Sentiment Analysis Revisited: A Multi-Metric Comparative Study

Kamal Walji[1], Allae Erraissi[2], Abdelali ZAKRANI[3], Mouad Banane[4]
Laboratory of Artificial Intelligence & Complex Systems Engineering, Hassan II University,
ENSAM and Casablanca, Morocco[1, 3, 4]
Chouaib Doukkali University, El Jadida, Morocco[2]

***Abstract*—Sentiment analysis is a fundamental task in natural language processing with wide-ranging applications, from customer feedback monitoring to healthcare and social media analytics. While recent research has mainly emphasized predictive accuracy, computational efficiency has remained largely overlooked, despite its importance for large-scale and real-time deployment. This study addresses this gap by conducting a comparative evaluation of classical machine learning algorithms (Logistic Regression, Naïve Bayes, Random Forest) and deep learning architectures [Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM)]. Experiments were carried out on two benchmark datasets, IMDB and Yelp Polarity, with evaluation based on accuracy, precision, recall, F1-score, training time, and a novel Efficiency Score. Results on IMDB show that Logistic Regression and LSTM both achieved 88% accuracy, but with radically different costs: Logistic Regression trained in 0.25 seconds, whereas LSTM required more than 2600 seconds. On Yelp Polarity, Logistic Regression improved to 91.6% accuracy, outperforming LSTM (86.2%) while remaining over 300 times faster. By integrating both predictive metrics and efficiency measures, the Efficiency Score highlighted the practical advantages of Logistic Regression and Naïve Bayes in resource-constrained environments. This dual evaluation framework demonstrates that classical models remain highly competitive when both accuracy and efficiency are considered, providing a practical alternative to computationally expensive neural architectures and offering practitioners clear guidelines for model selection under real-world constraints.**

Keywords*—Sentiment analysis; natural language processing; machine learning; deep learning logistic regression; random forest; Naive Bayes; LSTM; CNN; Efficiency Score

I. Introduction

Sentiment analysis, also known as opinion mining, has emerged as a central task in natural language processing (NLP) due to the exponential growth of digital platforms such as social networks, e-commerce, and online review systems. It is widely applied across domains such as marketing, reputation management, healthcare, and education, where it supports decision-making by extracting and classifying opinions or emotions from text [1]. However, the inherently noisy, ambiguous, and context-dependent nature of textual data continues to pose major challenges for classification algorithms [2].

Traditional machine learning approaches such as Logistic Regression, Naïve Bayes, and Random Forest have long been popular for sentiment classification. Their effectiveness stems from simple vector space representations like Bag-of-Words and TF-IDF, which enable efficient training and competitive accuracy. Yet, their inability to capture contextual dependencies limits their performance on longer or more complex texts [3]. In contrast, deep learning models—including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks—have achieved significant progress by modeling local and sequential relationships in text [4]. Despite these advances, neural models are computationally expensive, requiring substantial resources and long training times, which restrict their deployment in real-time or resource-constrained settings.

More recently, Transformer-based architectures such as BERT [6], DistilBERT [7], and ALBERT [8] have set new performance standards in sentiment analysis by providing superior contextual representations. However, their efficiency challenges—particularly in terms of memory requirements, inference latency, and energy consumption—remain well documented [9]. As a result, the practical question of balancing predictive performance and computational cost has become increasingly relevant.

In this study, we address this gap by conducting a systematic comparison of classical machine learning algorithms and neural architectures for sentiment analysis, evaluated through a multi-metric framework that integrates both predictive performance (Accuracy, Precision, Recall, and F1-score) and efficiency measures (Training Time and Efficiency Score). While the first set of metrics provides a rigorous view of classification quality, the second emphasizes feasibility for real-world applications where resources are limited and rapid response is critical.

The contributions of this work are threefold:

1) We provide a controlled empirical benchmark of classical algorithms (Logistic Regression, Naïve Bayes, Random Forest) and neural architectures (CNN, LSTM) on two benchmark datasets (IMDB and Yelp Polarity).

2) We introduce an Efficiency Score, defined as the ratio of accuracy to training time, as a composite indicator for jointly assessing predictive power and computational cost.

3) We highlight the robustness and practicality of classical models, demonstrating that they remain highly competitive when evaluated across multiple metrics, making them suitable for real-time and resource-constrained applications.

By establishing efficiency as a key criterion alongside accuracy, precision, recall, and F1-score, this work contributes a comprehensive evaluation framework that moves beyond traditional accuracy-centered approaches, providing both researchers and practitioners with a practical methodology for model selection in sentiment analysis.

To situate this study within the broader context of existing approaches, the following section reviews related work on classical machine learning, deep learning, and Transformer-based models for sentiment analysis, with a focus on their strengths, limitations, and reported evaluation practices.

