Proceedings of the Informing Science + Information Technology Education Conference

An Official Publication of the Informing Science Institute
InformingScience.org

InformingScience.org/Publications

July 20 – 28, 2025

Comparative Evaluation of Traditional, Lexicon-Based, and Transformer Models for Sentiment Classification

Philip Obiorah*	School of Computing, University of Buckingham, Buckingham, United Kingdom	philip.obiorah@buckingham.ac.uk
Grace Diri	Ignatius Ajuru University of Education, Port Harcourt, Nigeria	grace.diri@iaue.edu.ng
Hongbo Du	School of Computing, University of Buckingham, Buckingham, United Kingdom	hongbo.du@buckingham.ac.uk

* Corresponding author

Abstract

Aim/Purpose This paper presents a comparative analysis of sentiment classification models, focusing on the performance differences between lexicon-based sentiment analysis, traditional machine learning techniques, and transformer-based deep learning approaches across multiple benchmark datasets.

Background While numerous methods exist for sentiment analysis, a systematic comparison across modeling paradigms and datasets remains limited. This study addresses this gap by evaluating representative models on IMDB, Yelp Polarity, and Amazon Polarity datasets using unified metrics and preprocessing pipelines.

Methodology The evaluation includes (a) VADER as a lexicon-based baseline, (b) Logistic Regression, Decision Tree and Naive Bayes models built on three types of features: Bag of Words, TF-IDF, and Word2Vec as the traditional machine learning alternatives, and (c) two transformer models, DistilBERT-SST-2 and RoBERTa-Sentiment, without fine-tuning as the deep learning methods. Performance is assessed using Accuracy, Precision, Recall, and F1 Score.

Accepted by Editor Michael Jones │ Received: March 1, 2025 │ Revised: April 10, 2025 │ Accepted: April 29, 2025.

Cite as: Obiorah, P., Diri, G., & Du, H. (2025). Comparative evaluation of traditional, lexicon-based, and transformer models for sentiment classification. In M. Jones (Ed.), *Proceedings of InSITE 2025: Informing Science and Information Technology Education Conference*, Article 23. Informing Science Institute. https://doi.org/10.28945/5490

Contribution This study offers a comprehensive, side-by-side comparison of conventional and modern sentiment classification techniques, identifying model-feature synergies and highlighting limitations in generalizability across the domains.

Findings Test results show that VADER, although efficient, underperforms on complex texts. Traditional models particularly Logistic Regression paired with TF-IDF yield strong performance.

Recommendations for Practitioners DistilBERT demonstrates superior performance across all datasets, while RoBERTa's performance suffers due to domain mismatch.

Recommendations for Researchers When computational resources are constrained, TF-IDF with Logistic Regression offers a competitive alternative to deep learning. For best performance, transformer models like DistilBERT should be prioritized.

Impact on Society The study can improve applications in customer feedback systems, social media monitoring, and public opinion analysis, contributing to more responsive and adaptive services.

Future Research Future work should explore domain-specific fine-tuning strategies and evaluate multilingual sentiment performance, especially for transformer models.

Keywords nature language processing, sentiment analysis, text classification, traditional machine learning, deep learning

INTRODUCTION

Sentiment analysis, also known as opinion mining, is a branch of study in natural language processing (NLP) that aims to locate and extract subjective information from unstructured text. Today, social media, customer reviews, and discussion forums are some of the digital platforms that offer content to be analyzed. Businesses, as well as researchers, are trying to develop effective strategies for gauging public sentiment. The ability to understand customer opinions, sentiments towards products, brand perception as well as the wellbeing of targeted demographic groups is important for various industries, finance institutions, political organisations, healthcare providers, and others.

The selection of an analytical method remains one of the most significant problems in sentiment analysis. Traditional approaches based on Machine Learning (ML) algorithms have gained popularity due to their ease of use and interpretability. Examples include Support Vector Machines (SVM), Naïve Bayes (NB), and Logistic Regression (LR). However, the traditional approaches often do not perform well with linguistic structures, contextual dependencies, and evolving language patterns. With nuanced sentiments and contextual details, Deep Learning (DL) models show higher performance. Examples of these are Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNN), and Bidirectional Encoder Representations from Transformers (BERT) (L. Zhang et al., 2018). While these models are accurate, they are computationally expensive and require large datasets to train on, which can be difficult for budget-constrained environments. DL models also face difficulties in model explainability and interpretability (Biecek & Samek, 2024).

