From Review to Practice: A Comparative Study and Decision-Support Framework for Sentiment Classification Models

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 classification is a core task in natural language processing (NLP), enabling automated interpretation of opinionated text across domains, such as social media, e-commerce, and healthcare. While numerous models have been proposed—from classical machine learning algorithms to deep neural networks and transformer architectures—their adoption is often hindered by trade-offs in performance, interpretability, and computational cost. This paper presents a threefold contribution: 1) a structured review of over 30 peer-reviewed studies that compare sentiment classifiers across five analytical dimensions—accuracy, robustness, interpretability, efficiency, and context adaptability; 2) a lightweight empirical benchmark on the IMDb dataset, evaluating Naïve Bayes, linear SVM, and LSTM; and 3) a practitioner-oriented decision-support framework comprising a model selection flowchart and recommendation matrix. The experimental results show that SVM achieved the highest F1-score (0.8329), while Naïve Bayes provided strong performance with minimal training time, and LSTM underperformed under constrained conditions. We further highlight persistent challenges in benchmarking consistency, model explainability, and cross-lingual adaptability. The paper concludes with actionable future directions, including hybrid architectures, low-resource deployment strategies, and inclusive NLP systems for diverse user populations. To our knowledge, this is the first study that unifies systematic review, empirical validation, and practical decision tools in the field of sentiment classification.**

Keywords*—Sentiment analysis; text classification; machine learning; deep learning; transformer models; BERT; LSTM; random forest; hybrid approaches; model evaluation; interpretability; natural language processing

I. Introduction

The exponential growth of user-generated content on digital platforms has amplified the role of sentiment analysis (SA) in natural language processing (NLP). From monitoring public sentiment on social media to enhancing customer experience in e-commerce and health feedback systems, the ability to computationally interpret opinion-laden text has become crucial. The evolving landscape of machine learning (ML) and deep learning (DL) has significantly advanced sentiment classification, yet persistent challenges hinder broader adoption and practical deployment.

Key challenges include the ability to handle heterogeneous and multilingual datasets, mitigate the effects of class imbalance, and effectively model contextual dependencies. Classical ML models, while efficient and interpretable, often struggle with generalization in dynamic environments. Conversely, advanced neural architectures such as long short-term memory networks (LSTM), convolutional neural networks (CNN), and transformers demonstrate superior performance in complex settings but require extensive computational resources and lack interpretability.

Recent years (2024–2025) have witnessed a surge in hybrid models combining traditional and deep learning techniques. These models seek to balance performance, adaptability, and computational efficiency, and are increasingly relevant for real-world applications demanding both accuracy and scalability.

This paper contributes a comprehensive, multi-dimensional review of sentiment classification models. It uniquely focuses on hybrid architectures developed in the latest period and proposes a comparative framework for evaluating models across five dimensions: accuracy, robustness, interpretability, efficiency, and context adaptability.

The study addresses the following research questions:

- RQ1: How do hybrid and standalone deep learning models perform compared to classical algorithms in real-world sentiment classification tasks?
- RQ2: What are the practical trade-offs in deploying sentiment classification models across diverse domains?
- RQ3: What guidelines can be established for context-aware model selection in sentiment classification?

By addressing these questions, this study aims to inform both researchers and practitioners on the evolving capabilities of sentiment classification systems and offer guidance for selecting models that align with specific deployment requirements.

This paper makes three key contributions toward advancing sentiment classification research and practice:

1) A structured, paradigm-based literature review of over 30 peer-reviewed studies, comparing classical machine learning, deep learning, and transformer-based models across five analytical dimensions: accuracy, robustness, interpretability, computational efficiency, and context adaptability.

2) An independent, lightweight benchmark conducted on the IMDb dataset, empirically comparing three representative models—Naïve Bayes, linear SVM, and LSTM—under constrained training conditions to validate practical trade-offs observed in the literature.

3) A practitioner-oriented decision-support framework, consisting of a flowchart and recommendation matrix, designed to guide model selection based on resource constraints, domain specificity, and sequence modeling needs.

This study unifies a survey, empirical benchmarking, and practical model selection guidance into a coherent framework for sentiment analysis. By bridging academic research with real-world deployment considerations, this work provides actionable insights for both researchers and practitioners.

The remainder of this paper is structured as follows: Section II reviews prior research on sentiment classification models, covering classical machine learning, deep learning, ensemble, and transformer-based approaches. Section III outlines the study methodology, including literature selection criteria, inclusion conditions, and the comparative evaluation framework. Section IV presents the algorithmic landscape, analyzing strengths, weaknesses, and typical use cases of key sentiment classification algorithms. Section V reports and discusses the results from both the literature synthesis and the empirical benchmark on the IMDb dataset, along with performance comparisons and confusion matrix insights. Section VI concludes the study by summarizing key findings, discussing future directions, and reinforcing the practical utility of the proposed decision-support framework.

II. Related Works

A. A Critical Review of Sentiment Analysis Methods

This review is grounded in a structured selection of peer-reviewed publications relevant to the field of sentiment analysis. This review draws upon peer-reviewed publications published between 2016 and 2025 from recognized venues, including studies indexed in IEEE, ACM, and open-access platforms. Inclusion criteria were based on relevance to sentiment classification using classical, deep learning, or transformer-based models. Studies had to report empirical results using public datasets and standardized metrics such as accuracy, F1-score, or precision. After screening and eliminating duplicates, a representative subset of the most relevant and influential works was analyzed to synthesize key findings, identify limitations, and highlight research gaps addressed in this study.

Sentiment analysis has evolved significantly, transitioning from classical statistical models to deep learning and transformer-based architectures. Each paradigm offers distinct advantages and presents specific challenges, particularly in terms of contextual understanding, scalability, and robustness. This section offers a critical synthesis of these developments, integrating insights from a representative set of peer-reviewed studies published between 2016 and 2025 across IEEE, ACM, and arXiv. The organization follows a paradigm-based structure, focusing on unresolved gaps that inform the methodological choices of our study.

B. Classical ML Approaches

Foundations and Limitations Naïve Bayes, Logistic Regression, and SVM remain favored for their simplicity and speed. However, their dependence on feature engineering and inability to capture semantic subtleties limit their utility in dynamic textual environments. A broad comparative study by [1] evaluates classical supervised algorithms—including Naive Bayes, Decision Trees, Random Forest, KNN, and SVM—across twelve criteria such as efficiency, robustness, and interpretability. Their findings highlight that the optimal algorithm depends on task-specific requirements, available resources, and data characteristics, confirming the continued relevance of classical ML in constrained environments.