The remainder of this paper is organized as follows. Section II provides a review of related work, covering classical, deep, and Transformer-based sentiment classification approaches. Section III presents the datasets, preprocessing steps, selected models, and evaluation framework, including the proposed Efficiency Score. Section IV reports and analyzes the experimental results obtained from applying the models to benchmark datasets. Finally, Section V concludes the study and discusses potential directions for future research.

II. Related Works

Machine learning models have long been widely applied in sentiment analysis. Logistic Regression has consistently demonstrated competitive performance across diverse domains, typically achieving accuracies above 80% [7]. Random Forest has also shown robustness in handling noisy and heterogeneous data [10], [12], [13] [14],[15], while Naïve Bayes remains valued for its lightweight training and stability across contexts [16], [17]. However, the main limitation of these models is their inability to capture contextual relationships between words, which reduces their effectiveness on longer or more complex text. Moreover, prior studies evaluating these models have often relied solely on accuracy, neglecting complementary predictive metrics such as precision, recall, and F1-score.

Deep learning architectures have significantly advanced sentiment analysis by incorporating sequential and contextual modeling. Convolutional Neural Networks (CNNs) [3], [22] were adapted to capture local dependencies, while Long Short-Term Memory (LSTM) networks have consistently achieved strong performance in modeling long-range dependencies [4],[5], [19], [20]. For example, some studies have demonstrated the effectiveness of LSTM models when initialized with word embeddings, such as Word2Vec [5] or GloVe [21]. Studies such as Kaya and Fidan (2020) reported accuracies above 90% on Turkish and IMDB reviews using LSTM. However, these improvements come at the cost of significantly higher computational demands, which limit their deployment in real-time or resource-constrained settings. Importantly, most evaluations of deep learning models emphasize accuracy and, at best, F1-score, without systematically assessing training time or computational feasibility [28] [29].

More recently, Transformer-based models such as BERT [6], DistilBERT [8], and ALBERT [26] have achieved state-of-the-art accuracy in multiple sentiment benchmarks. DistilBERT, for instance, reduces computation while retaining approximately 95% of BERT's performance, whereas ALBERT improves memory efficiency through parameter sharing. Nevertheless, even lightweight Transformers remain resource-intensive compared to classical algorithms, raising concerns about their scalability for large-scale or low-latency applications [9], [25].

Infrastructure-level approaches for distributed NLP processing have also been considered [27], offering insights into how metadata management and storage layers can impact large-scale model deployment. While some studies mention inference cost, few provide a systematic comparison that integrates efficiency metrics alongside predictive performance.

Table I summarizes representative studies in sentiment analysis, highlighting the predominant reliance on accuracy as the principal evaluation criterion. Although a few works have reported precision, recall, or F1-score, computational efficiency is rarely benchmarked rigorously, leaving a gap between theoretical performance and practical applicability.

TABLE I SUMMARY OF REPRESENTATIVE SENTIMENT ANALYSIS STUDIES

Author(s) & Year	Models tested	Dataset	Evaluation criteria	Limitations
Kim (2014) [23]	CNN	Movie reviews	Accuracy	Ignores computational cost
Tang et al. (2015) [24]	GRNN, LSTM	IMDB	Accuracy	No efficiency evaluation
Kaya & Fidan (2020)[20]	LSTM	Turkish reviews, IMDB	Accuracy, F1	Lacks comparison with lightweight models
Abdirahman et al. (2023) [30]	ML (SVM, NB), DL (LSTM)	Somali dataset	Accuracy	Efficiency not considered
Jahan et al. (2024) [31]	ML (SVM, RF, LR, NB)	Twitter, Facebook posts	Accuracy, Precision, Recall	No efficiency analysis
Varone et al. (2023) [32]	Word embeddings + DL	Arabic reviews	Accuracy, F1	Lacks efficiency study
Sanh et al. (2019) [8]	DistilBERT	GLUE, SST-2	Accuracy	Efficiency mentioned but not benchmarked
Lan et al. (2020) [26]	ALBERT	GLUE, sentiment datasets	Accuracy	Memory efficiency only partially discussed

As summarized in Table I, prior research has predominantly focused on predictive accuracy [33], with limited consideration of complementary performance metrics and little to no integration of computational efficiency [18]. While deep learning and Transformer-based models achieve state-of-the-art accuracy, their high training and inference costs limit their suitability for large-scale or real-time deployment. Conversely, classical algorithms provide faster training and lower resource consumption but are frequently dismissed as mere baselines without sufficient attention to their efficiency advantage.