This study analyzes the reviews of literature regarding the use of both traditional and deep learning approaches in machine learning sentiment analysis in order to determine which is more effective. It looks at various studies in order to provide a comprehensive understanding of the strengths, limitations, and practical applications of different sentiment analysis techniques. In addition to the literature synthesis, the study offers a critical review of key existing work, identifying methodological trends and gaps in comparative evaluations. Building on this foundation, we implement a broad empirical assessment of selected sentiment analysis models across three diverse benchmark datasets —

IMDB, Yelp Polarity, and Amazon Polarity. The evaluation spans traditional machine learning methods, lexicon-based tools, and deep learning approaches, all tested under a unified framework using consistent preprocessing and performance metrics. This integrated approach enables a fair comparison of techniques and provides insights into their relative effectiveness across different data domains.

The rest of the paper is organised as follows. The Related Work section outlines the working principles behind existing main approaches in sentiment analysis. The Materials and Methods section describe the datasets used for this investigation and the workflow followed for the empirical study. The Evaluation and Comparative Analysis section illustrates the results of the empirical study and further analyzes the performance of various methods and solutions. The final section concludes the paper summarizing the main findings and outlining the future work.

RELATED WORK

SENTIMENT ANALYSIS AND SENTIMENT CLASSIFICATION APPROACHES

The computational study of people's views, sentiments, emotions, assessments, and attitudes regarding entities such as products, services, organizations, individuals, issues, events, themes, and their qualities is known as sentiment analysis or opinion mining (Liu et al., 2019). To make the proper judgment, the Sentiment Analysis must process data and understand its polarity. Sentiment polarity is a key feature of text that is often classified as either positive or negative, but not always strictly binary; multiple polarities in a range may exist. A document containing both positive and negative opinions has a mixed polarity whereas an objective text that presents facts without personal feelings or opinions has no polarity or remain neutral (Mejova, 2009). As a branch of Sentiment Analysis, sentiment classification particularly refers to the process of extracting and identifying useful features from source materials and accordingly developing effective classification models to categorize a given piece of text into one of the predefined sentiment polarities. Such classification models can then track the public's feelings about a specific product, event, or issue. In other words, the enormous volume of raw text data can be converted into useful information to support decision making.

Sentiment classification techniques can be categorized into two main types: lexicon analysis and machine learning (Patel & Choksi, 2015). The lexicon-based approach is based on pre-built sentiment lexicons or dictionaries that contain words and their associated scores for sentiment polarity levels. A sentiment lexicon is a collection of lexical properties that are classified as positive or negative based on their semantic orientation (Bonta et al., 2019). In this approach, a model is trained to recognize the sentiment of a text based on a collection of labeled training data. Many ML approaches, including decision trees, SVMs and artificial neural networks, may be used to train such models. When analyzing complicated text data, ML-based algorithms can be more accurate than rule-based analysis, but they also require more labeled training data and may be computationally more expensive (Ali, 2023). A Decision Tree (DT), for example, is a supervised ML technique for classification and regression. It splits the training dataset using greedy depth-first or breadth-first search until all data items belong to the same class, which can be used for sentiment classification.

Due to their capacity to recognize intricate textual patterns, deep learning models, particularly deep learning neural networks, have become very popular in recent years. Recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models, such as BERT (see later) and GPT, are frequently employed. These models that have been pre-trained on enormous volumes of text data are capable of encoding the context and meaning of text through sophisticated neural networks, allowing them to achieve state-of-the-art accuracy for a wide range of NLP tasks, including sentiment analysis. These models, however, need large amounts of computing resources and may not be appropriate for all use situations.

Transformer, a sophisticated neural network architecture, was introduced by (Vaswani et al., 2017) for sequence transduction to improve machine translation efficiency. At the heart of transformers is an attention mechanism that allows the model to focus on various sections of the input sequence while predicting each word in the output sequence, much like what humans do when reading or listening. A common transformer model includes an encoder for processing input and a decoder for producing the output. Both the encoder and decoder are composed of numerous identical layers, including attention mechanisms and fully linked neural network. Unlike prior sequence-to-sequence models, transformers process complete sequences concurrently instead of sequentially, significantly improving training efficiency. The Transformer achieved high performance on the WMT2014 English-German and English-French datasets, achieving a BLEU score of 28.4 and 41.8 respectively. One limitation of this architecture is that it is computationally intensive.

Introduced in 2018 (Devlin et al., 2019), Bidirectional Encoder Representations from Transformers (BERT) marked a significant breakthrough in NLP. By pre-training on extensive text corpora and then fine-tuning for specific tasks, BERT transformed the way NLP models are developed; unlike earlier models that read text sequentially, BERT understands the context of a word by looking at both its left and right surroundings, offering a major advancement in contextual representation. Soon afterwards, a Robustly Optimized BERT Pretraining Approach (RoBERTa) was further developed by Liu et al. (2019) as an optimization of the basic BERT. Key hyperparameters in BERT were adjusted, including the elimination of the next-sentence pretraining objective and training with significantly larger sub-batches and adjustment of the learning rates.