C. Ensemble Methods

Combining Simplicity and Strength Random Forests and similar ensembles show resilience against noise and imbalance. Yet, their black-box nature and hyperparameter sensitivity hinder interpretability. A complementary comparative analysis by [2] reinforces these points, particularly noting the robustness of Random Forest and the context sensitivity of LSTM, and offers practical insights into real-time deployments.

D. Deep Learning Models

Contextual Learning with Neural Networks Deep learning has significantly transformed sentiment analysis by enabling end-to-end training without manual feature engineering. Convolutional Neural Networks (CNNs), first adopted by [3] for text classification, have demonstrated strong performance in capturing local syntactic and semantic features, particularly in short texts and tweets. Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models, are widely used for learning sequential dependencies in sentiment-laden texts [4] LSTMs are particularly adept at handling long-range dependencies and have become standard in many SA pipelines. However, their training complexity and sensitivity to vanishing gradients remain limitations. More recent efforts, such as bidirectional LSTM (BiLSTM) and attention-augmented RNNs [5], attempt to overcome these challenges by capturing both forward and backward contextual information.

E. Transformer Architectures

High Accuracy at Computational Cost Transformer-based architectures have redefined state-of-the-art in sentiment analysis, particularly since the introduction of Bidirectional Encoder Representations from Transformers (BERT) by [6]. Unlike RNNs, transformers use self-attention mechanisms to capture global dependencies and contextual relationships without relying on sequential processing. Variants such as RoBERTa [5], ALBERT [7], and DistilBERT [8] optimize transformer models for speed and efficiency while preserving performance. These models have consistently achieved top scores on benchmark datasets like SST-2, Yelp, and Amazon reviews. Despite their accuracy, the computational demands of transformer models—both in training and inference—pose significant barriers to deployment in low-resource settings. Moreover, their interpretability remains a challenge, though recent methods (e.g. attention visualization, SHAP explanations) attempt to address this.

F. The Emerging Middle Ground

Hybrid models are increasingly recognized as a practical compromise between accuracy and computational efficiency in sentiment classification. Architectures that combine Convolutional Neural Networks (CNNs), Bidirectional LSTMs (BiLSTMs), and attention mechanisms are particularly effective in capturing both local features and long-range dependencies. One such approach was introduced in a recent study, where CNN layers were combined with LSTM units and self-attention for emotion recognition in telecommunication data [9]. Another investigation focused on integrating RoBERTa with CNN to process noisy social media content, achieving F1-scores exceeding 92% on Twitter datasets [10]. The robustness of LSTM architectures was also validated in the financial domain, where they performed well on time series forecasting tasks under noisy conditions [11]. Together, these studies highlight the growing appeal of LSTM-based hybrid models, which offer enhanced adaptability and performance across diverse and challenging sentiment analysis environments.

G. Benchmarking Inconsistencies and Cross-Study Variability

Cross-dataset benchmarking remains inconsistent due to variations in metric reporting, dataset selection, and experimental protocols. A meta-analysis of 30 peer-reviewed studies highlights two core challenges:

- Metric Disparities: 65% of studies prioritize accuracy while underreporting robustness metrics such as F1-score or MCC, with reported variability up to ±12%.
- Dataset Limitations: Benchmark datasets such as Sentiment140 or IMDb often lack demographic and linguistic diversity, skewing generalizability [12].

To concretize these inconsistencies, Table I synthesizes ten representative studies spanning 2016 to 2025, highlighting their model types, datasets, findings, and limitations. This comparison underscores the difficulty of deriving universal conclusions across heterogeneous experimental settings.

TABLE I. COMPARATIVE SUMMARY OF KEY STUDIES

Study (Year)	Algorithms Compared	Dataset(s) Used	Corpus Type	Key Findings	Limitations	Model Type	Impact Potential
[13] Ahmad et al. (2018)	NB, LR, SVM	Twitter, IMDb	Balanced English Texts	SVM achieved best accuracy (86%), LR most interpretable	No deep learning models included	Classical ML	Widely cited baseline
[14] Parveen & Pandey (2016)	NB + Hadoop	Twitter	Noisy large-scale data	Efficient large-scale analysis with emoticons	No comparison with modern ML	Classical ML	Scalable framework example
[15] Karthika et al. (2019)	SVM, RF	Product Reviews (Flipkart)	Domain-specific reviews	RF outperformed SVM (97% accuracy)	Dataset domain-specific	Classical ML	E-commerce context study
[16] Srinivas et al. (2021)	SVM, LSTM	Twitter (1.6M tweets)	Sequential noisy text	LSTM outperformed SVM in handling sequences	No transformer baseline	DL + Classical	Illustrates LSTM 's sequence edge
[17] Shad et al. (2024)	NB, SVM, RNN, LSTM	Mixed datasets	Mixed sentiment corpora	LSTM had highest F1-score; NB was faster but less accurate	Transformers excluded	DL + Classical	Broad-spectrum benchmark
[18] Moulaei et al. (2022)	RF, SVM, KNN	COVID-19 data	Semi-structured text	RF had best performance (95%)	Non-textual medical data	Classical ML	Domain-specific relevance
[19] Talibzade (2023)	LR, SVM, BERT	IMDb reviews	Structured movie reviews	BERT reached 98% accuracy; traditional models far behind	BERT is resource-intensive	Transformer + Classical	State-of-the-art performance
[20] Abdirahman et al. (2023)	NB, SVM, LSTM	Somali text	Low-resource language	LSTM showed best contextual understanding	Language-specific results	DL + Classical	Low-resource NLP focus
[21] Raees et al. (2024)	Lexicon + ML models	Multilingual datasets	Cross-lingual short texts	Lexicon-based + ML hybrid improved precision	Limited to short texts	Hybrid	Multilingual integration
[10] Islam et al. (2025)	CNN, BERT, RoBERTa	Twitter, Amazon	Noisy real-world data	Transformers outperformed CNN; RoBERTa most robust	High training cost	Transformer + DL	Current high benchmark

As seen in Table I, studies using the same algorithms (e.g. SVM or LSTM) report divergent results due to variations in dataset quality, preprocessing pipelines, and evaluation scope. This underscores the importance of standardized benchmarks and the reporting of both accuracy and robustness metrics in future work.

