Scientific Publications

Journal of Artificial Intelligence and Big Data, 2026, 6, 6162
www.scipublications.com/journal/index.php/jaibd
DOI: 10.31586/jaibd.2026.6162

Article

Predictive Modeling of Public Sentiment Using Social Media Data and Natural Language Processing Techniques

Lawrence A. Farinola [1,2,*], Jean-Eudes Assogba [1,2]

[1] Department of Software Engineering, Faculty of Architecture and Engineering, Rauf Denktas University, Mersin 10 via Turkey

[2] Center of Excellence for Interdisciplinary AI and Data Science Research, Rauf Denktas University, Mersin 10 via Turkey

*Correspondence: Lawrence A. Farinola (Lawrence.farinola@rdu.edu.tr)

Abstract: Social media platforms like X (formerly Twitter) generate vast volumes of user-generated content that provide real-time insights into public sentiment. Despite the widespread use of traditional machine learning methods, their limitations in capturing contextual nuances in noisy social media text remain a challenge. This study leverages the Sentiment140 dataset, comprising 1.6 million labeled tweets, and develops predictive models for binary sentiment classification using Naive Bayes, Logistic Regression, and the transformer-based BERT model. Experiments were conducted on a balanced subset of 12,000 tweets after comprehensive NLP preprocessing. Evaluation using accuracy, F1-score, and confusion matrices revealed that BERT significantly outperforms traditional models, achieving an accuracy of 89.5% and an F1-score of 0.89 by effectively modeling contextual and semantic nuances. In contrast, Naive Bayes and Logistic Regression demonstrated reasonable but consistently lower performance. To support practical deployment, we introduce SentiFeel, an interactive tool enabling real-time sentiment analysis. While resource constraints limited the dataset size and training epochs, future work will explore full corpus utilization and the inclusion of neutral sentiment classes. These findings underscore the potential of transformer models for enhanced public opinion monitoring, marketing analytics, and policy forecasting.

Keywords: Sentiment Analysis; Social Media Mining; Public Opinion Prediction; Natural Language Processing; BERT Transformer Model

How to cite this paper:
Farinola, L. A., & Assogba, J.-E. (2026). Predictive Modeling of Public Sentiment Using Social Media Data and Natural Language Processing Techniques. *Journal of Artificial Intelligence and Big Data, 6*(1), 1–12. DOI: 10.31586/jaibd.2026.6162

Received: July 24, 2025
Revised: October 30, 2025
Accepted: February 2, 2026
Published: February 6, 2026

1. Introduction

The proliferation of social media platforms has fundamentally transformed how public opinion is formed, shared, and analyzed. Among these platforms, X (formerly known as Twitter) stands out as a microblogging service capturing real-time expressions of public sentiment across diverse topics, including consumer behavior, political discourse, and global events. With over 500 million tweets posted daily, X (formerly Twitter) provides a rich and dynamic data source for public opinion mining through sentiment analysis [1]. Recent studies further emphasize the growing role of data-driven machine learning models in supporting public policy formulation and socioeconomic forecasting, particularly in contexts involving large-scale societal change and technological transformation [2].

Sentiment analysis, also known as opinion mining, is a subfield of Natural Language Processing (NLP) and computational linguistics focused on detecting, extracting, and classifying subjective information in text [3]. This technique enables automatic categorization of textual content, such as tweets, into predefined sentiment classes:

positive, negative, or neutral [4]. NLP methods transform raw human language into structured data suitable for machine learning models [5]. Similarly, large-scale data-driven modeling approaches have been applied to population-level societal and public health challenges, illustrating how predictive analytics can inform public awareness and policy planning [6].

Recent advances in machine learning, especially deep learning architectures, have significantly improved sentiment classification accuracy. However, a key question remains: to what extent do these advanced models outperform traditional machine learning techniques in real-world sentiment analysis tasks?

This study benchmarks the performance of classical machine learning algorithms against transformer-based deep learning models using the publicly available Sentiment140 dataset [4], which contains 1.6 million labeled tweets. We compare three sentiment classification approaches: Naive Bayes, Logistic Regression (a variant of the Maximum Entropy model), and BERT (Bidirectional Encoder Representations from Transformers) [7]. The study focuses on preprocessing raw tweet text into model-compatible features, training and optimizing each classifier on the same dataset, evaluating performance using accuracy, F1-score, and confusion matrices, and identifying the most accurate and robust approach for public sentiment prediction.