This gap underscores the need for a comprehensive evaluation framework that jointly considers accuracy, precision, recall, F1-score, training time, and efficiency. In this work, we directly address this issue by benchmarking classical models (Logistic Regression, Naïve Bayes, Random Forest) against neural architectures (CNN, LSTM) across two benchmark datasets (IMDB and Yelp Polarity). Unlike most comparative studies, we explicitly integrate efficiency metrics into the evaluation, providing a more holistic perspective and practical guidelines for model selection under real-world computational constraints.

III. Methodology

To ensure a fair and reproducible comparison between classical and neural approaches to sentiment analysis, this study followed a structured experimental pipeline covering dataset selection, preprocessing, feature extraction, model implementation, training, and evaluation.

A. Datasets

The two widely used benchmark datasets were employed to ensure robustness and cross-domain validation:

- IMDB Movie Reviews: This dataset contains 50,000 annotated reviews, evenly distributed between positive and negative sentiments, with 25,000 reviews used for training and 25,000 for testing Maas et al. [11]. Its balanced structure makes it a standard benchmark for sentiment analysis.
- Yelp Polarity Reviews: To test generalization across domains, a subset of 25,000 reviews (20,000 for training and 5,000 for testing) was extracted. The dataset consists of highly polarized reviews, making it a complementary benchmark to IMDB.

Both datasets were chosen due to their public availability, balanced class distribution, and wide adoption in related work.

B. Preprocessing

Textual Preprocessing was designed to standardize inputs and reduce noise:

1) Text normalization: Lowercasing and removal of punctuation, numbers, and special characters.

2) Tokenization: Splitting sentences into tokens.

3) Stop-word removal (for classical models only).

4) Sequence handling: For neural models, all sequences were truncated or padded to a fixed length of 200 tokens.

5) Feature representation:

- For classical models: TF-IDF vectors were extracted with a vocabulary limited to the 10,000 most frequent terms to reduce sparsity.
- For neural models: Tokens were mapped to 50-dimensional GloVe embeddings, allowing semantic relationships between words to be captured.

C. Feature Extraction and Models

Two families of models were evaluated:

- Machine learning algorithms:
  - Logistic Regression (with L2 regularization).
  - Naïve Bayes (multinomial version).
  - Random Forest (with balanced class weights).
- Deep learning algorithms:
  - Convolutional Neural Network (CNN) adapted for text classification using 1D convolutions, batch normalization, and dropout.
  - Long Short-Term Memory (LSTM) with dropout layers and L2 penalties to mitigate overfitting.

These models were selected based on three primary criteria:

- Methodological diversity: The set includes both classical models that rely on sparse, vectorized inputs (TF-IDF, Bag-of-Words) and deep neural architectures capable of modeling local and sequential dependencies in text. This ensures that the evaluation covers different learning paradigms.
- Empirical popularity: All five models are frequently used in both academic and applied sentiment analysis tasks. They are well-established benchmarks and form the backbone of many open-source sentiment analysis pipelines and toolkits.
- Practical relevance: These models exhibit a wide spectrum of computational complexity and scalability. While deep models offer high accuracy at greater cost, classical models like LR and NB are known for their efficiency and interpretability—critical in real-time or resource-constrained applications.

Hyperparameters were chosen based on prior literature and preliminary experiments to ensure fair comparisons.

D. Training and Optimization

Traditional models were trained using 5-fold cross-validation to reduce sampling bias. Neural models were optimized with the Adam optimizer, using a batch size of 64. Early stopping was applied when validation accuracy did not improve for two consecutive epochs, and the learning rate was dynamically adjusted using the ReduceLROnPlateau strategy. Dropout rates between 0.5 and 0.6 and L2 penalties were employed to prevent overfitting.

All experiments were conducted on Google Colab Pro, which provides a cloud-based environment with access to both CPU and GPU resources. Specifically, training was performed using an NVIDIA Tesla T4 GPU (16 GB VRAM) with 12 GB of RAM. Hardware specifications are reported to contextualize training times and to ensure reproducibility of results.

E. Evaluation Metrics

The effectiveness of the models was assessed using four widely adopted classification metrics:

- Accuracy: Overall correctness of predictions.
- Precision: Ratio of correctly predicted positive instances to all predicted positives.

- Recall: Ratio of correctly predicted positive instances to all actual positives.
- F1-score: Harmonic mean of precision and recall.

To complement predictive performance, training time (in seconds) was systematically recorded as a measure of computational efficiency. Unlike prior studies, this work introduces a dual evaluation perspective, where accuracy and efficiency are analyzed together.

Additionally, we propose a simple composite indicator, the Efficiency Score, defined as:

$$\text{Efficiency Score} = \frac{\text{Accuracy Training Time}}{\text{Training Time (s)}}$$

This metric highlights the trade-off between predictive power and computational cost, allowing for a more practical comparison of models.

