Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.35246

Traditional Machine Learning and DistilBERT for Sentiment Analysis

Lekang Sun

Beijing-Dublin International College, Beijing University of Technology, Beijing, China
lekang.sun@ucdconnect.ie

Abstract. Sentiment analysis identifies opinion polarity in textual data and supports decision-making in review-rich digital environments. This study compares three approaches for binary sentiment classification on the Internet Movie Database (IMDb) movie review dataset, namely Term Frequency-Inverse Document Frequency (TF-IDF)+Naive Bayes, TF-IDF+Logistic Regression, and DistilBERT, a distilled version of Bidirectional Encoder Representations from Transformers (BERT). The comparison focuses on the performance-efficiency trade-off rather than accuracy alone. Models are evaluated by accuracy, precision, recall, F1 score, training time, and inference time under 20%, 50%, and 100% training-data settings. A representative rule-based error analysis is also conducted. Under the full training-data setting, DistilBERT achieves the highest F1 score of 0.914496, ahead of Logistic Regression at 0.883284 and Naive Bayes at 0.846688. However, relative to Logistic Regression, DistilBERT is approximately 175.69 times slower in training and 51.05 times slower in inference. DistilBERT trained with only 20% of the training data still exceeds the full-data F1 scores of both traditional baselines. The findings indicate that model selection should jointly consider effectiveness, efficiency, data scale, and error characteristics.

Keywords: Sentiment Analysis, DistilBERT, Traditional Machine Learning, IMDb Movie Reviews, Error Analysis

1. Introduction

The rapid growth of online platforms, social media, e-commerce websites, and review systems has created large volumes of opinion-rich text. Sentiment analysis, also called opinion mining, aims to identify the emotional or evaluative orientation of such texts. In practical systems, this task can support public-opinion monitoring, customer-feedback analysis, product improvement, and decision-making in digital services. Because of these applications, sentiment analysis remains an important task in natural language processing [1].

Existing studies show that sentiment analysis has moved from mainly feature-based methods toward deep learning and pre-trained language models. Cui et al. reviewed 9,714 publications from 2002 to 2022 and found that the research focus has gradually shifted from traditional feature engineering to deep learning and pre-trained models [2]. Bordoloi and Biswas also point out that sentiment classification depends on several design choices, including preprocessing, feature representation, and the classifier itself. They further summarize sentiment analysis at document,

DOI: 10.54254/2755-2721/2026.35246

sentence or phrase, word, and aspect levels [3]. Bashiri and Naderi compare several Transformer models on 22 sentiment-analysis datasets and report that strong contextual models do not always provide the same balance between performance and efficiency across datasets [4]. These studies suggest that model choice should not be judged only by final accuracy.

This paper therefore compares TF-IDF + Naive Bayes, TF-IDF + Logistic Regression, and DistilBERT on IMDb movie reviews. The comparison is designed around four questions: how well the models classify sentiment, how much time they require for training and inference, how their performance changes when the training data is reduced, and what types of reviews remain difficult. The purpose is to examine whether the additional performance of DistilBERT justifies its higher computational cost in a practical sentiment-analysis setting.

2. Methodology

2.1. Dataset

This study uses the IMDb Large Movie Review Dataset for binary sentiment classification [5]. The task is to determine whether a review expresses positive or negative sentiment. The dataset contains 25,000 training reviews and 25,000 testing reviews with a balanced class distribution. IMDb reviews are appropriate for this comparison because they are document-level texts that often include contextual transitions, mixed evaluation, and long-form opinion expression.

2.2. Models and representation

Traditional models use Term Frequency-Inverse Document Frequency (TF-IDF) representation [6]. TF-IDF + Naive Bayes serves as a lightweight probabilistic baseline. Naive Bayes is widely used in text classification because of its simple conditional-independence structure and efficient training behavior [7]. TF-IDF + Logistic Regression is used as the stronger linear baseline. Large-scale linear classification is effective for sparse high-dimensional text features, and Logistic Regression provides a practical balance between performance and efficiency [8].

DistilBERT is selected as the compact contextual comparator. BERT established deep bidirectional language representations for downstream NLP tasks, while DistilBERT reduces model size through knowledge distillation while retaining much of BERT's language understanding capacity [9, 10]. DistilBERT receives the original review text rather than the cleaned traditional-model input so that richer contextual information can be preserved.