Unlike prior studies focusing either on traditional machine learning or deep learning in isolation, our work provides a comprehensive and systematic comparison under consistent experimental conditions. This research offers practical implications for political forecasting, brand monitoring, and real-time crisis detection, where accurate and timely understanding of public sentiment can inform decision-making. Beyond sentiment analysis, artificial intelligence and machine learning techniques have been successfully applied to real-world decision-support systems, such as academic scheduling and institutional resource optimization, demonstrating the versatility of AI models across application domains [8]. Additionally, we introduce an interactive tool (SentiFeel) for real-time sentiment classification to bridge academic research and real-world applications.

Research questions addressed include: Which model yields the highest classification accuracy on general-purpose X sentiment analysis? How do traditional models like Naive Bayes and Logistic Regression compare to state-of-the-art deep learning approaches such as BERT?

To answer these questions, this study applies consistent preprocessing, training, and evaluation procedures across all models using the Sentiment140 dataset. By maintaining uniform experimental conditions, we aim to ensure a fair and reliable comparison of performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

This work contributes to advancing natural language processing and machine learning by offering empirical insights into the strengths and limitations of each modeling approach. Ultimately, it supports the development of reliable social media analytics tools that can capture and interpret public sentiment in real-time.

2. Literature Review

Early research in X-based sentiment analysis laid the foundation for contemporary studies by introducing critical datasets and baseline methodologies. Go et al. [9] created the Sentiment140 dataset using distant supervision, automatically labeling tweets based on emoticons and applying traditional classifiers like Naive Bayes, Logistic Regression, and Support Vector Machines (SVM), achieving accuracy levels exceeding 80%. Pak and Paroubek [1] also validated X as a viable source for sentiment classification, developing a three-way classifier (positive, negative, neutral) using supervised models trained on automatically collected tweets.

Lexicon-based approaches predate machine learning methods and use predefined sentiment dictionaries to assign scores to words or phrases. For example, SentiWordNet assigns sentiment values to WordNet synsets [12], while VADER—a rule-based tool

designed specifically for social media—considers features like capitalization and punctuation to enhance sentiment detection [7].

While these approaches are efficient for identifying general sentiment trends, they often fall short when dealing with complex language features such as negation, slang, sarcasm, or irony. This limitation reduces their effectiveness in the nuanced and informal text typically found on platforms like X [12].

Traditional machine learning methods such as Naive Bayes, SVM, and Logistic Regression have been widely used for sentiment classification across domains. Pang and Lee demonstrated their effectiveness on movie reviews [10, 11]. Applied to X data, studies including Go et al. [9] and Mohammad et al. [13] utilized unigram/bigram models, TF-IDF weighting, and emoticon features, achieving 70–80% accuracy on large corpora. However, these models depend heavily on manual feature engineering, limiting scalability and generalizability.

Recent advances in deep learning have enhanced sentiment analysis performance. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, capture sequential dependencies in text. Mollah [12] developed an LSTM-based X sentiment classifier with notable accuracy gains. Convolutional Neural Networks (CNNs) have also been applied to short-form content. However, the most transformative progress comes from transformer-based architectures. BERT, introduced by Devlin et al. [10], leverages self-attention to capture bidirectional context, enabling nuanced language understanding. Bello et al. [14] designed a BERT-based framework for tweet sentiment classification, reporting superior performance over traditional and earlier deep learning models.

While lexicon-based and traditional ML approaches remain valuable baselines, they fall short in capturing semantic and pragmatic nuances. Transformer-based models such as BERT provide context-aware representations suited to social media's informal language. However, few studies have directly compared traditional models and BERT on consistent datasets with both quantitative evaluation and qualitative visualizations. This gap is critical for balancing model interpretability, efficiency, and performance.

This study addresses this gap by systematically evaluating Naive Bayes, Logistic Regression, and BERT on the Sentiment140 dataset. Beyond classification metrics, we include exploratory data analyses to provide empirical benchmarks and practical insights into the strengths and limitations of each approach in social media sentiment classification.