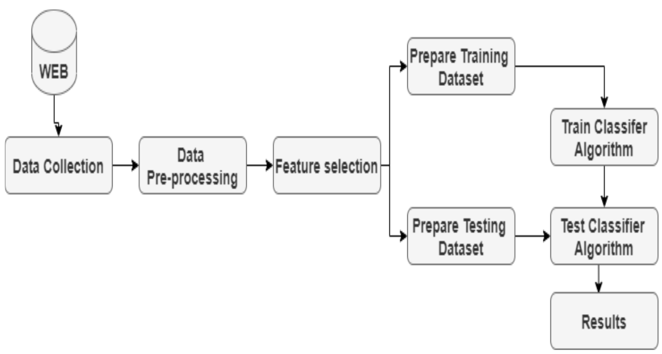

Fig. 1. Experimental pipeline for sentiment analysis.

The experimental workflow followed in this study is illustrated in Fig. 1. The pipeline provides a high-level overview of the process, starting with dataset selection, followed by data preprocessing and feature extraction, which transform raw text into machine-readable representations. The data is then divided into training and testing subsets, enabling the evaluation of multiple classifiers. The final stage involves measuring both predictive performance and computational efficiency. This structured pipeline ensures that results are reproducible and that trade-offs between accuracy and efficiency can be systematically analyzed.

IV. Results and Comparative Analysis

The experimental findings are presented in this section, combining results on the IMDB and Yelp Polarity datasets. In addition to predictive performance, training time was measured to highlight the trade-off between accuracy and efficiency, which is often overlooked in prior work.

A. IMDB Results

The comparative performance of classical and neural models on the IMDB dataset is presented in Table II. Logistic Regression (LR) and LSTM both achieved the highest accuracy (88%), but with vastly different computational requirements: LR completed training in only 0.25 seconds, whereas LSTM required over 2600 seconds. Random Forest and Naïve Bayes achieved slightly lower accuracy (85% and 84%, respectively) while remaining more computationally efficient. CNN obtained 82% accuracy but required more than 750 seconds of training.

TABLE II. Performance of Classical and Neural Models on IMDB Dataset

Model	Accuracy	Precision	Recall	F1-score	Training Time (s)	Efficiency Score
Logistic Regression	0.88	0.89	0.87	0.88	0.25	3.52
Random Forest	0.85	0.86	0.85	0.85	10.32	0.0824
Naïve Bayes	0.84	0.83	0.86	0.84	0.12	7.0
CNN	0.82	0.83	0.80	0.81	751.91	0.0011
LSTM	0.88	0.89	0.85	0.87	2603.62	0.0003

These results clearly demonstrate the importance of considering computational efficiency alongside predictive performance. Although neural networks are often considered superior, the findings reveal that Logistic Regression matches LSTM in accuracy while being orders of magnitude faster. Moreover, Naïve Bayes achieves the highest Efficiency Score, making it a compelling option in time-critical environments.

To better visualize these trade-offs, Fig. 2 combines predictive performance metrics (accuracy, precision, recall, and F1-score) with training time. The figure highlights the stark efficiency advantage of classical models compared to neural architectures, especially in large-scale or real-time applications.

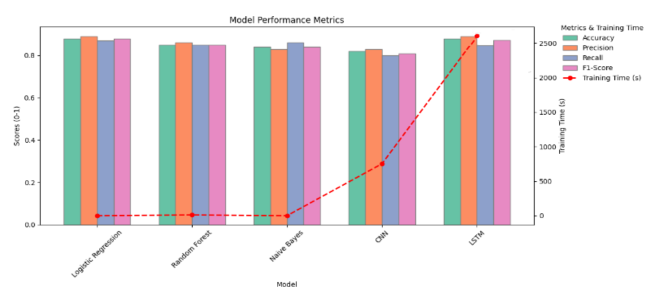

Fig. 2. Predictive performance and training time of classical and neural models on IMDB dataset.

B. Results on Yelp Polarity

To assess cross-domain generalization, Logistic Regression (LR) and LSTM were further evaluated on the Yelp Polarity dataset. As shown in Table III, LR significantly outperformed LSTM in both accuracy and efficiency. LR achieved 91.6% accuracy and an F1-score of 0.916 in less than one second, whereas LSTM obtained 86.2% accuracy and an F1-score of 0.839, requiring approximately 179 seconds of training.