Existing Work in Sentiment Analysis

A wide range of studies has explored sentiment analysis using traditional machine learning, lexicon-based approaches, and deep learning models. Severyn and Moschitti (2015) introduced a deep learning system using Convolutional Neural Networks (CNNs) for Twitter sentiment analysis. Their novel weight initialization technique enabled the model to achieve top performance on the Twitter'15 test set for phrase-level sentiment classification and a second-place ranking in message-level sentiment tasks. However, performance varied across datasets, with reduced accuracy on LiveJournal'14, highlighting the dependence of model generalization on the diversity and quality of training data

In a comparative study, Wang (2024) evaluated sentiment classification using traditional machine learning methods such as Support Vector Machines (SVM) and Naive Bayes (NB), alongside deep learning models like Bidirectional Encoder Representations from Transformers (BERT) and its optimized variant RoBERTa. RoBERTa achieved the highest accuracy (87.44%) and F1-score (0.8746), demonstrating strong performance in capturing nuanced sentiments. However, transformer models required significant training time, approximately three hours and computational power. In contrast, lexicon-based methods like Valence Aware Dictionary and sEntiment Reasoner (VADER) offered quick and interpretable outputs, though they performed poorly on complex or sarcastic texts.

Traditional ML models such as SVM and Naive Bayes have historically been effective for sentiment analysis due to their simplicity and low computational cost. However, as noted by L. Zhang et al. (2018), deep learning models, including CNNs, outperform traditional approaches in accuracy (e.g., 90.1% vs. 83.5% for CNN vs. logistic regression), albeit with increased computational overhead.

Kumari and Singh (2024) similarly found that BERT reached 92.3% accuracy, outperforming SVM's 86.7%, but with greater inference time and resource needs. They also noted that model performance is highly sensitive to dataset characteristics. Deep learning models excel with complex language (e.g., sarcasm in tweets), while traditional models are effective on simpler texts like product reviews. For instance, Chen (2025) developed a real-time Twitter sentiment tracking system using Logistic Regression, favoring speed over deep contextual understanding.

Dake and Gyimah (2023) examined sentiment analysis on open-ended survey feedback using models such as SVM, Naive Bayes, J48, and Random Forest. SVM achieved the highest accuracy of 63.79%, validated through k-fold cross-validation with real-world predictive accuracy reaching 92%, demonstrating practical applications in educational contexts.

Basarslan and Kayaalp (2021) compared traditional and deep learning methods using datasets from Internet Movie Database (IMDB), Yelp, and Twitter. BERT-based models outperformed text representation methods like Term Frequency–Inverse Document Frequency (TF-IDF) and Word2Vec, achieving accuracies between 84% and 98%. Traditional models ranged between 75% and 90%.

Ng et al. (2023) benchmarked eight models including Light Gradient Boosting Machine (LightGBM), SVM, Deep Neural Networks (DNN), Long Short-Term Memory (LSTM), BERT, and RoBERTa on e-commerce product reviews. SVM delivered the best performance, with an accuracy of 73.98% and the highest precision, recall, and F1 scores. However, the study did not extend to social media contexts.

Ashbaugh and Zhang (2024) compared traditional models such as NB and logistic regression against CNNs and Recurrent Neural Networks (RNNs) for customer feedback analysis. Deep learning models provided more accurate sentiment classification but required larger datasets and struggled to detect subtle emotional signals.

Chen (2025) examined multiple deep learning architectures including LSTM, a hybrid LSTM-CNN, and BERT, using the IMDB dataset. While these models excelled in binary sentiment classification, they faced challenges in multilingual environments and raised privacy concerns when applied to real-time data pipelines.

Dang et al. (2020) examined deep learning for social media sentiment analysis. Deep models improved accuracy but struggled with informal language, slang, and sarcasm on platforms like Twitter and Facebook.

Oumaima et al. (2024) reviewed both traditional and deep models on social media. BERT and RoBERTa performed well on unstructured data, while VADER worked best on structured text. They highlighted the difficulty of extracting meaning from noisy, informal text.

Xu and Song (2023) conducted experiments on online commentary using SVM, CNN, and a hybrid BERT+CNN model. The hybrid achieved the best accuracy, recall, and F1 score, with LSTM excelling in recall and Gated Recurrent Units (GRU) scoring highest in F1. These results underscore the potential of combining architectures for better sentiment classification on social platforms.

In their review, Ganie and Dadvandipour (2022) compared various sentiment analysis techniques across datasets and text representation methods. They emphasized the importance of aligning model selection with dataset characteristics and noted that transforming raw social media content into meaningful features remains a major obstacle.

Furthermore, Gul et al. (2025) explored advanced transformer architectures—including BERT-base, A Lite BERT (ALBERT), RoBERTa-base, and eXtreme Language Understanding (XLNet)—for three tasks: target-oriented opinion extraction, aspect-level sentiment analysis, and document-level sentiment classification. While Fine-Tuned RoBERTa (FT-RoBERTa) achieved the best results in aspect-level tasks, their Target-Oriented Opinion Word Extraction (TOWE) model required manually annotated aspects, highlighting a current limitation in automatic aspect detection.