2.3. Experimental design

Traditional baselines are implemented with a scikit-learn pipeline [11]. Before feature extraction, the review text is lowercased, HTML line-break tags are removed, non-alphabetic characters are filtered, and extra whitespace is merged. TF-IDF is then used to represent the cleaned reviews. The vectorizer removes English stop words, keeps the 20,000 most frequent features, uses unigram and bigram features, ignores terms that appear in fewer than two documents, and removes terms that appear in more than 95% of documents. These settings are used to keep the feature space large enough for review-level sentiment classification while avoiding an unnecessarily large vocabulary.

Naive Bayes and Logistic Regression are trained on the TF-IDF features. Logistic Regression is trained with a maximum of 1,000 optimization iterations, and the random seed is fixed at 42 for reproducibility. DistilBERT is fine-tuned from the pretrained distilbert-base-uncased checkpoint for

DOI: 10.54254/2755-2721/2026.35246

two epochs, with a batch size of 16, a learning rate of 2e-5, and a weight decay of 0.01. For DistilBERT, a validation split is created from the training data and used for checkpoint selection. The checkpoint with the best validation F1 score is selected. During tokenization, each review is truncated or padded to a maximum length of 256 tokens.

The evaluation includes accuracy, precision, recall, F1 score, training time, and inference time. These metrics are used because the task requires both predictive quality and computational efficiency. To test the effect of training-data size, each model is trained with 20%, 50%, and 100% of the original training set, while the 25,000-review test set is kept unchanged for final evaluation.

The numerical results are supplemented with a rule-based error analysis. All test reviews misclassified by at least one model are first grouped into Error Patterns according to which models made incorrect predictions. Thirty samples are then drawn from each Error Pattern, giving 210 representative cases. These samples are assigned preliminary Error Categories and Possible Reasons. This analysis is exploratory rather than exhaustive, but it helps identify recurring sources of difficulty that are not visible from aggregate metrics alone.

All experiments are conducted in a Jupyter Notebook environment on a Windows 10 machine with Python 3.10.20, a 16-logical-core AMD64 CPU, 13.86 GB RAM, and an NVIDIA GeForce RTX 3050 Laptop GPU with 4 GB memory. The TF-IDF baselines are implemented with scikit-learn 1.7.2 and executed on CPU. DistilBERT fine-tuning and inference are performed with PyTorch 2.5.1 using CUDA acceleration. The reported training and inference times are therefore hardware-dependent wall-clock measurements in this experimental environment.

3. Results

3.1. Full training-data performance

Table 1 summarizes the final classification and efficiency outcomes under the full training-data setting. DistilBERT achieves the best predictive performance, with Accuracy = 0.913720 and F1 = 0.914496. Logistic Regression is the strongest traditional baseline, reaching F1 = 0.883284, while Naive Bayes records F1 = 0.846688. Relative to Logistic Regression, DistilBERT improves F1 by approximately 3.12 percentage points, but it requires substantially greater training and inference time.

Table 1. Final performance of the three models under the full training-data setting

Model	Accuracy	Precision	Recall	F1	Train Time (s)	Inference Time (s)
TF-IDF + Naive Bayes	0.849840	0.864842	0.829280	0.846688	9.649062	4.096136
TF-IDF + Logistic Regression	0.882840	0.879952	0.886640	0.883284	10.305945	4.292568
DistilBERT	0.913720	0.906341	0.922800	0.914496	1810.676478	219.141574

Note: The reported times are hardware-dependent wall-clock measurements. The scikit-learn TF-IDF baselines were executed on CPU, whereas DistilBERT was executed on a CUDA-enabled GPU.

3.2. Performance under different training-data sizes

The data-size experiment shows a stable ranking across reduced supervision. With 20% of the training data, F1 scores are 0.830359 for Naive Bayes, 0.857255 for Logistic Regression, and 0.897070 for DistilBERT. At 50%, the corresponding scores rise to 0.840883, 0.872116, and 0.905369. Under the full dataset, they reach 0.846688, 0.883284, and 0.914496. Figure 1 shows that

DOI: 10.54254/2755-2721/2026.35246

all three models improve as the training-data ratio increases. DistilBERT consistently achieves the highest F1 score, rising from 0.897 at 20% data to 0.914 at 100% data. Logistic Regression remains the strongest traditional baseline, while Naive Bayes gives the lowest F1 scores across all settings.