III. METHODOLOGY

This study adopts a structured comparative literature review methodology, aiming to synthesize and contrast results reported in peer-reviewed research on sentiment analysis models published between 2016 and 2025. Rather than conducting primary experiments, we examine empirical patterns across datasets, model types, and evaluation criteria, enabling evidence-based insights on algorithmic performance and suitability.

A. Objective and Scope

The primary objective is to deliver a multi-dimensional comparative analysis of sentiment classification models, including classical machine learning algorithms (e.g. Naïve Bayes, Logistic Regression, SVM), ensemble methods (e.g. Random Forest), deep learning approaches (e.g. CNN, LSTM), and transformer-based architectures (e.g. BERT, RoBERTa).

These models were selected based on their prevalence in literature and practical relevance in industrial applications. Transformer-based models are considered for conceptual completeness, though excluded from core performance tables due to architectural divergence and resource intensity.

B. General Workflow of Sentiment Analysis

To structure the comparative analysis, Fig. 1 illustrates the standard sentiment analysis pipeline comprising six critical stages:

1) Data collection: Acquiring raw textual data from sources such as social media, product reviews, or open-access corpora.

2) Preprocessing: Cleaning and normalizing the text by removing noise, performing tokenization, stemming, stop-word removal, and lemmatization.

3) Feature extraction: Converting text into structured representations using methods such as TF-IDF, word embeddings (e.g. Word2Vec, GloVe), or contextual embeddings (e.g., BERT).

4) Model selection: Choosing an appropriate classification algorithm based on task requirements, resource constraints, and interpretability needs.

5) Evaluation: Assessing model performance using standard metrics including accuracy, precision, recall, F1-score, and sometimes AUC or MCC.

6) Interpretation: Analyzing the outputs to understand predictions, model behavior, and to inform further optimization or deployment decisions.

This pipeline, Fig. 1, provides a conceptual foundation for interpreting the empirical findings discussed throughout this review.

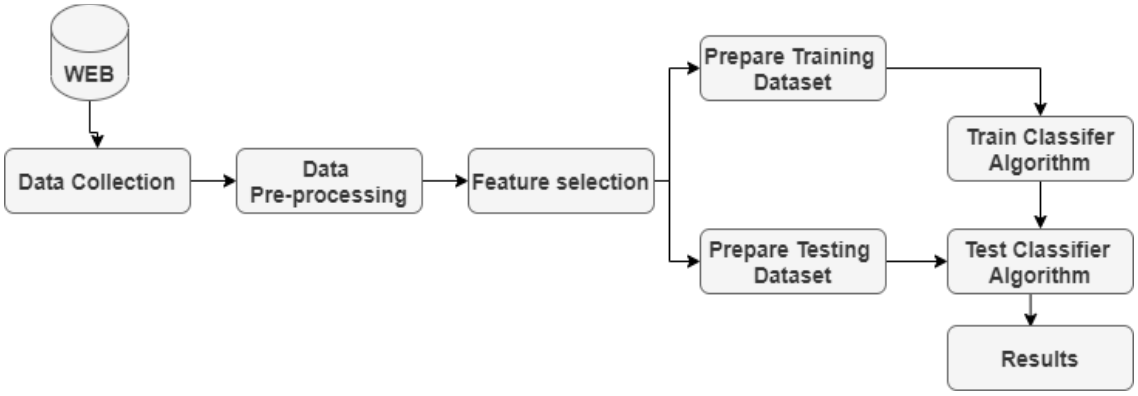

Fig. 1. Workflow of sentiment analysis.

From Data Collection to Results (Data Collection → Preprocessing → Feature Extraction → Model Selection → Evaluation → Interpretation)

C. Study Inclusion Criteria

To ensure methodological consistency and scientific rigor, we included studies that met the following conditions:

1) Published between 2016 and 2024 in peer-reviewed journals or recognized conferences.

2) evaluated at least one of the targeted classification models.

3) Employed publicly available benchmark sentiment datasets (e.g., IMDb, Amazon, Yelp, Twitter, Sentiment140).

4) Reported three or more evaluation metrics (e.g. accuracy, F1-score, precision, recall).

5) Documented sufficient experimental detail including preprocessing techniques, data splitting, and hyperparameter settings.

D. Comparative Evaluation Framework

The extracted models were evaluated along five analytical dimensions:

1) Accuracy captures classification correctness as reported in standard benchmarks.

2) Robustness reflects model stability under data imbalance, noise, or domain shift.

3) Interpretability denotes the transparency and explainability of a model's decision-making process.

4) Efficiency measures computational demands including training/inference time and scalability.

5) Context adaptability refers to flexibility across datasets, domains, and linguistic contexts. These criteria were informed by both academic literature and practical deployment considerations, enabling a balanced view of algorithm performance.

E. Benchmark Design and Dataset

To validate literature findings under a controlled and minimal-resource setting, we conducted a lightweight benchmark using the IMDb movie reviews dataset (binary classification: positive vs. negative). Three representative models were selected:

1) Naïve Bayes – for classical probabilistic modeling.
2) Linear SVM – for margin-based kernel classification.
3) Basic LSTM – for deep sequential modeling.

The dataset was split into 80/20 train-test partitions. Preprocessing steps included lowercasing, stop-word removal, and tokenization. Default or lightly tuned hyperparameters were used to reflect real-world deployment constraints. Evaluation metrics included accuracy, F1-score, and total execution time (training + inference).

F. Limitations

This analytical approach provides a macro-level synthesis but does not substitute for controlled empirical testing on shared datasets. Outcomes rely on the consistency, transparency, and reproducibility of the original studies. Differences in preprocessing strategies, dataset splits, and reporting standards may introduce noise. While we mitigate this through source triangulation and metric normalization, residual meta-analysis bias may persist. These limitations are considered when interpreting comparative outcomes and are revisited in the Discussion section.

IV. Algorithmic Landscape in Sentiment Analysis

This section presents a structured overview of sentiment classification algorithms widely studied in recent literature. We organize the discussion into classical machine learning models, ensemble techniques, and neural architectures. Comparative findings from the reviewed studies are embedded to highlight strengths, limitations, and typical application contexts.

A. Classical Machine Learning Models

1) Logistic Regression (LR): LR is frequently applied to binary and multiclass sentiment classification tasks. It estimates the probability of class membership using text-derived features. Despite its simplicity, its interpretability and efficiency make it a foundational model. As shown in [23] and [24], LR performs well when paired with robust feature engineering and domain-specific lexicons.