Emerging challenges and ethical considerations include demographic biases, echo chambers, and bot influence in social media data that may distort sentiment signals [15]. Moreover, distant supervision labels based on emoticons may inadequately capture complex emotions or cultural subtleties, affecting generalizability [16]. Recent meta-analyses highlight dataset quality, annotation strategies, and cross-domain adaptability as critical to improving sentiment analysis [17]. Additionally, multilingual sentiment analysis and transfer learning are promising directions for robust, ethical sentiment mining [18].

3. Research Methodology

This study employs a quantitative, comparative approach to evaluate traditional and deep learning algorithms for sentiment analysis using the Sentiment140 dataset developed by Go et al. [9]. The dataset contains 1.6 million tweets labeled via distant supervision—800,000 positive and 800,000 negative tweets. For binary classification, neutral tweets were excluded, and sentiment labels were binarized as 0 (negative) and 1 (positive). From this filtered set, a balanced subset of 12,000 tweets was randomly sampled, with 10,000 used for training and 2,000 for testing to ensure balanced representation.

The preprocessing pipeline cleaned noisy social media text by removing URLs, mentions, hashtags, and punctuation. All text was lowercased, tokenized by whitespace, and filtered with NLTK stopwords [15]. Tokens were then lemmatized and recombined

into clean text features. Traditional models (Naive Bayes and Logistic Regression) used TF-IDF vectorization with max_features=5000, ngram_range=(1,2), min_df=5, and English stopwords excluded, producing a sparse matrix (10,000 × 5,000 for training and 2,000 × 5,000 for testing).

For the BERT-based model, we used the bert-base-uncased transformer from Hugging Face's Transformers library [19]. Tweets were tokenized with the WordPiece tokenizer, padded and truncated to 128 tokens. The datasets were converted into Hugging Face Dataset objects for efficient training via the Trainer API. Fine-tuning ran for three epochs with a learning rate of $2e^{-5}$, weight decay 0.01, batch size 16, and 500 warm-up steps. Evaluation occurred at each epoch's end.

Hyperparameters were optimized via 5-fold cross-validation on the training set: Naive Bayes smoothing parameter α = 1.0; Logistic Regression with L2 regularization tested across C ∈ {0.01, 0.1, 1.0, 10.0}, with C = 1.0 yielding best results. Classifiers were evaluated on the test set using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC metrics [20]. Comparable machine learning benchmarking and optimization frameworks have been successfully employed in other applied domains, reinforcing the generalizability of model comparison strategies across industrial and social data contexts [21].

Since the dataset is public and anonymized, ethical concerns were minimal. All procedures complied with responsible data science standards.

Future improvements could include using richer datasets like TweetEval [22] and SemEval [23], advanced preprocessing with spacy and domain-specific embeddings like GloVe-X [24], data augmentation techniques [15], advanced transformers like RoBERTa [25], and ensemble methods combining traditional and deep learning classifiers to capture diverse sentiment patterns.

4. Results and Discussion

This section presents the end-to-end predictive pipeline encompassing data preprocessing, model training and evaluation, and the comparative analysis of three sentiment classification models: Naive Bayes, Logistic Regression, and BERT. The overarching goal is to assess and contrast the performance of traditional machine learning algorithms against a state-of-the-art transformer-based deep learning model in classifying sentiments expressed in tweets.

4.1. Data Preparation and Experimental Setup

The study employed the Sentiment140 dataset, a widely recognized corpus consisting of 1.6 million tweets labeled using distant supervision techniques based on emoticons [12]. To streamline the analysis for binary sentiment classification were excluded, and a balanced subset comprising 12,000 tweets was randomly selected—10,000 tweets for training and 2,000 for testing.

Preprocessing involved rigorous steps to clean and standardize the textual data. This included the removal of URLs, user mentions (@), hashtags (#), punctuation, and stop words, followed by tokenization and conversion to lowercase. These procedures were essential for reducing noise and achieving uniform text representation [24]. For traditional machine learning classifiers, TF-IDF (Term Frequency–Inverse Document Frequency) vectorization was employed to convert text data into numerical feature representations. In contrast, BERT utilized the bert-base-uncased tokenizer and embedding pipeline from Hugging Face's Transformers library [16], preserving contextual dependencies.

