparse text data.

Linear Support Vector Machine learns a linear decision boundary and seeks a large-margin separation between classes. It is widely applicable to high-dimensional sparse representations such as TF-IDF. The suitability of linear SVMs for high-dimensional text classification has also been established in earlier text-classification research [10].

VII. EXPERIMENTAL METHODOLOGY

The experiment uses the original fixed IMDb train-test split. Raw reviews are cleaned by removing HTML markup, converting text to lowercase, and normalizing whitespace. A single TF-IDF vectorizer is fitted only on the training reviews and is then used to transform both training and testing reviews. This prevents information from the test set from influencing feature construction.

The TF-IDF configuration is `ngram_range=(1,2)`, `min_df=2`, `max_df=0.95`, `sublinear_tf=True`, `max_features=100000`, and `strip_accents=unicode`. Logistic Regression is configured with `max_iter=1000`, `solver=liblinear`, and `random_state=42`. Multinomial Naive Bayes uses its standard implementation. Linear SVM uses `LinearSVC` with `random_state=42`.

After training, every classifier predicts the same 25,000 test reviews. Accuracy, precision, recall, F1-score, and confusion matrices are calculated from the predictions.

VIII. EVALUATION METRICS

Accuracy measures the proportion of correctly classified test examples. Precision measures the proportion of positive predictions that are correct. Recall measures the proportion of actual positive examples identified correctly. F1-score is the harmonic mean of precision and recall. Confusion matrices are used to examine true-negative, false-positive, false-negative, and true-positive predictions.

X. RESEARCH GAP AND CONTRIBUTION

Recent sentiment-analysis research increasingly emphasizes deep learning, pretrained language models, and large language models [1], [2], [7]. These approaches can capture richer semantic and contextual information, but their experiments often involve different datasets, feature representations, preprocessing pipelines, or optimization procedures. As a result, reported performance values are not always directly comparable across studies.

The contribution of this study is a controlled and reproducible baseline in which the IMDb dataset, preprocessing steps, TF-IDF representation, train-test split, classifier settings, and evaluation metrics are explicitly fixed while the classifier is varied. This design isolates the effect of the classifier under one common representation. The measured results provide a transparent reference point for future experiments involving other datasets, neural models, or transformer-based representations.

XI. LIMITATIONS

The dataset represents movie reviews and may not generalize directly to other domains such as products, finance, healthcare, or social media. The task is restricted to binary sentiment classification. TF-IDF is based on statistical term



weighting and does not explicitly represent deep semantic relationships or long-range context. The study does not compare the classical models directly with transformer-based language models. The selected preprocessing and TF-IDF hyperparameters form one reproducible configuration and are not an exhaustive hyperparameter search.

XII. FUTURE SCOPE

Future work can evaluate the same reproducible framework across product, application, social-media, and customer-feedback datasets. Additional experiments can compare word embeddings and transformer representations against TF-IDF, investigate multilingual and Indian-language sentiment analysis, examine aspect-based sentiment classification, and evaluate the trade-off between classification quality and computational cost.

XIII. CONCLUSION

This study presented and experimentally evaluated a reproducible TF-IDF-based framework for binary sentiment classification. Using the supplied IMDb dataset and a common feature pipeline, three classical classifiers were compared on the fixed 25,000-review test set. Linear SVM achieved the highest measured accuracy (90.196%) and F1-score (90.153%), while Logistic Regression produced closely competitive performance (89.936% accuracy). Multinomial Naive Bayes obtained lower performance under the same TF-IDF configuration (87.280% accuracy). The findings show that classical linear models can provide strong sentiment-classification baselines while remaining relatively simple and reproducible. The resulting framework can serve as a controlled baseline for future experiments involving additional domains and modern NLP models.

XIV. REPRODUCIBILITY AND RESEARCH INTEGRITY

The reported numerical results were generated by executing the described pipeline on the supplied IMDb dataset. No performance values were copied from another paper. The dataset split, preprocessing, TF-IDF configuration, classifier settings, and evaluation procedure are stated so that the experiment can be reproduced. Figures in this paper were generated from the measured predictions produced by the experiment.

The manuscript uses original wording and cites the underlying research and dataset. The additional literature sources were included to strengthen the related-work discussion and research gap. Before journal submission, the author should perform the journal-required similarity/originality check and review the final manuscript and references.

Characteristic	Description
Dataset	Large Movie Review Dataset (IMDb)
Total labeled reviews	50,000
Training reviews	25,000
Testing reviews	25,000
Training positive / negative	12,500 / 12,500
Testing positive / negative	12,500 / 12,500
Task	Binary sentiment classification
Language	English
Domain	Movie reviews

TABLE I. CHARACTERISTICS OF THE IMDb DATASET

IX. RESULTS AND DISCUSSION

The experiment was executed on the supplied IMDb dataset using the fixed 25,000-review test set. The following values are measured results from the described pipeline.