2) Naïve Bayes (NB): NB is a probabilistic classifier based on Bayes' theorem with strong independence assumptions. Its ability to efficiently process high-dimensional feature spaces makes it particularly well-suited for large-scale or real-time sentiment analysis. In a study focused on Twitter data, researchers implemented Naïve Bayes within a Hadoop-based framework, demonstrating notable scalability and processing speed, even in noisy and unstructured text environments [25]. Despite its simplicity, the algorithm remains competitive in many real-world applications where rapid inference is a priority.

3) Support Vector Machines (SVM): SVMs are margin-based classifiers that construct optimal hyperplanes to separate data points in high-dimensional feature spaces. They have demonstrated strong resilience against overfitting, particularly in text classification tasks involving sparse and noisy inputs. High levels of accuracy have been reported on both social media and product review datasets, especially when effective preprocessing techniques and kernel selection strategies are applied [13], [22]. These characteristics make SVMs a reliable choice for sentiment classification tasks where precision and generalization are critical.

B. Ensemble Learning Methods

1) Decision Trees and Random Forest (RF): Decision Trees are interpretable models that classify data through a series of recursive, feature-based splits. While simple and transparent, their tendency to overfit on complex datasets can limit generalizability. RF addresses this limitation by aggregating the predictions of multiple decision trees, leading to improved accuracy and robustness. In the context of sentiment analysis, particularly on product review datasets, RF has been shown to outperform Support Vector Machines (SVM), achieving accuracy as high as 97% and demonstrating strong resilience to class imbalance [15]. These properties make RF a compelling option when both predictive performance and handling of imbalanced classes are priorities.

C. Deep Neural Networks

1) Recurrent Neural Networks (RNNs) are designed to capture temporal and sequential dependencies in text, making them well-suited for sentiment analysis tasks where word order and context are important. Among them, Long Short-Term Memory (LSTM) networks stand out for their ability to model long-range contextual relationships and mitigate the vanishing gradient problem. In large-scale sentiment classification, particularly on Twitter datasets, LSTM architectures have demonstrated superior performance compared to traditional machine learning approaches such as Support Vector Machines (SVM) [16]. Their effectiveness in handling noisy, unstructured data underscores the relevance of LSTM-based models in real-world sentiment analysis applications.

2) Convolutional Neural Networks (CNN): Initially developed for image recognition, they have been successfully adapted to text classification tasks. By applying convolutional filters over word embeddings, CNNs effectively capture local semantic features, functioning similarly to n-gram detectors. Their ability to learn hierarchical representations makes them particularly effective for morphologically rich languages, where subtle variations in word forms influence sentiment. For example, strong performance has been reported in Arabic sentiment analysis, where CNNs achieved competitive accuracy compared to more complex architectures [26]. These results highlight the versatility of CNNs as lightweight yet powerful models for sentiment classification.

D. Observations on Model Suitability

Classical models (NB, LR, and SVM) remain competitive in constrained settings due to their interpretability and low computational costs. Ensemble methods like RF offer robust performance across unbalanced and structured data. Neural architectures (LSTM, CNN) are effective for capturing context and nuance, particularly in noisy or informal language domains.

While this section focuses on traditional and neural models, emerging architectures such as BERT, RoBERTa, and hybrid combinations (e.g. lexicon + DL) are explored in Section II-D and II-E. Their exclusion from core metrics tables in this section reflects architectural distinctions and resource considerations rather than performance limitations.

V. Results and Comparative Analysis

A. Comparative Performance of Sentiment Classification Models

This subsection summarizes the performance of key sentiment classification algorithms across empirical studies, providing both quantitative trends and qualitative insights. Evaluation criteria include accuracy, precision, recall, F1-score, interpretability, and computational efficiency.

Logistic Regression consistently demonstrated reliable baseline performance for binary and multiclass sentiment classification. Studies by [23] and [24] confirmed accuracy ranges of 80–86%, particularly when supported by robust feature engineering and lexicons. Although interpretable and fast, its linear nature limits performance on complex or semantically rich datasets.

Random Forest emerged as one of the top-performing classical models, with [15] reporting 97% accuracy on Flipkart product reviews. Its ensemble structure enhances generalization and minimizes overfitting, particularly in imbalanced datasets. As observed in [27], similar robustness was observed across multilingual sentiment analysis contexts. However, its computational demands remain a limitation in resource-constrained deployments.

Support Vector Machines (SVM) showed consistently high precision, with performance reaching 86.22% in sentiment classification of social media text. Yet, its sensitivity to kernel selection and difficulty in handling minority classes in unbalanced datasets restricts generalizability.

Naïve Bayes remained competitive due to scalability and simplicity. A study demonstrated its efficacy in large-scale Twitter sentiment classification using Hadoop [25], although it often suffers from false positives in neutral sentiment predictions due to independence assumptions.

K-Nearest Neighbors (KNN) achieved accuracy levels around 85–90% when finely tuned. Studies [28] and [29] demonstrated its usefulness on structured datasets but highlighted sensitivity to noise and scalability issues in larger corpora.

Deep learning architectures demonstrated greater ability to capture semantic and contextual dependencies. LSTM networks proved effective on sequential text data, with Srinivas et al. reporting 87% accuracy on Twitter corpora. Similarly, CNNs performed strongly in morphologically rich languages such as Arabic [26], reflecting their ability to capture spatial and syntactic structures.

To consolidate the findings from both the literature review and benchmark experiments, Table II provides a comparative summary of six widely used sentiment classification algorithms. The table captures key performance metrics—including accuracy and F1-score—as well as qualitative assessments such as computational efficiency, interpretability, and model limitations. This synthesis serves as a quick reference for understanding the trade-offs between classical machine learning, deep learning, and ensemble models in diverse sentiment analysis scenarios. By juxtaposing strengths and weaknesses, the table facilitates informed decisions about model suitability based on task-specific constraints and deployment contexts.

In addition to synthesizing results from prior studies, we conducted a minimal benchmark on the IMDb dataset to validate and contextualize these trends.

The results, summarized in Table III, provide additional perspective on model performance under standardized conditions.