TABLE III PERFORMANCE COMPARISON OF LOGISTIC REGRESSION AND LSTM ON YELP POLARITY DATASET

Model	Accuracy	Precision	Recall	F1-score	Training Time (s)	Efficiency Score
Logistic Regression	0.916	0.916	0.916	0.916	0.95	0.9642
LSTM	0.862	0.84	0.838	0.839	179.0	0.0048

These findings highlight that LR is not only more efficient but also more robust across domains. While LSTM matched LR on IMDB, its performance dropped on Yelp, suggesting a sensitivity to dataset characteristics. Yelp reviews are often shorter and more polarized, making them well-suited to sparse TF-IDF representations used by LR. By contrast, LSTM's reliance on sequential dependencies may lead to overfitting on domain-specific structures, limiting transferability.

Fig. 3 provides a visual overview, showing both predictive metrics and training time. It illustrates how LR maintains strong predictive accuracy with minimal computational cost, while LSTM incurs substantially higher training overhead with inferior predictive outcomes.

C. Efficiency Score Analysis

To provide a more integrated view of the trade-off between predictive performance and computational efficiency, an Efficiency Score was introduced, defined as the ratio of accuracy to training time. This composite metric highlights models that achieve strong predictive results at minimal computational cost.

On the IMDB dataset (Table II), Naïve Bayes recorded the highest Efficiency Score (7.0), owing to its extremely fast training time combined with reasonable accuracy (84%). Logistic Regression followed closely with an Efficiency Score of 3.52, striking a strong balance between accuracy and efficiency. By contrast, CNN and LSTM achieved near-zero scores (0.0011 and 0.0003), indicating that their heavy computational requirements outweighed their accuracy levels.

On the Yelp Polarity dataset (Table III), Logistic Regression again dominated with an Efficiency Score of 0.9642, significantly outperforming LSTM (0.0048). These results confirm that Logistic Regression is not only efficient but also robust across datasets, maintaining superior predictive performance while requiring only a fraction of the training cost.

Fig. 4 further illustrates the comparison by plotting Accuracy against Training Time on a logarithmic scale. The plot shows that classical models cluster in the region of high accuracy and low cost, while neural models shift toward high cost with limited accuracy benefits. This Pareto-like distribution emphasizes that Logistic Regression and Naïve Bayes represent the most practical choices under real-world computational constraints.

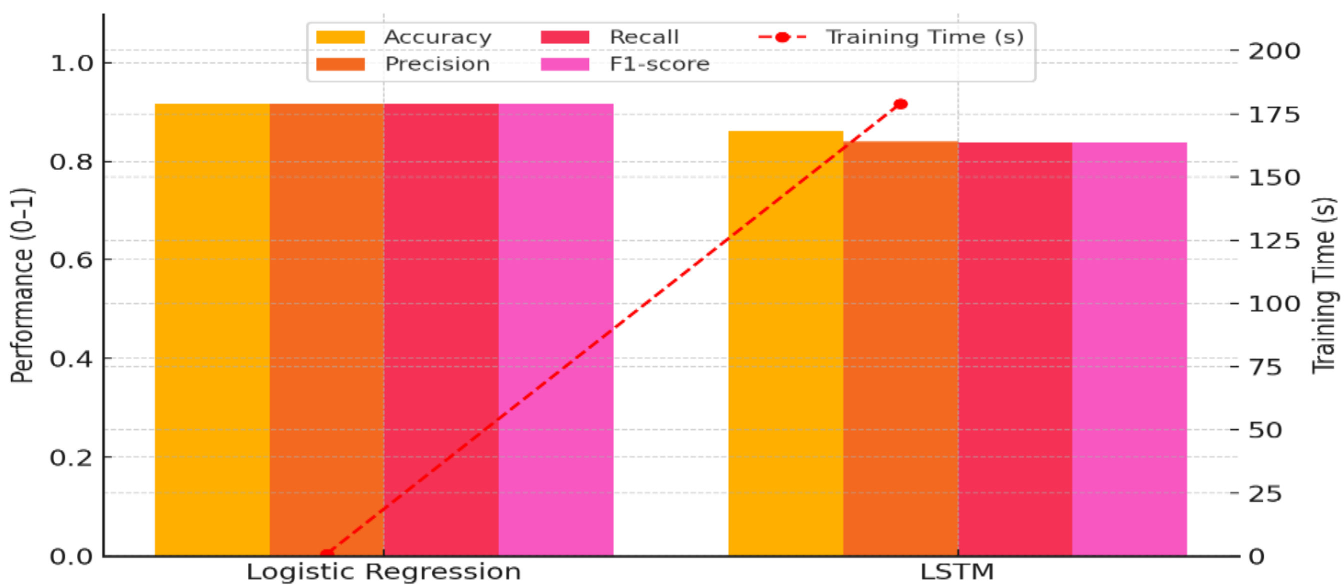

Fig. 3. Predictive performance and training time of logistic regression and LSTM on Yelp polarity dataset.