Classifier	Accuracy	Precision	Recall	F1-Score
Logistic Regression	89.936%	89.588%	90.376%	89.980%



Multinomial Naive Bayes	87.280%	89.371%	84.624%	86.933%
Linear SVM	90.196%	90.550%	89.760%	90.153%

The literature comparison indicates a clear practical gap: many studies compare models while also changing datasets, representations, preprocessing, or optimization procedures. This makes direct model-level conclusions difficult. The present work therefore focuses on a controlled baseline experiment in which the major pipeline components are fixed.

TABLE II. MEASURED PERFORMANCE OF THE THREE CLASSIFIERS ON THE IMDb TEST SET

Linear SVM achieved the highest test accuracy of 90.196% and the highest F1-score of 90.153%. Logistic Regression followed closely with 89.936% accuracy and 89.980% F1-score. Multinomial Naive Bayes achieved 87.280% accuracy and 86.933% F1-score. The results indicate that the two linear decision-based approaches were stronger than Multinomial Naive Bayes under this fixed TF-IDF configuration.

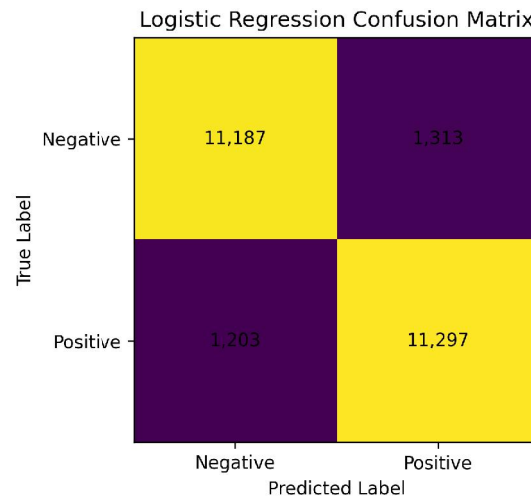


Fig. 1. Confusion matrix for Logistic Regression.

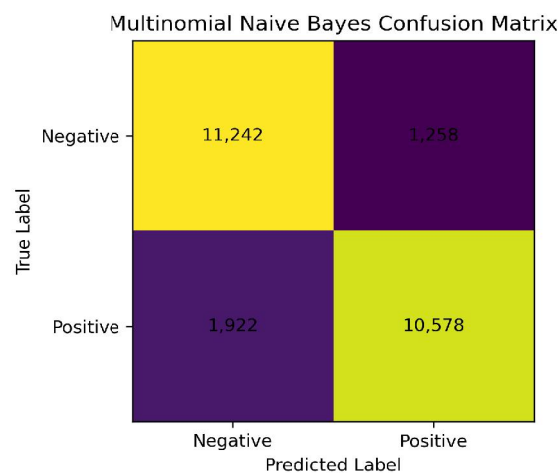


Fig. 2. Confusion matrix for Multinomial Naive Bayes.



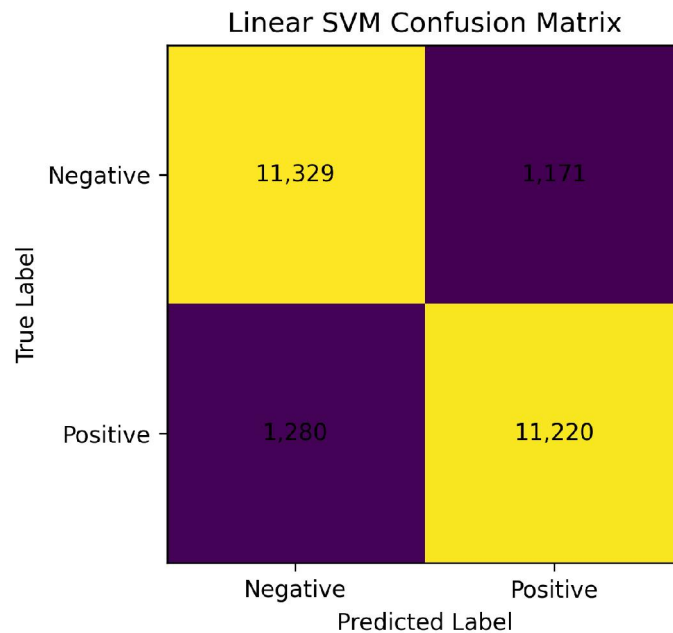


Fig. 3. Confusion matrix for Linear SVM.