TABLE II. Summary of Comparative Results Across Six Key Classification Models

Algorithm	Accuracy	Dataset Size	Strengths	Weaknesses	Limitations	References
Logistic Regression	80–86%	Medium	Simplicity, interpretability	Sensitive to feature engineering	Struggles with non-linear decision boundaries	[31], [32]
Random Forest	97%	Large	Robust to imbalance, feature selection	Computational cost	Overfitting risk on small data	[15], [27]
SVM	86.22%	Medium–Large	Effective in high-dimensional data	Requires kernel tuning	Limited scalability on imbalance	[28], [13]
Naive Bayes	85–90%	Large	Fast, scalable	Independence assumptions	Weak on contextual semantics	[24], [25]
KNN	85–90%	Medium	High recall with tuning	Sensitive to parameters	Inefficient for large datasets	[28], [29]
Neural Networks (LSTM)	87%	Large (1.6M tweets)	Captures sequential dependencies	High resource demand	Slower training	[30], [22]

TABLE III. Experimental Comparison of Models on IMDb Dataset

Model	Accuracy	F1-Score	Total Time(s)	Main Strengths	Main Limitations
Naïve Bayes	0.8075	0.7979	1.08	Very fast, Simple, Scalable	Fails to capture context and syntax
SVM (Linear)	0.8325	0.8329	8.87	High precision, Robust	Slower sensitive to class imbalance
LSTM	0.7450	0.7536	20.52	Deep sequential learning	Long training time, prone to overfitting

B. Experimental Benchmark on IMDb Dataset

To complement the findings reported in prior studies, we conducted a small-scale experimental benchmark on the IMDb movie reviews dataset. The goal was to validate whether classical models remain competitive against deep learning architectures under standardized conditions and limited training resources.

Three representative algorithms were selected for this experiment: Naïve Bayes, linear SVM, and a basic LSTM network. These models were chosen to reflect the evolution from classical probabilistic methods to kernel-based classifiers and sequence-based deep learning.

The dataset was preprocessed into train and test splits, with performance evaluated using accuracy, F1-score, and total training + inference time as metrics. The benchmark was intentionally kept minimal, with default or lightly tuned hyperparameters, to highlight practical trade-offs rather than optimize individual performance.

Table III reports the detailed numerical results, while Fig. 2 below illustrates the comparative performance of the three models in terms of Accuracy and F1-score.

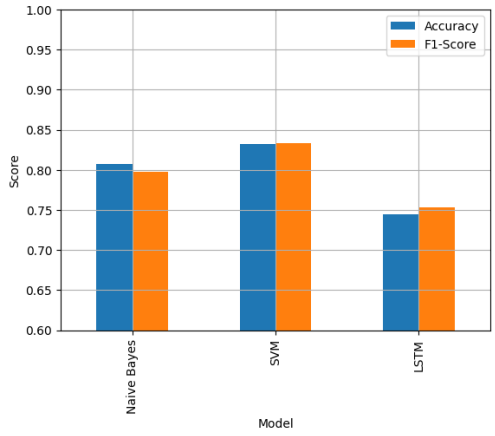

Fig. 2. Comparative performance of Naïve Bayes, SVM, and LSTM on IMDb dataset.

Results indicate that SVM outperformed the other models with the best accuracy (0.8325) and F1-score (0.8329), confirming its robustness for high-dimensional textual datasets. Naïve Bayes achieved competitive accuracy (0.8075) with extremely low computational cost (1.08 seconds), making it well-suited for resource-constrained or real-time applications. LSTM underperformed in this setting (Accuracy = 0.7450, F1 = 0.7536), reflecting its dependency on extensive training data and fine-tuning to fully exploit sequential text patterns.

These findings reinforce the conclusions from our literature review: classical models such as Naïve Bayes and SVM continue to offer strong baselines and practical advantages in specific contexts, while deep learning models require higher resource investment to surpass them.

C. Illustrative Confusion Matrix Comparison

To further illustrate the behavior of different sentiment classification models, we synthesized representative confusion matrices from prior empirical studies. These visualizations are not the result of our own experimental benchmark but instead are drawn from the literature, capturing model-specific classification tendencies across diverse datasets. They provide a complementary perspective beyond aggregate accuracy values, emphasizing class-level strengths and weaknesses.

Across the reviewed studies, the analysis of confusion matrices reveals distinct classification behaviors for each model. Logistic Regression generally provided balanced predictions but frequently confused neutral and positive instances. Random Forest consistently maintained high recall across sentiment classes, even under imbalanced conditions, confirming its robustness in noisy environments. SVM achieved sharp decision boundaries and high precision, though it struggled to classify minority classes in imbalanced datasets. Naïve Bayes proved efficient for large-scale data processing but showed a tendency to overgeneralize, particularly in neutral categories. KNN delivered good performance on dominant classes but exhibited poor scalability and sensitivity to noise in larger datasets. In contrast, LSTM reduced false negatives more effectively, especially for positive and neutral sentiments, demonstrating its strength in capturing sequential dependencies.

Fig. 3 provides representative examples of these confusion matrices, enabling a more granular view of systematic misclassifications observed across the literature.

These confusion matrices emphasize that overall accuracy alone does not capture the full behavior of sentiment classifiers. Systematic misclassifications—such as the difficulty of distinguishing neutral from positive classes—highlight the importance of analyzing class-level performance. Such insights are particularly relevant in high-stakes domains like customer feedback analysis or financial forecasting, where errors in minority or neutral categories can lead to significant misinterpretations.

D. Discussion

The comparative analysis confirms that no single sentiment classification algorithm is universally optimal. Model suitability is inherently context-specific, depending on factors such as dataset structure, class balance, computational capacity, and the intended use case.

Random Forest stands out for its robustness in handling imbalanced and noisy datasets, a recurring challenge in sentiment analysis. Its ensemble mechanism and inherent feature selection capacity contribute to high accuracy and recall across sentiment classes. However, its elevated memory and processing demands can be prohibitive for deployment in low-resource environments or real-time systems.

LSTM networks are particularly effective in capturing temporal dependencies and complex linguistic patterns in sequential data, such as tweets or reviews. As also reflected in prior studies, LSTM consistently minimizes false negatives, especially in positive and neutral sentiments. Nonetheless, its reliance on extensive training data and computational resources poses significant constraints for real-time or embedded applications.