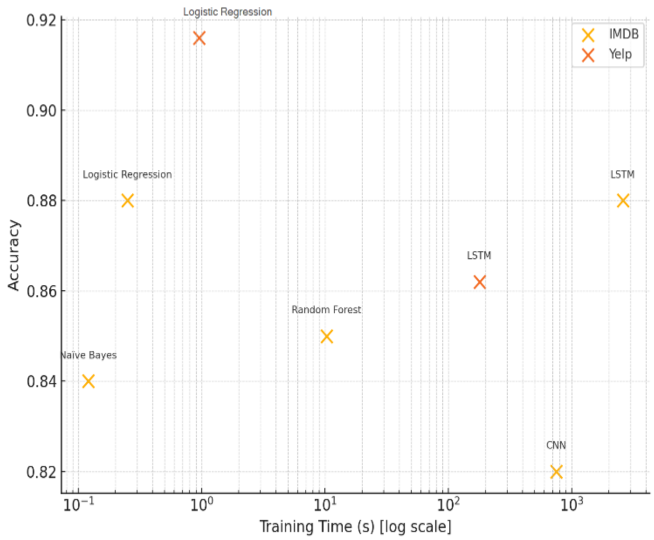

Fig. 4. Accuracy vs. training time (log scale) for IMDB and Yelp datasets.

D. Discussion

The comparative results on IMDB and Yelp demonstrate that accuracy alone is insufficient to evaluate sentiment analysis models. While LSTM and CNN are theoretically capable of capturing sequential dependencies and richer contextual patterns, their computational demands make them impractical for real-time or large-scale applications. In both datasets, Logistic Regression consistently emerged as the most efficient and, on Yelp, even more accurate than LSTM.

Several insights can be drawn from these findings:

1) Generalization across domains: Logistic Regression generalized more effectively than LSTM on Yelp. One possible explanation is that TF-IDF features provide robust, domain-independent representations of word importance, whereas LSTM relies heavily on sequential dependencies. This reliance may have caused LSTM to overfit to dataset-specific structures in IMDB, reducing its transferability to Yelp's shorter, more polarized reviews.

2) Impact of feature engineering: Classical models benefited greatly from TF-IDF vectorization, which captures sentiment-bearing terms with high precision. In contrast, neural models depended on pre-trained GloVe embeddings, which may not fully capture the nuances of Yelp or IMDB reviews without extensive fine-tuning. This suggests that effective feature engineering can offset, and sometimes surpass, the advantages of deep architectures, particularly when resources are limited.

3) Sensitivity to noise and class imbalance: The robustness of Logistic Regression also stems from its relative insensitivity to noisy or imbalanced data. Naïve Bayes, though fast, is more vulnerable to noisy tokens and strong independence assumptions, while Random Forest struggles with high-dimensional sparse features. Neural networks can mitigate some noise through embeddings, but require large, balanced datasets to perform optimally. This makes them less reliable in real-world scenarios where data often contains noise and imbalance.

From a practical standpoint, these results indicate that simpler models such as Logistic Regression remain highly competitive in modern sentiment analysis pipelines. They provide an optimal balance of accuracy and efficiency, making them suitable for applications such as real-time social media monitoring, customer feedback analysis, and deployment on mobile or embedded devices. Deep neural networks, while powerful, should be reserved for contexts where computational resources are abundant and domain-specific sequential modeling is critical.

E. Limitations

While the results of this study highlight the competitiveness of classical models in sentiment analysis, several limitations must be acknowledged:

1) Restricted dataset scope: Only two benchmark datasets (IMDB and Yelp Polarity) were used. Although these are widely adopted, they may not fully represent real-world scenarios such as noisy, informal, or multilingual data (e.g. Twitter or cross-lingual reviews).

2) Limited model coverage: The comparison focused on classical machine learning algorithms (Logistic Regression, Naïve Bayes, Random Forest) and selected deep learning architectures (CNN, LSTM). More recent Transformer-based architectures, such as BERT, DistilBERT, or ALBERT, were not included, even though they represent the state of the art in sentiment analysis.

3) Evaluation metrics: Computational efficiency was evaluated primarily in terms of training time. Other important factors, such as inference latency, memory footprint, and energy consumption were not measured. These aspects are increasingly relevant for real-time and sustainable AI applications.

4) Pre-trained embeddings: Neural models were initialized with GloVe embeddings, which may not capture domain-specific nuances without further fine-tuning. Alternative embeddings (e.g. contextual embeddings from Transformers) could yield different outcomes.

These limitations do not undermine the validity of the findings but indicate that the study represents a controlled benchmark rather than an exhaustive evaluation. Addressing them in future work would broaden the applicability of the results and further strengthen the conclusions.