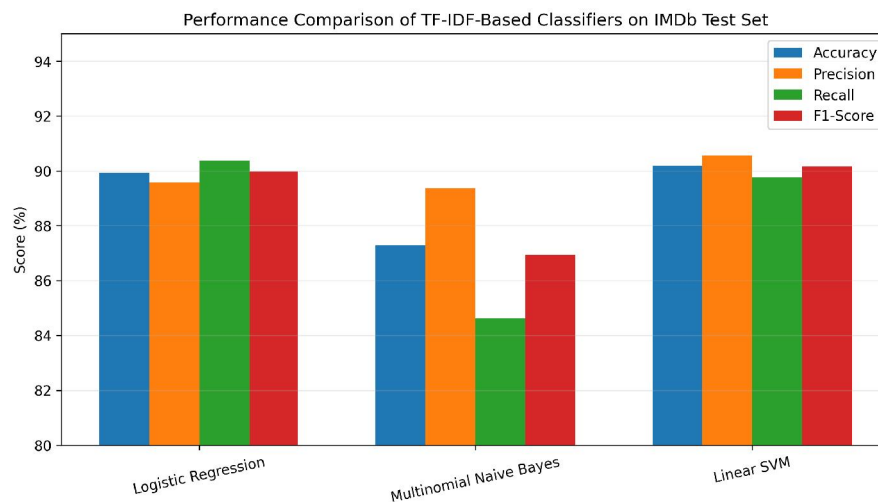


Fig. 4. Comparison of measured classifier performance.

Logistic Regression produced 11,187 true negatives, 1,313 false positives, 1,203 false negatives, and 11,297 true positives on the 25,000-review test set.

Multinomial Naive Bayes produced 11,242 true negatives, 1,258 false positives, 1,922 false negatives, and 10,578 true positives on the 25,000-review test set.

The confusion matrices provide additional insight into the differences between the classifiers. Linear SVM produced the lowest number of false positives among the three models and a substantially lower number of false negatives than Multinomial Naive Bayes. Logistic Regression produced a similar error pattern to Linear SVM, which is consistent with both models being linear classifiers operating on the same sparse TF-IDF representation.



A qualitative inspection of representative Linear SVM errors shows several recurring types of difficult reviews. Some reviews contain mixed positive and negative opinions, where positive words occur in an overall negative review or vice versa. Other errors involve sarcasm or indirect criticism, where the literal words do not match the intended sentiment. Some reviews contain long plot descriptions followed by a short opinion, making the sentiment signal relatively weak. Context-dependent expressions and comparisons between films also create difficulty for a bag-of-words representation. These observations explain an important limitation of TF-IDF: it represents the importance of terms and short n-grams but does not explicitly model deeper semantic relationships or long-range context. Therefore, the results support the use of classical models as strong and accessible baselines, while also motivating future comparison with contextual representations such as transformer models. This error analysis is qualitative and is intended to identify representative error patterns rather than claim a statistical distribution of error types.

The IMDb dataset used in the experiment is the original Large Movie Review Dataset introduced by Maas et al. It contains 25,000 labeled training reviews and 25,000 labeled test reviews, with balanced positive and negative classes in each split [5].

Linear SVM produced 11,329 true negatives, 1,171 false positives, 1,280 false negatives, and 11,220 true positives on the 25,000-review test set.

REFERENCES

- [1] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 6, Art. no. 100059, 2024, doi: 10.1016/j.nlp.2024.100059.
- [2] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 4, Art. no. 102048, 2024.
- [3] M. S. Islam, M. N. Kabir, N. A. Ghani, K. Z. Zamli, N. S. A. Zulkifli, M. M. Rahman, and M. A. Moni, "Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach," *Artificial Intelligence Review*, vol. 57, Art. no. 62, 2024.
- [4] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.
- [5] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon, USA, 2011, pp. 142–150.
- [6] J. G. Brandão, A. P. Castro Junior, V. M. G. Pacheco, C. G. Rodrigues, O. M. O. Belo, A. P. Coimbra, and W. P. Calixto, "Optimization of machine learning models for sentiment analysis in social media," *Information Sciences*, vol. 694, Art. no. 121704, 2025, doi: 10.1016/j.ins.2024.121704.
- [7] H. Bashiri and H. Naderi, "Comprehensive review and comparative analysis of transformer models in sentiment analysis," *Knowledge and Information Systems*, vol. 66, no. 12, pp. 7305–7361, 2024, doi: 10.1007/s10115-024-02214-3.
- [8] A. U. El Majid and R. Nuari, "Performance comparison of BERT metrics and classical machine learning models (SVM, Naive Bayes) for sentiment analysis," *INOVTEK Polbeng – Seri Informatika*, vol. 10, no. 2, pp. 741–752, 2025, doi: 10.35314/wmh3rg23.
- [9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: Sentiment classification using machine learning techniques," in *Proc. ACL-02 Conf. Empirical Methods in Natural Language Processing*, vol. 10, 2002, pp. 79–86, doi: 10.3115/1118693.1118704.
- [10] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. 10th European Conference on Machine Learning (ECML)*, 1998, pp. 137–142, doi: 10.1007/BFb0026683.