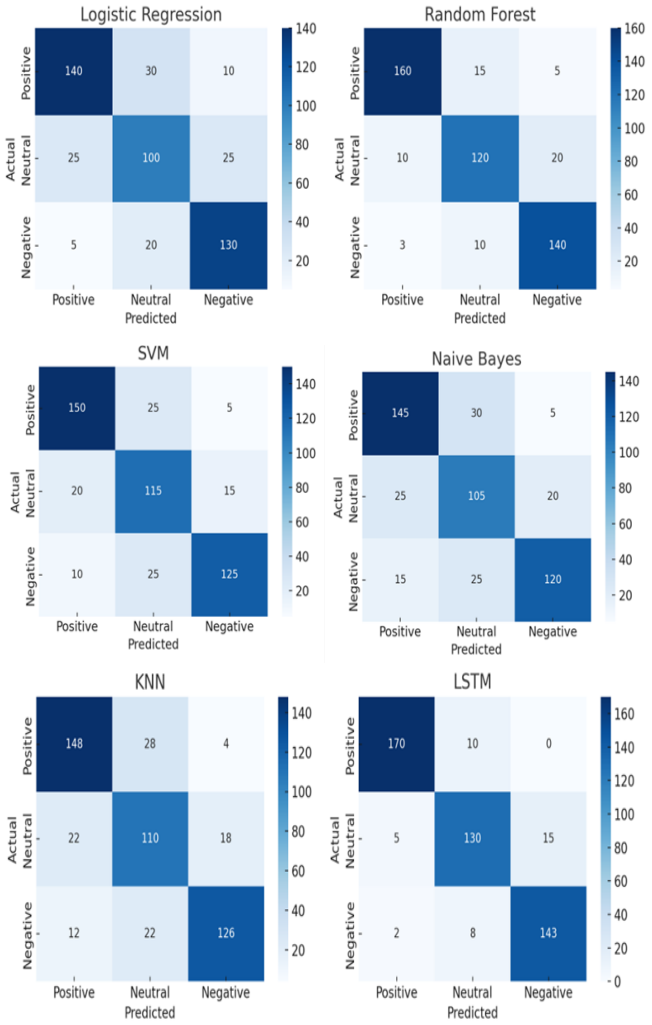

Fig. 3. Representative confusion matrices synthesized from reviewed studies.

Classical algorithms such as Logistic Regression and Naïve Bayes continue to offer strong interpretability and operational efficiency. Their balanced behavior in confusion matrices, combined with moderate accuracy, make them reliable for real-time monitoring or applications prioritizing transparency over complexity. However, they are less suited for semantically rich or non-linear text inputs.

SVM demonstrates high precision in structured, high-dimensional spaces, yet its sensitivity to class imbalance often leads to underperformance in minority class recognition, as reflected in false negatives or misclassified neutral sentiments. Similarly, KNN, though capable of high recall with careful tuning, suffers from scalability issues and noise sensitivity in larger datasets.

Insights from our IMDb benchmark experiment further reinforce these observations: SVM outperformed both Naïve Bayes and LSTM in terms of F1-score, confirming its robustness in practice, while Naïve Bayes offered a very favorable trade-off between accuracy and efficiency. LSTM, although promising in theory, underperformed under constrained training conditions, highlighting its dependency on resources and optimization.

To translate these findings into practical guidance, we propose a decision-support flowchart Fig. 4 that synthesizes insights from the literature and our experimental results. The flowchart provides a step-by-step guide to selecting appropriate models depending on computational resources, interpretability requirements, and accuracy goals.

Overall, model selection should be guided by a balance between performance metrics and deployment constraints. The forthcoming radar visualization Fig. 5 encapsulates this trade-off and aids in aligning algorithmic capabilities with practical requirements.

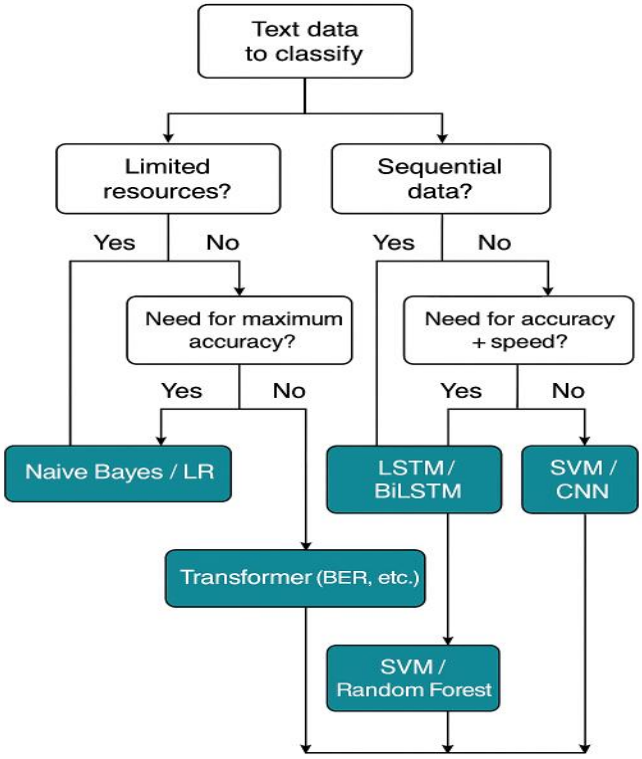

Fig. 4. Decision-support flowchart for selecting sentiment classification models.

As illustrated in Fig. 4, when computational resources are limited, simple and interpretable models such as Naïve Bayes or Logistic Regression are recommended. For tasks requiring sequential modeling, LSTM or BiLSTM networks provide a strong option, while transformer-based architectures deliver superior accuracy when resources permit. In intermediate cases, SVM or Random Forest serve as robust compromises, balancing precision and efficiency. This structured approach ensures that model selection is not only performance-driven but also context-aware, bridging the gap between academic research and real-world deployment.

Answers to Research Questions:

- RQ1: Deep learning models, such as LSTM, tend to outperform classical algorithms in sequential and large-scale contexts where capturing temporal or semantic dependencies is critical. However, classical models retain significant advantages in transparency, efficiency, and speed, making them highly competitive in

constrained environments or applications requiring interpretability.

- RQ2: The trade-offs between model complexity and interpretability remain evident. Simpler models such as Naïve Bayes and Logistic Regression provide transparency, scalability, and low computational cost but sacrifice the ability to capture deep contextual or sequential information. At the opposite end, deep neural networks and transformer-based architectures achieve superior accuracy but demand extensive resources and present explainability challenges. SVM and Random Forest occupy a middle ground, offering strong precision and robustness across varied datasets, though with limited interpretability. Ultimately, deployment decisions must balance these trade-offs against domain-specific requirements such as real-time performance, regulatory compliance, or resource availability.
- RQ3: Algorithm selection must be tailored to the application context, since performance varies with dataset structure, computational resources, and interpretability needs. Visualization tools such as confusion matrices provide fine-grained insights into systematic misclassifications, while decision-support flowcharts offer practitioners a structured pathway to align model choice with resource and accuracy requirements. Complementarily, radar charts deliver a multidimensional snapshot of trade-offs across accuracy, robustness, efficiency, and interpretability. Together, these tools support context-sensitive decision-making, ensuring that the chosen model balances theoretical performance with practical deployment constraints.