MATERIALS AND METHODS

DATASETS

We utilized three publicly available benchmark datasets for sentiment classification. The IMDB movie reviews, Yelp Polarity reviews, and Amazon Polarity product reviews, each offering binary sentiment labels and distinct textual and structural characteristics in terms of size, text length, and vocabulary richness. Table 1 provides a comparative overview including sample size, average text length, word count, label format, and data source. IMDB reviews are longer and more detailed, while Amazon reviews are typically concise and accompanied by titles. Yelp reviews strike a balance in length and content structure. All datasets maintain balanced binary labels for consistency across evaluation.

Table 1. Overview of Dataset Properties

PROPERTIES	IMDB POLARITY	YELP POLARITY	AMAZON POLARITY
Samples (Train/Test)	25k / 25k	560k / 38k	3.6M / 400k
Avg. Text Length (chars)	1325	726	405
Avg. Word Count	234	133	74
Duplicates	Yes (low)	None	None
Special Char Ratio	~4%	~3.8%	~3.1%
Class Balance	Balanced	Balanced	Balanced
Original Labels	Positive / Negative	1–5 star ratings (3-star reviews removed)	1–5 star ratings (3-star reviews removed)
Labels Used	Positive, Negative	Positive, Negative	Positive, Negative
Label Type	Binary	Binary	Binary
Data Source	Maas et al., 2011 (ACL)	X. Zhang et al., 2015 (NeurIPS)	X. Zhang et al., 2015 (NeurIPS)

All sentiment classification tasks in this study were formulated as a binary classification problem — distinguishing between positive and negative sentiment. To achieve this, datasets and tools that originally contained more than two sentiment categories were systematically mapped to a binary format. For the Yelp Polarity and Amazon Polarity datasets, which were originally derived from 1-to-5 star rating systems, only highly polarized reviews were retained: ratings of 1 and 2 stars were mapped to the negative class, while ratings of 4 and 5 stars were mapped to the positive class. Neutral reviews (3-star ratings) were excluded during dataset construction to ensure clear sentiment polarity, as per the approach defined in the polarity versions introduced by X. Zhang et al. (2015). For the VADER sentiment tool, which outputs a compound score ranging from -1 (most negative) to +1 (most positive), a thresholding method was applied to align with the binary classification framework. Following the authors' recommendation (Hutto & Gilbert, 2014), scores equal to or above +0.05 were categorized as positive, while scores equal to or below -0.05 were considered negative. Neutral scores between these thresholds were excluded to maintain binary class integrity.

In the case of pretrained transformer models such as DistilBERT and RoBERTa, sentiment predictions were used directly without further fine-tuning, relying on the model's built-in binary classification heads trained on domain-specific corpora. The IMDB dataset was natively binary (positive or negative reviews), so no label transformation was required. This conversion process ensured label consistency across all resources, enabling a fair and controlled comparison of model performance under a unified binary classification setting.

METHODS

Data preprocessing

Before supplying raw text to feature extraction techniques and machine learning models, the preprocessing pipeline was carefully designed to organise and clean it. A detailed explanation of each transformation used on the text data is as follows.

1. Text Normalization

The raw review texts were first converted entirely to lowercase. This step ensured uniformity across words that might otherwise be treated differently due to case sensitivity. For instance, "Excellent" and "excellent" would be considered the same after this transformation, eliminating redundancy and reducing the vocabulary size.

2. Removal of Noise and Irrelevant Characters

All characters in the text that were not English alphabet letters or whitespace were removed. This included punctuation marks (like !, ., ?), digits (such as ratings like 10/10), and special characters (like @, #, *, etc.). The aim was to eliminate visual and syntactic clutter that typically does not contribute positively to sentiment classification, especially in traditional models such as Logistic Regression or Naive Bayes. This step had a direct impact on the "Average Special Character Ratio" measured in the data analysis, which quantified the proportion of such characters in the raw (pre-cleaning) dataset. The ratio served as an indicator of textual noise in different datasets and justified the need for cleaning.

3. Tokenization (Splitting Text into Words)

After cleaning, the text was split into individual words using standard tokenization techniques. Tokenization is the process of identifying word boundaries so that the text can be analyzed at the word level. For example, the sentence "The food was amazing!" would become the sequence: ["the", "food", "was", "amazing"]. This step laid the groundwork for downstream techniques like stopword removal and vectorization

4. Stopword Elimination

Stopwords — extremely common words that carry little meaningful information — were removed from the tokenized text. Examples include "the", "is", "in", "at", "of", etc. These words, while necessary for grammatical structure, often dilute the sentiment signal in text classification tasks. The decision to remove them was based on the goal of emphasizing words that are more sentiment-revealing (e.g., "terrible", "great", "boring"), which have a much higher predictive value.