V. Conclusion and Future Work

A. Conclusion

This study presented a comparative evaluation of classical machine learning algorithms and deep learning architectures for sentiment analysis, with a dual emphasis on predictive performance and computational efficiency. On the IMDB dataset, Logistic Regression (LR) and LSTM achieved similar accuracy (88%), but with drastically different computational costs: LR required only 0.25 seconds, whereas LSTM exceeded 2600 seconds. To assess generalizability, additional experiments were conducted on the Yelp Polarity dataset, where Logistic Regression not only maintained but improved its performance (91.6% accuracy, F1 = 0.916), clearly outperforming LSTM while remaining more than 300 times faster.

These findings underscore a fundamental trade-off in sentiment analysis: classical models, particularly Logistic Regression, remain highly competitive because they combine strong accuracy with exceptional efficiency, making them suitable for real-time and resource-constrained applications. In contrast, deep learning models such as LSTM provide richer contextual modeling but at prohibitive computational costs, limiting their practicality in large-scale or rapid-deployment settings. CNNs, while capable of capturing local dependencies, proved less effective overall, reinforcing the view that convolution alone is insufficient for modeling long textual sequences.

Beyond reaffirming the relevance of classical approaches, this study contributes a dual evaluation framework that integrates efficiency alongside accuracy, addressing a gap in the literature where predictive metrics alone have typically dominated. By introducing the Efficiency Score as a composite indicator, the work provides practitioners with a practical guideline for selecting sentiment analysis models that balance predictive power against computational feasibility.

B. Future Work

Future research will extend this analysis along several directions. First, the evaluation will be broadened by incorporating a wider range of datasets beyond IMDB and Yelp, including noisy and short-text corpora such as Twitter streams, as well as multilingual collections (e.g. Arabic and French reviews). This will allow for a more rigorous assessment of model robustness across diverse and challenging real-world contexts.

Second, Transformer-based architectures such as BERT, RoBERTa, and lightweight variants like DistilBERT and ALBERT will be included in future benchmarks. Comparing these models with both classical and recurrent approaches will provide valuable insights into whether their superior contextual modeling justifies the significant computational overhead, especially in scenarios where efficiency is a critical requirement.

Third, future studies will adopt a broader view of computational efficiency. While this study measured efficiency primarily in terms of training time, other dimensions such as inference latency, memory consumption, and energy usage are increasingly important in practice. Incorporating these factors will contribute to a more holistic evaluation framework and align sentiment analysis research with the growing field of sustainable AI.

Finally, lightweight optimization techniques such as model pruning, quantization, and knowledge distillation will be explored to reduce the resource requirements of deep learning architectures without substantially sacrificing accuracy. These methods could enable the development of hybrid frameworks that combine the interpretability and efficiency of classical models with the representational power of neural networks, thus bridging the gap between theoretical performance and practical deployment.

REFERENCES

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

[2] Cambria, E., Schuller, B., Xia, Y., Havasi, C.: New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems 28(2), 15–21 (2017).

[3] Ganie, S.H., Dadvandipour, Z.: Exploring the impact of informal language on sentiment analysis models for social media text using convolutional neural networks. International Journal of Advanced Computer Science 15(3), 105–115 (2023).

[4] Zhang, T.: LSTM network-based sentiment analysis for textual data. Journal of Information Technology Research 28(4), 305–320 (2023).

[5] Feng, Z., Zhao, J., Chen, L.: Sentiment analysis of user comments using LSTM and Word2Vec embeddings. Data Science Journal 12(2), 200–215 (2023).

[6] Abia, V.M., Johnson, T.K.: A comparative study of sentiment analysis techniques for Nigerian opinions using random forest and logistic regression. Applied Intelligence 42(1), 112–130 (2024).

[7] Bahtiar, N., Rahman, A., Putri, S.R.: Comparison of logistic regression and naive Bayes for sentiment analysis on Indonesian marketplaces. Procedia Computer Science 210, 1520–1528 (2023).

[8] Negi, A., Pandey, S., Ranjan, K.: Aspect-based sentiment analysis using CNN-LSTM. Computational Linguistics and Natural Language Processing 8(3), 180–195 (2022).

[9] A. Erraissi et A. Belangour, « A big data security layer meta-model proposition », Advances in Science, Technology and Engineering Systems, vol. 4, no 5, p. 409-418, 2019, doi: 10.25046/aj040553.

[10] Rudra, R.I., Gopalakrishnan, A.K.: Sentiment analysis of consumer reviews using machine learning approach. IEEE Journal, pp. 1–7 (2023).

[11] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proc. ACL-HLT, pp. 142–150 (2011).

[12] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. NAACL-HLT, pp. 4171–4186 (2019).