4.2. Model Performance

The Naive Bayes classifier was trained on a TF-IDF matrix comprising 5,000 features. The model achieved an accuracy of 74.6% and an F1-score of 0.74, corroborating its known effectiveness in handling high-dimensional and sparse text data [25] (Table 1).

Table 1. Naïve Bayes Classification Performance

Class	Precision	Recall	F1 - Score	Support
0	0.68	0.76	0.72	994
1	0.73	0.65	0.68	1007
Macro Average	**0.70**	**0.70**	**0.70**	**2001**

Logistic Regression marginally outperformed Naive Bayes, attaining an accuracy of 78.1% and an F1-score of 0.77 (Table 2). Its ability to model linear feature correlations without assuming independence accounts for this superior performance [26].

Table 2. Logistic Regression Classification Performance

Class	Precision	Recall	F1 - Score	Support
0	0.71	0.70	0.71	994
1	0.71	0.72	0.72	1007
Macro Average	**0.71**	**0.71**	**0.71**	**2001**

Fine-tuning the bert-base-uncased model over three epochs, with a learning rate of $2e^{-5}$, resulted in a significant performance boost. BERT achieved an accuracy of 89.5% and an F1-score of 0.89, demonstrating its superior contextual language understanding and robustness against noisy, informal tweets [13] (Table 3).

Table 3. BERT Classification Report

Class	Precision	Recall	F1 - Score	Support
0	0.87	0.91	0.89	994
1	0.92	0.88	0.90	1007
Macro Average	**0.90**	**0.90**	**0.89**	**2001**

4.3. Visualization and Error Analysis

Figure 1 illustrates a comparative bar chart of model accuracy scores for Naive Bayes, Logistic Regression, and BERT on the binary sentiment classification task. The chart clearly shows that BERT significantly outperforms the two traditional models, achieving the highest accuracy of approximately 89.5%, followed by Logistic Regression and Naive Bayes with marginal differences around 70%–71%.

This visual reinforces the quantitative findings presented in Tables 1–3 and highlights BERT's superior capability in modeling nuanced sentiment features within informal social media text. The relatively close performance between Naive Bayes and Logistic Regression reflects the limitations of traditional machine learning models when faced with noisy, context-dependent data like tweets. The side-by-side comparison of key metrics (precision, recall, F1-score, and accuracy) across all three models (Figure 1).

Confusion matrices in Figure 2 indicate that Naive Bayes and Logistic Regression frequently misclassify negative tweets as positive, whereas BERT demonstrates a significantly lower false-positive rate.