5. Final Text Reconstruction

After cleaning, tokenizing, and filtering out stopwords, the remaining words were reassembled into a space-separated string. This version of the text was then used for feature extraction (Bag of Words, TF-IDF, or Word2Vec). The cleaned output retained only meaningful, lowercase words without punctuation, numbers, or stopwords.

Traditional machine learning methods

In this study, we applied three supervised traditional machine learning algorithms — Logistic Regression (LR), Decision Tree (DT), and Naive Bayes (NB) — for sentiment classification. To represent the textual data numerically, we evaluated three widely used feature extraction techniques: Bag of Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF) and Word2Vec embeddings. The preprocessing pipeline for these models involved several stages. First, the raw text was normalized by converting all characters to lowercase and trimming excess whitespace. Next, irrelevant elements such as punctuation marks, digits, and special characters were removed to reduce noise. The cleaned text was then tokenized into individual words, and common stop-words such as "the," "is," and "in" were filtered out to retain only sentiment-relevant terms. Finally, the processed tokens were reassembled into space-separated strings suitable for vectorization. Feature extraction for BoW and TF-IDF was performed using Python's scikit-learn and NLTK libraries while Word2Vec embeddings were generated using Gensim, enabling the models to capture both frequency-based and semantic representations of the input text.

Lexicon-based method

The VADER sentiment analyzer was used as a lightweight rule-based baseline. It operates directly on raw text and assigns compound polarity scores without any training phase. Texts were classified using recommended thresholds for compound scores: Positive: ≥ +0.05. Negative: ≤ –0.05. Neutral scores were discarded to maintain binary classification alignment**.**

Transformer-based deep learning

We evaluated two transformer models—DistilBERT-SST-2 and RoBERTa-Sentiment—accessed via Hugging Face's Transformers library. Pretrained versions were used without additional fine-tuning. These models are designed to infer sentiment directly from text using contextual embeddings and classification heads trained on domain-specific corpora (e.g., SST-2, Twitter).

EXPERIMENTAL SETUP AND PERFORMANCE EVALUATION METRICS

All experiments were conducted on a MacBook Pro equipped with an Apple M1 Pro chip, 16 GB of unified memory, and running macOS Ventura (version 13.5.1). The trained models were run in inference mode only, without task-specific fine-tuning, and computations were accelerated using Apple's Metal Performance Shaders (MPS) backend in PyTorch. The system provided sufficient processing power for training traditional models, conducting inference on deep models, and processing all datasets efficiently.

To evaluate model performance, we adopted four widely accepted classification metrics: Accuracy, Precision, Recall, and F1 Score. These metrics offer a balanced view of both overall correctness and the ability to identify sentiment-specific patterns. Accuracy measures the proportion of correctly predicted instances out of the total predictions. Precision assesses how many of the predicted positive instances are actually positive, whereas Recall evaluates how many of the actual positive instances are correctly identified. F1 Score is the harmonic mean of precision and recall, balancing the trade-off between recall and precision. The formulae for the performance metrics are given as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\textbf{F1 Score} = \mathbf{2} \times \frac{\textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}}$$

where true positives (**TP**), true negatives (**TN**), false positives (**FP**), and false negatives (**FN**) refer to the four respective possible outcomes of sentiment predictions. These metrics were computed using scikit-learn's built-in evaluation functions to ensure consistency across model types and datasets. This evaluation framework facilitated a robust and standardized comparison of model performance across various sentiment classification techniques. The complete implementation, including source code and dataset access details, is available in the Appendix.

EVALUATION AND COMPARATIVE ANALYSIS

This section presents a detailed comparative analysis of sentiment classification models across the three benchmark datasets. We use tables to summarize the performance of the trained models when they are tested using the designated test sets already split in the benchmark datasets.

VADER: Lexicon-Based Analysis

Table 2 summarizes the accuracies of the VADER (Valence Aware Dictionary and sEntiment Reasoner) tool across the three datasets. The tested accuracy is the lowest for the Amazon dataset while the tested accuracy for Yelp dataset reaches the highest with the accuracy for IMDB dataset in the middle. All accuracy levels are similar to each other. While VADER performs well on short, informal texts, its performance suffered on longer reviews. Despite its simplicity and efficiency, its performance is comparatively lower than traditional and deep learning models (see later), indicating its limited effectiveness on more complex or lengthy review texts.