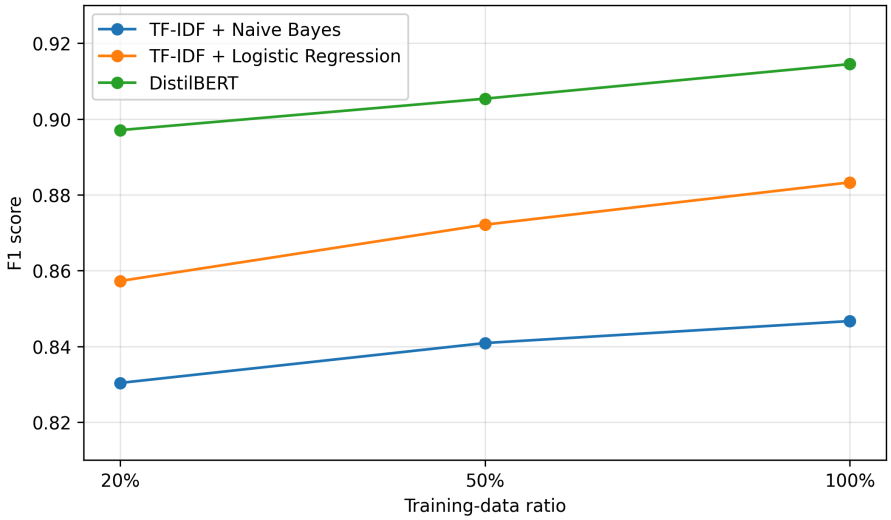

Figure 1. F1 score trends under 20%, 50%, and 100% training-data settings (photo credit: original)

3.3. Computational cost under different training-data sizes

Table 2 compares training and inference time across the three data-ratio settings. The TF-IDF baselines remain lightweight, their training time stays below eleven seconds even on the full dataset, and inference remains close to four seconds. DistilBERT exhibits a very different profile. Its training time increases from 715.901012 seconds at 20% data to 1810.676478 seconds at full data, while inference remains above 219 seconds in every setting. The training-cost gap widens as the fine-tuning data increases, while the inference-time gap remains consistently large because the same fixed test set is used.

Table 2. Training and inference time under different training-data sizes

Ratio	Training Size	Model	Train Time (s)	Inference Time (s)
20%	5,000	TF-IDF + Naive Bayes	2.265759	4.655697
20%	5,000	TF-IDF + Logistic Regression	2.322377	4.645489
20%	5,000	DistilBERT	715.901012	220.164963
50%	12,500	TF-IDF + Naive Bayes	5.536825	4.290586
50%	12,500	TF-IDF + Logistic Regression	5.082063	4.281756
50%	12,500	DistilBERT	1129.434779	222.123236
100%	25,000	TF-IDF + Naive Bayes	9.649062	4.096136
100%	25,000	TF-IDF + Logistic Regression	10.305945	4.292568
100%	25,000	DistilBERT	1810.676478	219.141574

Note: Timing measurements follow the same hardware-dependent setting described in Table 1.

DOI: 10.54254/2755-2721/2026.35246

3.4. Error-pattern evidence

Error-pattern statistics provide another perspective on model behavior. After excluding reviews correctly classified by all models, 5,465 misclassified reviews remain. Table 3 shows that the largest category is "Naive Bayes wrong, stronger models correct" at 25.86%, followed by "Traditional models wrong, DistilBERT correct" at 23.22%. At the same time, 16.01% of cases belong to "Traditional models correct, DistilBERT wrong," confirming that DistilBERT is not uniformly superior across every review type.

Representative rule-based categories indicate why aggregate scores should be supplemented with qualitative inspection. Table 4 summarizes the Error Category distribution among the 210 sampled errors. Negation / rhetorical framing accounts for 102 cases (48.57%), Genre-specific vocabulary for 75 cases (35.71%), and Long review / truncation risk for 15 cases (7.14%). Figure 2 shows that most representative errors are concentrated in two categories. Negation or rhetorical framing accounts for the largest share at 48.57%, followed by genre-specific vocabulary at 35.71%. Other categories, such as truncation risk, mixed sentiment, and weak implicit sentiment, occur much less frequently.