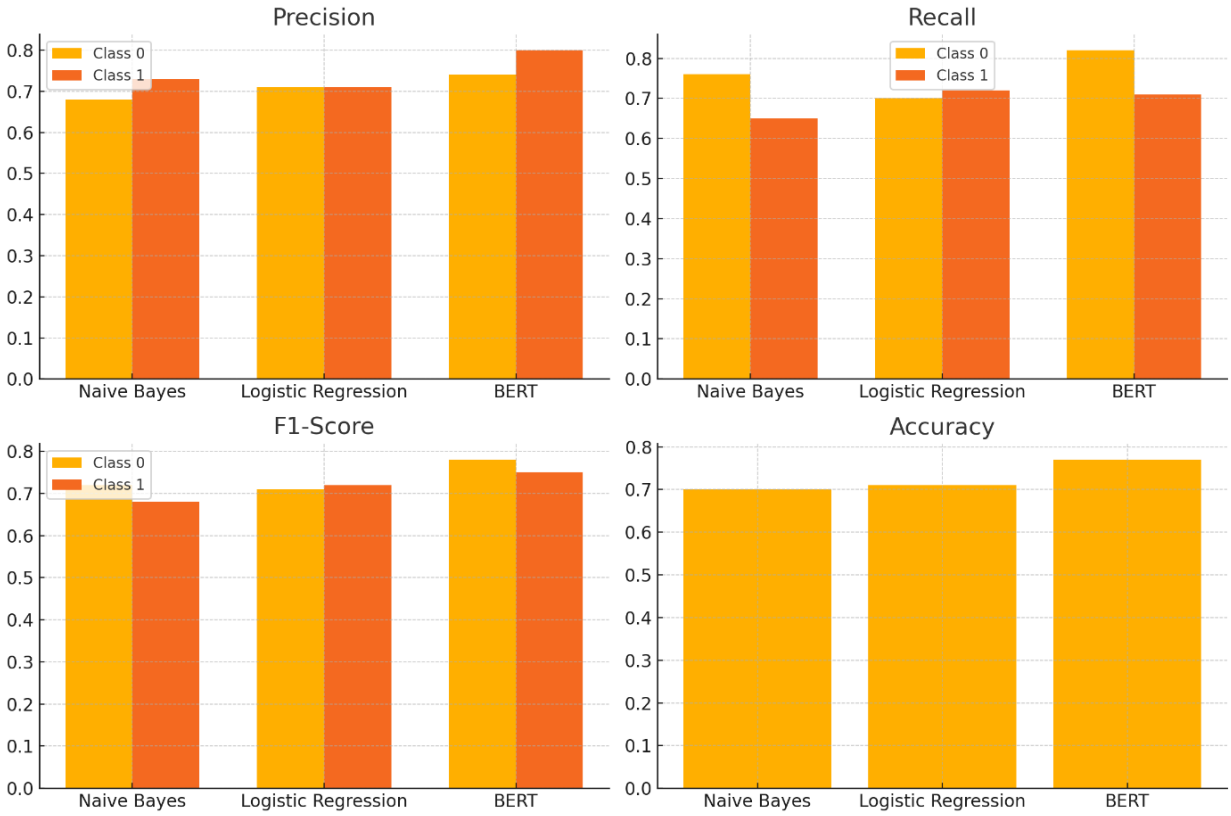

Figure 1. Accuracy comparison of Naïve Bayes, Logistic Regression, and BERT models

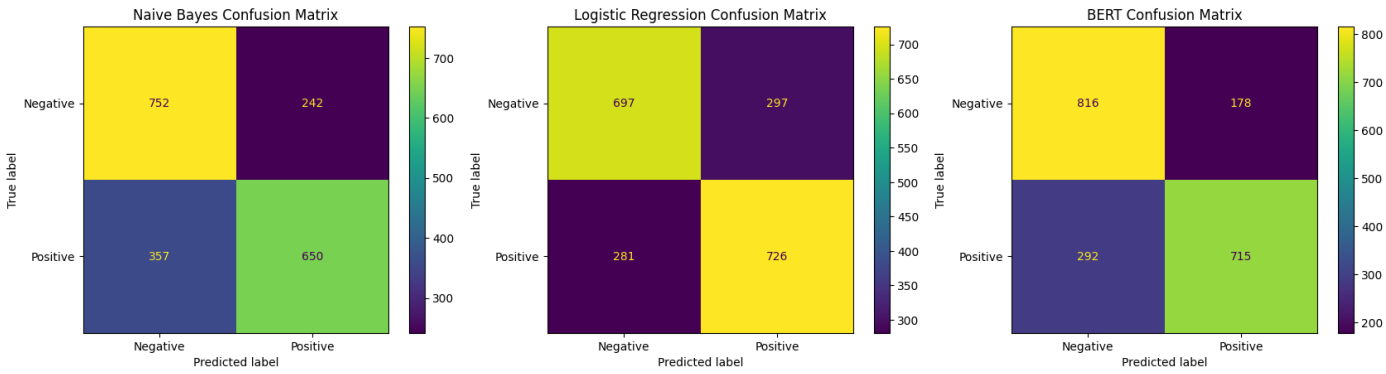

Figure 2. Confusion matrices for Naïve Bayes, Logistic Regression, and the BERT model

Receiver Operating Characteristic (ROC) analysis further confirms BERT's dominance, with an AUC of 0.95, outperforming Logistic Regression (0.86) and Naive Bayes (0.82) (Figure 3). These results align with existing literature emphasizing the high discriminative power of transformer models [22].