Table 2. Performance of VADER for Sentiment Analysis

DATASET	MODEL	ACCURACY
IMDB	VADER	0.6974
Yelp	VADER	0.7134
Amazon	VADER	0.6966

Traditional Machine Learning Models

Table 3 presents a detailed evaluation of the three traditional machine learning models when each model was trained on the three types of text features. The results show that Logistic Regression(LR) consistently outperforms other models, particularly when paired with TF-IDF features. Naive Bayes(NB) also performs well with BoW and TF-IDF but is incompatible with Word2Vec due to negative values. Decision Trees(DT) generally exhibit the lowest performance across all datasets and feature methods.

Table 3. Performance Comparison of Traditional Machine Learning Models Using Different Feature Extraction Techniques

DATASET	METHOD	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
IMDB	BoW	LR	0.8483	0.8484	0.8483	0.8483
		DT	0.7103	0.7103	0.7103	0.7103
		NB	0.8381	0.8389	0.8381	0.8380
	TF-IDF	LR	0.8800	0.8801	0.8800	0.8800
		DT	0.7079	0.7081	0.7079	0.7078
		NB	0.8424	0.8428	0.8424	0.8424
	Word2Vec	LR	0.8047	0.8048	0.8047	0.8047
		DT	0.6626	0.6627	0.6626	0.6626
Yelp	BoW	LR	0.9246	0.9246	0.9246	0.9245
		DT	0.7907	0.7907	0.7907	0.7907
		NB	0.8622	0.8634	0.8622	0.8621
	TF-IDF	LR	0.9268	0.9268	0.9268	0.9268
		DT	0.7926	0.7926	0.7926	0.7926
		NB	0.8752	0.8753	0.8752	0.8752
	Word2Vec	LR	0.9053	0.9053	0.9053	0.9053

DATASET	METHOD	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
		DT	0.7994	0.7994	0.7994	0.7994
Amazon	BoW	LR	0.8661	0.8663	0.8661	0.8661
		DT	0.7546	0.7546	0.7546	0.7546
		NB	0.8197	0.8197	0.8197	0.8197
	TF-IDF	LR	0.8666	0.8666	0.8666	0.8666
		DT	0.7479	0.7479	0.7479	0.7479
		NB	0.8204	0.8204	0.8204	0.8204
	Word2Vec	LR	0.8475	0.8475	0.8475	0.8475
		DT	0.7263	0.7263	0.7263	0.7263

Transformer-Based Models

Table 4 presents evaluation metrics for two transformer-based models (DistilBERT-SST-2 and RoBERTa-Sentiment) on the three benchmark datasets. DistilBERT-SST-2, a distilled version of BERT fine-tuned on SST-2, shows consistently high performance across all datasets, with IMDB achieving the highest F1 score of 0.9418. In contrast, RoBERTa-Sentiment, pre-trained on Twitter data, performs notably lower, particularly on Yelp and Amazon, indicating domain misalignment. Despite having high precision, RoBERTa's low recall results in imbalanced F1 scores.

Table 4. Transformer-Based Sentiment Classification Performance

DATASET	MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Yelp	DistilBERT-SST-2	0.85	0.8541	0.85	0.8499
	RoBERTa-Sentiment	0.42	0.7037	0.42	0.4956
Amazon	DistilBERT-SST-2	0.85	0.8518	0.85	0.8501
	RoBERTa-Sentiment	0.47	0.7823	0.47	0.5414
IMDB	DistilBERT-SST-2	0.89	1.0000	0.89	0.9418
	RoBERTa-Sentiment	0.79	1.0000	0.79	0.8827

Comparative Insights and Discussion

The comparative analysis across the IMDB, Yelp, and Amazon Polarity datasets reveals the strengths and limitations of traditional, deep learning, and lexicon-based sentiment classifiers. Logistic Regression with TF-IDF features consistently outperformed other traditional models, achieving accuracies of 0.8800 (IMDB), 0.9268 (Yelp), and 0.8666 (Amazon). In contrast, the lexicon-based VADER analyzer underperformed across all datasets, with accuracies of 0.6974, 0.7134, and 0.6966 respectively highlighting its limited effectiveness on long-form and nuanced reviews. McNemar's test confirmed statistically significant differences in prediction distributions between Logistic Regression and DistilBERT ($p < 0.001$), suggesting that even modest accuracy gaps reflect meaningful divergence in model outputs.

In particular, this comprehensive evaluation of sentiment classification models across three benchmark datasets yields several key observations regarding model behavior, feature efficacy, and generalization performance. Among traditional models, Logistic Regression consistently emerged as the most effective classifier, particularly when paired with TF-IDF feature representations. This combination delivered robust performance across all datasets, benefiting from TF-IDF's ability to emphasize informative terms while down-weighting frequent but less discriminative tokens. Naive Bayes also demonstrated competitive results with both BoW and TF-IDF features; however, it was unsuitable for Word2Vec due to its reliance on non-negative input distributions. Decision Trees, by contrast, generally underperformed relative to the other models. Their limited depth and inability to capture nuanced linguistic patterns likely contributed to this trend. Overall, the results reaffirm the practicality of TF-IDF as a lightweight, high-utility feature extraction method, particularly when deep learning resources are unavailable.