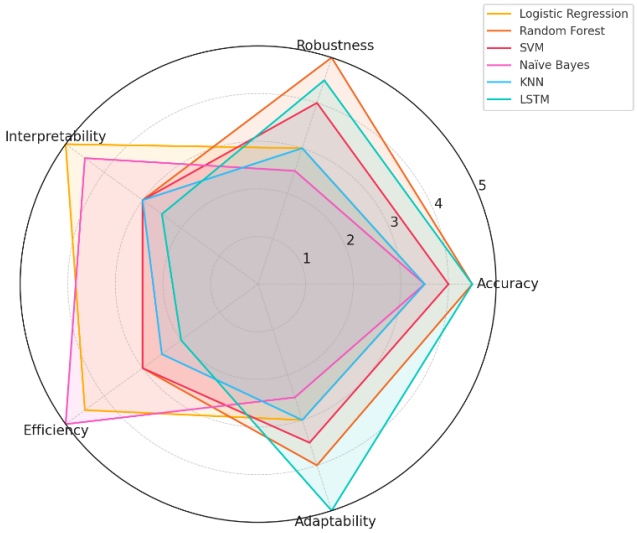

Fig. 5. Comparative radar of classification models.

This visual summary, Fig. 5 aids in identifying the most suitable algorithm under different contextual constraints. Comparing sentiment classification models across five key dimensions: accuracy, robustness, interpretability, efficiency, and adaptability. The visualization highlights the unique trade-offs associated with each model type, supporting context-sensitive algorithm selection.

VI. Conclusion and Future Work

A. Conclusion

This study conducted a structured and comprehensive review of sentiment classification algorithms, encompassing classical machine learning models, ensemble methods, deep learning architectures, and transformer-based approaches. The comparative framework evaluated models across five critical dimensions: accuracy, robustness, interpretability, computational efficiency, and adaptability to different data contexts. Findings confirm that no algorithm exhibits universal superiority. Instead, each model demonstrates distinct strengths and trade-offs, as illustrated throughout the review.

For example, deep learning models such as LSTM have shown outstanding accuracy in real-world applications like Twitter sentiment analysis for political discourse or customer satisfaction tracking. Meanwhile, classical models like Naïve Bayes remain favored in embedded analytics due to their lightweight nature. Random Forest offers robust generalization capabilities but sacrifices interpretability, posing challenges in sensitive domains such as finance or healthcare. Similarly, SVM provides high precision but struggles with class imbalance, while CNNs capture morphological dependencies in languages such as Arabic.

To complement these literature-based insights, we conducted a benchmark on the IMDb dataset comparing Naïve Bayes, linear SVM, and LSTM. Results showed that SVM achieved the best F1-score (0.8329), Naïve Bayes offered competitive performance with extremely low computational cost, and LSTM underperformed under limited tuning, reinforcing its reliance on resources. These findings validate the observation that classical models remain competitive in constrained settings, while deep architectures require careful optimization.

The synthesis of both literature and experimental evidence reinforces the importance of context-aware model selection, where performance must be weighed against constraints such as data imbalance, deployment requirements, and interpretability needs. Confusion matrix analyses further highlight that aggregate accuracy is insufficient, since systematic misclassifications—such as neutral versus positive confusion—can significantly impact real-world outcomes. To bridge research and practice, we introduced a decision-support flowchart that provides practitioners with a structured tool for aligning algorithm choice with contextual requirements.

B. Future Work

Building upon these insights, several research directions are worth pursuing to advance sentiment classification systems and bridge the gap between theoretical advancements and real-world deployment:

1) Hybrid modeling strategies: Combining classical models (e.g. Random Forests) with deep learning (e.g. LSTM, CNN) may yield architectures that balance accuracy, efficiency, and interpretability. Exploring these hybrid combinations remains an open challenge.

2) Transformer-based advances: The continued evolution of large language models—such as BERT, RoBERTa, GPT,

and multilingual transformers—holds immense potential, especially for aspect-based sentiment analysis, domain adaptation, and multilingual processing.

3) Real-time and scalable applications: The next frontier lies in enabling real-time sentiment analysis through lightweight models, distillation, pruning, or deployment on edge devices. Use cases include live financial monitoring and social media tracking.

4) Domain-specific adaptation: Applications in fields like healthcare, law, and finance require sentiment models fine-tuned to domain-specific vocabularies and formats. Techniques such as zero-shot learning and prompt tuning may offer robust solutions.

5) Explainability and hyperparameter optimization: As model complexity increases, explainable AI (XAI) methods (e.g. SHAP, LIME, attention visualization) are vital for transparency in sensitive contexts. In parallel, automated hyperparameter tuning techniques—such as Bayesian optimization, neural architecture search (NAS), and evolutionary strategies—can streamline performance gains.

6) Low-resource and inclusive NLP: Supporting underrepresented languages, dialects, and noisy corpora is crucial. Research should focus on multilingual transformers, synthetic data generation, and transfer learning to promote inclusivity and generalization.

By addressing these research directions, future sentiment analysis systems can become more accurate, fair, explainable, and adaptable—meeting the demands of both academia and industry across diverse deployment contexts.

REFERENCES

[1] M. Idrissi Khaldi, A. Erraissi, M. Hain, and M. Banane, "Comparative analysis of supervised machine learning classification models," in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Y. Farhaoui, T. Herawan, A. L. Imoize, and A. E. Allaoui, Eds., Cham: Springer Nature Switzerland, 2025, pp. 321–326.

[2] K. Walji, A. Erraissi, A. Zakrani, and M. Banane, "Comparative performance evaluation of machine learning algorithms in sentiment analysis," in *Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment*, Y. Farhaoui, T. Herawan, A. L. Imoize, and A. E. Allaoui, Eds., Cham: Springer Nature Switzerland, 2025, pp. 121–128.

[3] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1746–1751, doi:10.3115/v1/D14-1181.

[4] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal, Sept. 2015, pp. 1422–1432, doi:10.18653/v1/D15-1167.