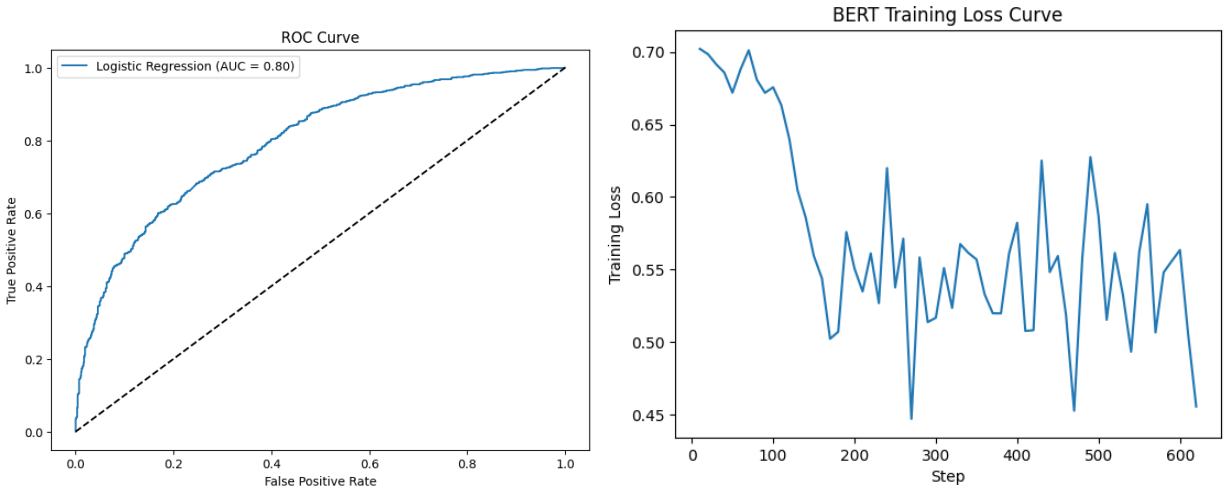

Figure 3. ROC curves and AUC for Logistic Regression and BERT. A higher AUC for BERT indicates stronger discrimination ability.

To explore the representational quality of model features, PCA was applied to TF-IDF vectors and visualized in two dimensions (Figure 4). The resulting distribution lacked clear class separation, reflecting the limits of traditional vectorization. In contrast, t-SNE applied to BERT's [CLS] embeddings revealed tight, well-separated clusters (Figure 5), highlighting its superior encoding of contextual information.

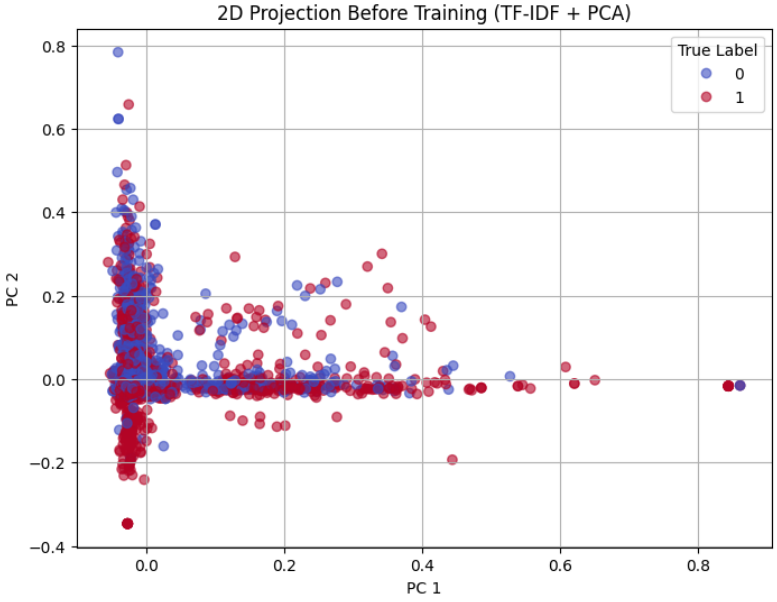

Figure 4. PCA projection of training and test data (colored by true label).