The VADER sentiment analyzer provided a useful lightweight baseline, but its performance lagged significantly behind both traditional and deep learning models. VADER's heuristic-driven design proved insufficient for complex or long-form text typical of review datasets such as IMDB and Amazon. Although optimized for short and informal texts, such as tweets or social media posts, VADER's generalizability is limited in domains that require deeper semantic understanding or context awareness. Its relatively better performance on Yelp data reflects this design alignment.

The DistilBERT-SST-2 model outperformed all other approaches, achieving the highest accuracy and F1 scores across datasets, most notably an F1 score of 0.9418 on IMDB. As a distilled version of BERT fine-tuned on the SST-2 corpus, it delivers a favorable balance between efficiency and performance. In contrast, RoBERTa-Sentiment, despite achieving perfect precision on certain datasets, exhibited substantially lower recall and F1 scores. This suggests a tendency toward conservative predictions and indicates that the model's training data—centered on Twitter content—did not generalize well to the more structured language in product and movie reviews.

Transformer-based models like DistilBERT demonstrated strong and consistent generalization without task-specific fine-tuning, surpassing both traditional and lexicon-based methods in accuracy scoring 0.89 on IMDB and 0.85 on both Yelp and Amazon. RoBERTa, however, exhibited inconsistent performance due to domain mismatch, achieving only 0.42 accuracy on Yelp, at times performing worse than VADER. Confidence intervals computed for DistilBERT and Logistic Regression on the IMDB dataset (±0.007) further underscore their statistically distinct performance. These findings affirm that while transformer models lead in raw accuracy, traditional classifiers remain strong, interpretable baselines, and lexicon-based tools like VADER are best suited for brief, informal text settings.

These findings emphasize the critical role of domain alignment in pretrained transformer models. While RoBERTa may excel in short-text, social media domains, DistilBERT's fine-tuning on sentence-level classification tasks like SST-2 made it more suitable for review-based sentiment analysis.

CONCLUSION

This paper has presented a comprehensive comparative evaluation of sentiment classification models across traditional machine learning, lexicon-based methods, and transformer-based deep learning approaches. By assessing models on three diverse datasets: IMDB, Yelp, and Amazon Polarity. We identified consistent patterns in performance and generalization. The results reaffirm that transformer models, especially DistilBERT, offer superior performance without task-specific fine-tuning, making them ideal for domains requiring high accuracy and contextual understanding. However, traditional models like Logistic Regression paired with TF-IDF remain strong contenders, especially when computational efficiency and model interpretability are prioritized. Lexicon-based tools such as VADER, while lightweight and fast, are best suited to short, informal text and struggle with nuanced or domain-specific sentiment.

Overall, this study highlights that model selection should align with the text complexity, resource availability, and specific application context. While deep learning sets the performance benchmark, well-optimized traditional methods offer a reliable alternative when simplicity and speed are key.

Contribution And Future Research

This study offers a comprehensive, side-by-side comparison of conventional and modern sentiment classification techniques, identifying model-feature synergies and highlighting limitations in generalizability across domains. The study can improve applications in customer feedback systems, social media monitoring, and public opinion analysis, contributing to more responsive and adaptive services. Future work should explore domain-specific fine-tuning strategies and evaluate multilingual sentiment performance, especially for transformer models.

This research opens avenues for further exploration into hybrid interpretability-performance models and zero-shot or few-shot sentiment classification using large language models.

References

Ali, M. (2023). *NLTK sentiment analysis tutorial for beginners*. DataCamp.

Ashbaugh, L., & Zhang, Y. (2024). A comparative study of sentiment analysis on customer reviews using machine learning and deep learning. *Computers, 13*(12), 340. https://doi.org/10.3390/computers13120340

Basarslan, M. S., & Kayaalp, F. (2021). Sentiment analysis on social media reviews datasets with deep learning approach. *Sakarya University Journal of Computer and Information Sciences, 4*(1), 35–49. https://doi.org/10.35377/SAUCIS.04.01.833026

Biecek, P., & Samek, W. (2024). Position: Explain to question not to justify. *Proceedings of the 41st International Conference on Machine Learning.* arXiv:2402.13914 https://arxiv.org/abs/2402.13914

Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A comprehensive study on lexicon based approaches for sentiment analysis. *Asian Journal of Computer Science and Technology, 8*(2249-0701), 1-6. https://doi.org/10.51983/ajcst-2019.8.S2.2037

Chen, S. (2025). Sentiment analysis techniques for deep learning classification and comparison. *Theoretical and Natural Science, 86*(1), 74–80. https://doi.org/10.54254/2753-8818/2025.20342

Dake, D. K., & Gyimah, E. (2023). Using sentiment analysis to evaluate qualitative students' responses. *Education and Information Technologies, 28*(4), 4629-4647. https://doi.org/10.1007/s10639-022-11349-1

Dang, N. C., Moreno-García, M. N., & De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. arXiv: Computation and Language. https://doi.org/10.3390/electronics9030483

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 4171-4186.