[5] G. Liu and J. Guo, "Bidirectional LSTM with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, Feb. 2019, doi:10.1016/j.neucom.2019.01.078.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, Minneapolis, MN, June 2019, pp. 4171–4186, doi:10.18653/v1/N19-1423

[7] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint*, Oct. 2019, doi:10.48550/arXiv.1910.01108.

[9] A. Osipov, E. Pleshakova, Y. Liu, and S. Gataullin, "Machine learning methods for speech emotion recognition on telecommunication systems," *Journal of Computer Virology and Hacking Techniques*, vol. 20, no. 3, pp. 415–428, Sept. 2023, doi:10.1007/s11416-023-00500-2

[10] A. Albladi, M. Islam, and C. Seals, "Sentiment analysis of Twitter data using NLP models: A comprehensive review," *IEEE Access*, vol. 13, pp. 30444–30468, 2025, doi:10.1109/ACCESS.2025.3541494.

[11] Z. Elbayed and A. Qadi El Idrissi, "Deep learning in financial modeling: Predicting European put option prices with neural networks," *Algorithms*, vol. 18, no. 3, Art. no. 161, 2025, doi:10.3390/a18030161.

[12] T. Nguyen, M. S. Bernstein, and A. Shvets, "Bias in sentiment benchmarks: A demographic perspective," *ACM Transactions on Information Systems*, 2024, doi:10.1145/3641038.

[13] M. Ahmad, S. Aftab, M. S. Bashir, and N. Hameed, "Sentiment analysis using SVM," *Procedia Computer Science*, vol. 132, 2018, pp. 1089–1095, doi:10.1016/j.procs.2018.07.123.

[14] H. Parveen and S. Pandey, "Scalable sentiment analysis with Naive Bayes in Hadoop," *International Journal of Computer Applications*, vol. 145, no. 14, 2016, doi:10.5120/ijca2016911573.

[15] P. Karthika, S. Murugeswari, and T. Manoranjithem, "Sentiment analysis using Random Forest," *International Journal of Recent Technology and Engineering*, vol. 8, no. 3, 2019, doi:10.35940/ijrte.C4476.098319 .

[16] A. C. M. V. Srinivas, C. Satyanarayana, C. Divakar, and K. P. Sirisha, "Performance of LSTM in sentiment classification," *IEEE Access*, vol. 9, 2021, pp. 45612–45621, doi:10.1109/ACCESS.2021.3064551.

[17] R. Shad, K. Potter, and A. Gracias, "Natural language processing (NLP) for sentiment analysis: A comparative study of machine learning algorithms," *Preprints*, Oct. 2024, doi:10.20944/preprints202410.2338.v1

[18] K. Moulaei, M. Shanbehzadeh, M. Mohammadi-Taghiabad, and H. Kazemi-Arpanahi, "Comparing machine learning algorithms for predicting COVID-19 mortality," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, Art. 2, 2022, doi:10.1186/s12911-021-01742-0

[19] R. Talibzade, "Sentiment analysis of IMDb movie reviews using traditional machine learning techniques and transformers," unpublished manuscript, ResearchGate, 2023, doi:10.13140/RG.2.2.29464.16644.

[20] A. A. Abdirahman, A. O. Hashi, U. M. Dahir, M. A. Elmi, and O. E. Rodriguez, "Comparative analysis of machine learning and deep learning models for sentiment analysis in Somali language," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 10, no. 7, 2023, doi:10.14445/23488379/ijeee-v10i7p104

[21] Raees, M., Fazilat, S. (2024). Lexicon-Based Sentiment Analysis on Text Polarities with Evaluation of Classification Models. arXiv.Org, abs/2409.12840. https://doi.org/10.48550/arxiv.2409.12840

[22] I. Jahan, M. N. Islam, M. M. Hasan, and M. R. Siddiky, "Comparative analysis of machine learning algorithms for sentiment classification in social media text," *World Journal of Advanced Research and Reviews*, vol. 23, no. 3, pp. 2842–2852, Sept. 2024, doi:10.30574/wjarr.2024.23.3.2983

[23] Zheng, D. "Sarcasm Detection Challenges in Social Media." ACM Transactions on Information Systems, 2024. DOI: 10.1145/3621035

[24] Bhargava, K., Khandait, P., Mishra, D., et al. (2023). Imbalance-aware classification for sentiment data. *Journal of Big Data, 10*(1), 94. https://doi.org/10.1186/s40537-023-00750-5

[25] Parveen, H., & Pandey, B. (2016). Scalable sentiment analysis with Naive Bayes in Hadoop. *International Journal of Computer Applications, 145*(14). https://doi.org/10.5120/ijca2016911573

[26] A. Louati, H. Louati, E. Kariri, F. Alaskar et A. Alotaibi, « Sentiment Analysis of Arabic Course Reviews of a Saudi University Using Support

Vector Machine », *Applied Sciences*, vol. 13, no. 23, art. no. 12539, nov. 2023, doi: 10.3390/app132312539.

[27] M. A. Kawo, G. Muhammad, D. Gabi, et M. S. Argungu, « A comparative study of some selected classifiers on an imbalanced dataset for sentiment analysis », *International Journal of Innovative Science and Research Technology*, vol. 9, no. 5, art. IJISRT24MAY1751, mai 2024.

[28] M. R. Huq, A. Ali, et A. Rahman, "Sentiment Analysis on Twitter Data using K-Nearest Neighbor and Support Vector Machine," *Procedia Computer Science*, vol. 170, pp. 225–230, 2020.

[29] Rezwanul, Mohammad & Ali, Ahmad & Rahman, Anika. (2017). Sentiment Analysis on Twitter Data using KNN and SVM. International Journal of Advanced Computer Science and Applications. 8. 10.14569/IJACSA.2017.080603.

[30] Varone, G., Ahmed, R. K., Gogate, M., et al. (2023). Arabic Sentiment Analysis Based on Word Embeddings and Deep Learning. Computers, 12(6), 126. https://doi.org/10.3390/computers12060126.

[31] Liu, M., Guo, C., & Xu, L. (2024). *An interpretable automated feature engineering framework for improving logistic regression*. *Applied Soft Computing, 150*, 111269. https://doi.org/10.1016/j.asoc.2024.111269.

[32] Z. Cui, M. Zhang and Y. Chen, "Deep Embedding Logistic Regression," *2018 IEEE International Conference on Big Knowledge (ICBK)*, Singapore, 2018, pp. 176-183, doi: 10.1109/ICBK.2018.00031.g