Table 3. Error Pattern summary among misclassified samples

Error Pattern	Count	Percentage
Naive Bayes wrong, stronger models correct	1,413	25.86%
Traditional models wrong, DistilBERT correct	1,269	23.22%
Traditional models correct, DistilBERT wrong	875	16.01%
All models wrong	824	15.08%
Logistic Regression wrong, others correct	626	11.45%
Only Logistic Regression correct	248	4.54%
Only Naive Bayes correct	210	3.84%
Total Misclassified Samples	5,465	100.00%

Table 4. Rule-based error category summary in the 210 representative samples

Error Category	Count	Percentage
Negation / rhetorical framing	102	48.57%
Genre-specific vocabulary	75	35.71%
Long review / truncation risk	15	7.14%
Ambiguous or mixed sentiment	6	2.86%
Weak or implicit sentiment	5	2.38%
Other context-dependent error	4	1.90%
Sentiment shift / contrast	3	1.43%
Total Representative Examples	210	100.00%

DOI: 10.54254/2755-2721/2026.35246

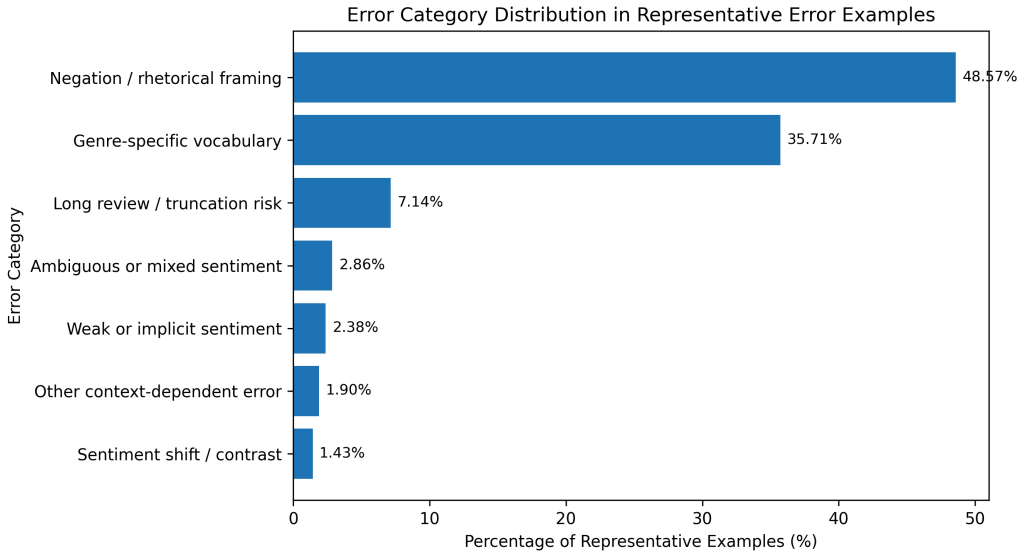

Figure 2. Percentage distribution of rule-based Error Categories in the 210 representative samples (photo credit: original)

4. Discussion

The results show that sentiment-model choice should be based on both predictive performance and computational cost. DistilBERT achieves the highest F1 score, but its advantage over Logistic Regression is not free. Under the full training-data setting, DistilBERT improves F1 by about 3.12 percentage points compared with Logistic Regression, yet it is about 175.69 times slower in training and 51.05 times slower in inference in the reported environment. Since the TF-IDF baselines are executed on CPU and DistilBERT uses GPU acceleration, the timing results should be interpreted as practical wall-clock measurements rather than a hardware-neutral benchmark. Even so, DistilBERT remains much slower, which supports the view that model performance should be considered together with latency, memory demand, and resource consumption [12-14]. For tasks where fast deployment or repeated retraining is important, TF-IDF + Logistic Regression may be a more practical choice.

The reduced-data experiment shows another side of DistilBERT. With only 20% of the training data, it still obtains an F1 score of 0.897070, which is higher than the full-data F1 scores of both traditional baselines. This suggests that pre-trained contextual representations can be useful when labeled data are limited, a pattern also discussed in low-resource sentiment studies [15]. However, this result should not be treated as a general rule. The experiment is limited to one balanced English movie-review dataset and one DistilBERT configuration.