[13] Fauzi, N., Rahman, M., Aditya, P.: Analysis of tweets related to climate change using Random Forest. In: Proc. Int. Conf. on Social Media Analysis, pp. 70–76. ACM, New York (2023).

[14] Syamsi, R., Hidayat, A., Kurniawan, A.: COVID-19 sentiment analysis on tweets using Random Forest. Journal of Computational Data Science 19(3), 80–91 (2023).

[15] Zheng, R.: Sentiment analysis on IMDb reviews with Random Forest. Computational Intelligence Journal 22, 86–95 (2023).

[16] Utomo, B.P., Kurniasih, T., Wijayanto, F.: Sentiment analysis of Smart Campus Unisbank application reviews using Naive Bayes. Indonesian Journal of Data Science 7(3), 55–63 (2023).

[17] Putri, N.A., Darmawan, R., Syahputra, H.: Sentiment analysis of tweets from three Indonesian cities using Naive Bayes. In: Proc. Int. Conf. on Data and Social Media Research, pp. 88–95 (2023).

[18] Walji, K., Erraissi, A., Zakrani, A., Banane, M. (2025). Comparative Performance Evaluation of Machine Learning Algorithms in Sentiment Analysis. Lecture Notes in Networks and Systems, vol 1397. Springer, Cham. https://doi.org/10.1007/978-3-031-90921-4_17.

[19] A. Erraissi et A. Belangour, « Meta-modeling of zookeeper and mapreduce processing », présenté à 2018 International Conference on Electronics, Control, Optimization and Computer Science, ICECOCS 2018, 2018. doi: 10.1109/ICECOCS.2018.8610630.

[20] Bilen, O., Horasan, Z.: Sentiment analysis on Turkish and IMDb datasets using LSTM. Journal of Computational Linguistics 32(1), 99–110 (2022).

[21] Kande, R.K.: Sentiment analysis on tweets using GloVe embeddings and LSTM. In: Proc. IEEE Conf. on Sentiment Analysis, pp. 101–112 (2024).

[22] Maharani, S., Yulianti, T., Rahmawati, L.: Sentiment analysis of fuel price hikes using CNN. International Journal of Social Media Analytics 13(3), 77–84 (2023).

[23] Kim, Y.: Convolutional neural networks for sentence classification. In: Proc. EMNLP, pp. 1746–1751 (2014).

[24] Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural networks for sentiment classification. In: Proc. EMNLP, pp. 1422–1432 (2015).

[25] Elbayed, Z., & Qadi El Idrissi, A. (2025). Deep Learning in Financial Modeling: Predicting European Put Option Prices with Neural Networks. Algorithms, 18(3), 161. https://doi.org/10.3390/a18030161.

[26] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A lite BERT for self-supervised learning of language representations. In: Proc. ICLR (2020).

[27] A. Erraissi et A. Belangour, « Capturing Hadoop storage big data layer meta-concepts », présenté à Advances in Intelligent Systems and Computing, 2019, p. 413-421. doi: 10.1007/978-3-030-11928-7_37.

[28] A. Erraissi et A. Belangour, « Meta-modeling of big data visualization layer using on-line analytical processing (OLAP) », International Journal of Advanced Trends in Computer Science and Engineering, vol. 8, no 4, p. 990-998, 2019, doi: 10.30534/ijatcse/2019/02842019.

[29] H. Ouchra, A. Belangour, et A. Erraissi, « An overview of GeoSpatial Artificial Intelligence technologies for city planning and development », présenté à 2023 5th International Conference on Electrical, Computer and Communication Technologies, ICECCT 2023, 2023. doi: 10.1109/ICECCT56650.2023.10179796.

[30] Abdirahman, A., et al. "Transfer learning for Somali sentiment analysis." Journal of African Languages and Linguistics, vol. 44, no. 3, 2023. DOI: 10.1515/jall-2023-0034.

[31] I. Jahan and T. Farah Sanam, "An Improved Machine Learning Based Customer Churn Prediction for Insight and Recommendation in E-commerce," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 1-6, doi: 10.1109/ICCIT57492.2022.10054771.

[32] Elhassan, N., Varone, G., Ahmed, R., Gogate, M., Dashtipour, K., Almoamari, H., El-Affendi, M. A., Al-Tamimi, B. N., Albalwy, F., & Hussain, A. (2023). Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning. Computers, 12(6), 126.

[33] Mouataz IDRISSI KHALDI, Allae ERRAISSI, Mustapha HAIN and Mouad BANANE. "In-Depth Comparison of Supervised Classification Models - Performance and Adaptability to Practical Requirements". International Journal of Advanced Computer Science and Applications (ijacsa) 16.8 (2025). http://dx.doi.org/10.14569/IJACSA.2025.0160862.