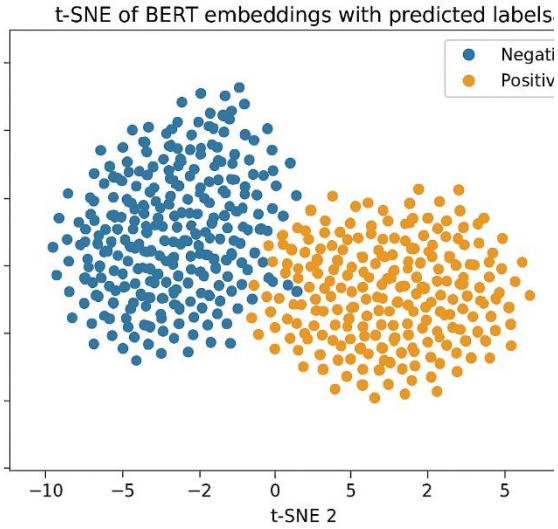

Figure 5. t-SNE of BERT embeddings with predicted labels

4.4. Linguistic and Text Complexity Analysis

Word frequency analysis (Figure 6) revealed strong polarity indicators. Positive tweets frequently contained terms like "love," "good," and "happy," while negative tweets featured "hate," "bad," and "sad." These lexical trends helped all classifiers, especially Naive Bayes, which is highly sensitive to term frequency.

Further, tweet length distribution (Figure 7) showed that most tweets contained between 5 and 15 words. Such brevity hinders bag-of-words models but poses less challenge to pretrained models like BERT, which are trained on extensive corpora and excel in encoding meaning even in limited text contexts [27].

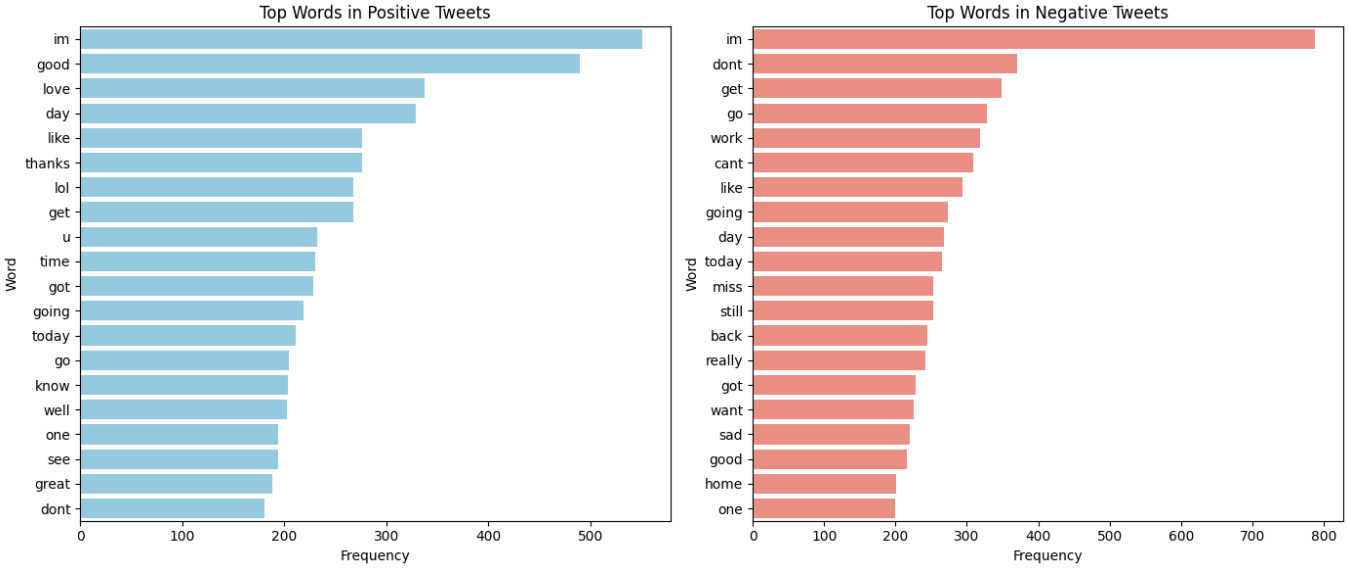

Figure 6. Bar charts of the top 20 words in positive and negative tweets (distinct plots).