The error analysis further shows why overall F1 is not enough for deployment decisions. Explainable sentiment-analysis research indicates that aggregate performance scores may hide systematic weaknesses in linguistically difficult cases [16]. In the representative errors, negation, rhetorical framing, and sarcasm-like polarity reversal appear frequently, which is consistent with previous studies on sentiment-analysis difficulties [17, 18]. Genre-specific vocabulary is also important in IMDb reviews, because words related to horror, violence, suspense, or darkness may describe the movie genre rather than the reviewer's final opinion. In addition, Long review / truncation risk cases suggest a possible limitation of using a maximum input sequence length of 256 tokens. Some long reviews may place the final sentiment after plot description or contrastive

DOI: 10.54254/2755-2721/2026.35246

comments, and this information may be lost after truncation. Since the error categories are based on representative rule-based annotation, future work should test larger manual coding, additional domains, and long-text or more efficient Transformer variants [19].

5. Conclusion

This paper compares TF-IDF + Naive Bayes, TF-IDF + Logistic Regression, and DistilBERT for binary sentiment classification on IMDb movie reviews. DistilBERT achieves the best predictive performance, while Logistic Regression remains a strong and efficient traditional baseline. Naive Bayes provides a lightweight reference model with lower computational cost.

The findings indicate that higher F1 score should be weighed against training and inference cost. DistilBERT keeps a clear advantage under 20%, 50%, and 100% training-data settings, showing the value of pre-trained contextual representations. At the same time, it requires much more computation than the TF-IDF baselines in the reported environment. The representative error analysis suggests that negation, rhetorical framing, genre-specific vocabulary, and long-review structure are important sources of difficulty. Therefore, practical sentiment-analysis systems should select models according to accuracy requirements, available resources, data scale, and expected error patterns. Future work can extend the comparison to other review domains, improve manual error verification, and test long-text or more efficient Transformer-based models.

References

[1] Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. Journal of King Saud University - Computer and Information Sciences, 36(4), Article 102048.
[2] Cui, J., Wang, Z., Ho, S.-B., & Cambria, E. (2023). Survey on sentiment analysis: Evolution of research methods and topics. Artificial Intelligence Review, 56, 8469–8510.
[3] Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. Artificial Intelligence Review, 56(11), 12505–12560.
[4] Bashiri, H., & Naderi, H. (2024). Comprehensive review and comparative analysis of transformer models in sentiment analysis. Knowledge and Information Systems, 66, 7305–7361.
[5] 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 ACL-HLT 2011 (pp. 142–150).
[6] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513–523.
[7] McCallum, A., & Nigam, K. (1998). A comparison of event models for naive Bayes text classification. In AAAI Workshop on Learning for Text Categorization (pp. 41–48).
[8] Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., & Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 9, 1871–1874.
[9] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171–4186).
[10] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. arXiv preprint arXiv: 1910.01108.
[11] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
[12] Chitty-Venkata, K. T., Mittal, S., Emani, M., Vishwanath, V., & Somani, A. K. (2023). A survey of techniques for optimizing transformer inference. Journal of Systems Architecture, 144, Article 102990.
[13] Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., & Shen, C. (2023). A survey on efficient training of transformers. In Proceedings of IJCAI-23 (pp. 6823–6831).
[14] Bolon-Canedo, V., Moran-Fernandez, L., Cancela, B., & Alonso-Betanzos, A. (2024). A review of green artificial intelligence: Towards a more sustainable future. Neurocomputing, 599, Article 128096.
[15] Chen, W., Lin, F., Li, G., & Liu, B. (2024). A survey of automatic sarcasm detection: Fundamental theories, formulation, datasets, detection methods, and opportunities. Neurocomputing, 578, Article 127428.

DOI: 10.54254/2755-2721/2026.35246

[16] Diwali, A., Saeedi, K., Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2024). Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis. IEEE Transactions on Affective Computing, 15(3), 837–846.
[17] Ilmawan, L. B., Muladi, & Prasetya, D. D. (2024). Negation handling for sentiment analysis task: Approaches and performance analysis. International Journal of Electrical and Computer Engineering, 14(3), 3382–3393.
[18] Nuci, K. P., Landes, P., & Di Eugenio, B. (2024). RoBERTa low resource fine tuning for sentiment analysis in Albanian. In Proceedings of LREC-COLING 2024 (pp. 14146–14151).
[19] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. ACM Computing Surveys, 55(6), Article 109, 1–28.