Ganie, A., & Dadvandipour, S. (2022). Traditional or deep learning for sentiment analysis: A review. Multidiszciplináris Tudományok, 12(1), 3–12. https://doi.org/10.35925/j.multi.2022.1.1

Gul, S., Asif, M., Saleem, K., & Imran, M. (2025). Advancing aspect-based sentiment analysis in course evaluation: A multi-task learning framework with selective paraphrasing. *IEEE Access, 13*, 7764-7779. https://doi.org/10.1109/ACCESS.2025.3527367

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the international AAAI conference on web and social media*, *8*(1), 216-225. https://doi.org/10.1609/icwsm.v8i1.14550

Kumari, S., & Singh, M. P. (2024). Machine learning-based election results prediction using Twitter activity. *SN Computer Science, 5*(7), Artricle 819. https://doi.org/10.1007/s42979-024-03180-x

Liu, Y., Ott, M., Goyal, N., Du, J., Josh, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A robustly optimized BERT pretraining approach.* https://doi.org/10.48550/arXiv.1907.11692

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics.*

Mejova, Y. (2009). Sentiment analysis: An overview. University of Iowa, Computer Science Department, 5, 1-34. https://www.academia.edu/291678/Sentiment_Analysis_An_Overview

Ng, J. X., Lim, K. M., Lee, C. P., Lim, Q. Z., Ooi, E. K. H., & Loh, N. K. N. (2023, August). Sentiment analysis using learning-based approaches: A comparative study. In *2023 11th International Conference on Information and Communication Technology (ICoICT*), 469-474. https://doi.org/10.1109/ICoICT58202.2023.10262604

Oumaima, B., Baïna, A., & Bellafkih, M. (2024). Deep learning or traditional methods for sentiment analysis: A review. In M. Ben Ahmed, A. A. Boudhir, R. El Meouche, R., İ. R. Karaş (Eds), *Innovations in Smart Cities Applications* Volume 7. The Proceedings of the 8th International Conference on Smart City Applications. https://doi.org/10.1007/978-3-031-53824-7_3

Patel, S. N., & Choksi, M. J. B. (2015). A survey of sentiment classification techniques. *Journal for Research, 1*(1), 15-20.

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 959-962. https://doi.org/10.1145/2766462.2767830

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30. https://arxiv.org/pdf/1706.03762.pdf

Wang, F. (2024). Comparative evaluation of sentiment analysis methods: From traditional techniques to advanced deep learning models. *Applied and Computational Engineering, 105*, 23-29. https://doi.org/10.54254/2755-2721/105/2024TJ0056

Xu, L., & Song, Y. (2023). Comparison of text sentiment analysis based on traditional machine learning and deep learning methods. *International Conference Civil Engineering and Architecture*, 692–695. https://doi.org/10.1109/ICCEA58433.2023.10135273

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8*(4), 1253. https://doi.org/10.1002/widm.1253

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level Convolutional Networks for Text Classification. In Advances in Neural Information Processing Systems.

APPENDIX

The complete codebase used for this study is publicly available at the following GitHub repository:

https://github.com/philipobiorah/Comparative_Study_of_Sentiment_analysis_techniques

All datasets used in this research; IMDB, Yelp Polarity, and Amazon Polarity were accessed via Hugging Face's datasets library. This ensured consistent preprocessing, easy integration with machine learning pipelines, and reproducibility.

AUTHORS

Philip Obiorah works as a Learning Development Coach in Computing at the School of Computing, University of Buckingham. His main research interests are data science, machine learning, generative AI, natural language processing, and data storytelling. He is particularly interested in applying AI techniques to real-life problems in practical applications such as sentiment analysis, topic classification, and disaster alert systems. He is an active member of the British Computer Society and leads the Google Developer Groups (GDG) Cloud Port Harcourt community.

Grace Diri is a Research Scientist in Artificial Intelligence, Machine Learning, and Computer Vision. She holds a Ph.D. in Artificial Intelligence and Machine Learning. Her main research interests include AI, Machine Learning, and their application to real-world problems. She is particularly focused on using these technologies for practical solutions such as distance learning style recommendation, facial age estimation, abnormal heart rate detection, and blockchain-based food traceability.

Hongbo Du works as a Professor in Computing at the School of Computing University of Buckingham. His main research interests are big data, data mining, machine learning and medical image analysis. He is particularly interested in AI and Machine Learning for real-life problems in practical applications.