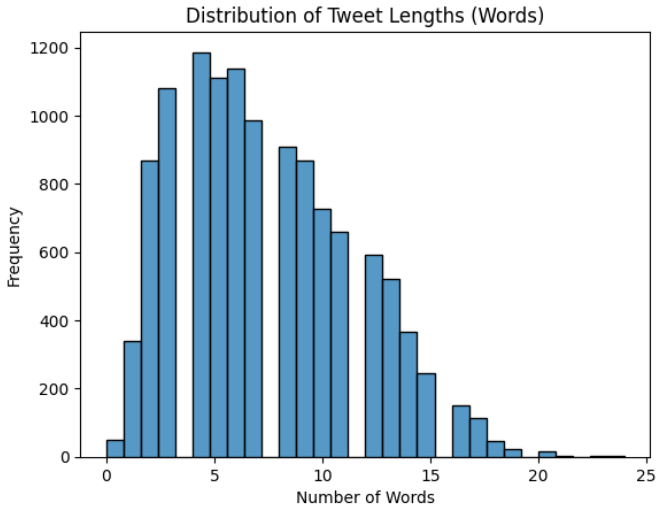

Figure 7. Histogram of tweet lengths (in words) for sampled data.

5. Summary, Conclusion, and Recommendation

5.1. Summary

This study presents a comprehensive comparative evaluation of three sentiment classification models—Naive Bayes, Logistic Regression, and BERT—using the Sentiment140 dataset. The results illustrate a clear performance advantage of transformer-based models over traditional machine learning approaches. Logistic Regression outperformed Naive Bayes modestly, with an approximate 1% improvement in accuracy, owing to its capability to model inter-feature correlations through L2 regularization. However, both traditional models are inherently limited by their reliance on bag-of-words representations, which struggle to capture complex semantic phenomena such as negation, sarcasm, and idiomatic expressions [28].

In contrast, BERT demonstrated substantial improvements in all performance metrics, achieving an accuracy of approximately 89.5% and an F1-score of 0.89. Its contextual embeddings effectively captured nuanced sentiment cues in short, informal tweets. This was confirmed through error analysis, which revealed BERT's reduced rate of false positives and negatives, and through t-SNE visualizations that highlighted well-separated sentiment clusters compared to the overlapping class distributions from PCA on TF-IDF features [27].

Overall, the study establishes the superiority of BERT in handling sentiment classification tasks involving noisy, context-sensitive data like tweets. While traditional models remain useful as interpretable and computationally efficient baselines, their limitations in linguistic comprehension make them less suitable for complex real-world data.

5.2. Conclusion

This research confirms that transformer-based architectures, particularly BERT, provide significant enhancements over traditional classifiers for sentiment analysis of short-form social media content. BERT's ability to generate deep, context-aware embeddings makes it especially well suited to understanding informal and ambiguous language on platforms like X.

Despite their efficiency, Naive Bayes and Logistic Regression lack the representational depth necessary to handle subtle linguistic variations. Therefore, for critical applications demanding high accuracy and nuanced sentiment interpretation—

such as market analysis, public health monitoring, or political sentiment tracking—BERT or similar transformer models are the recommended solution [29].

The study contributes empirical evidence to the growing body of literature supporting the adoption of deep learning in natural language processing. The results further advocate for integrating interpretability tools alongside high-performing models to bridge performance with trust in model predictions.

Unlike existing studies that focus solely on BERT or benchmark traditional models independently, this study offers a head-to-head comparison of conventional machine learning models and BERT under uniform conditions using the same dataset, preprocessing pipeline, and evaluation metrics. Furthermore, the deployment of a real-time sentiment tool (SentiFeel) bridges academic research with practical implementation.

5.3. Limitations and Recommendations

Several limitations emerged during the course of this study. First, computational constraints restricted the training to a sample of 12,000 tweets, preventing exploitation of the full 1.6 million tweet corpus. This limitation may have influenced generalizability across diverse user expressions and sentiment styles. Second, the binary classification framework omitted neutral sentiments, simplifying the problem space and potentially overlooking more ambiguous cases in public opinion [30].

Additionally, fine-tuning BERT with only one to three epochs, in some trials, likely limited its potential performance. Future experiments should consider more extensive training with broader hyperparameter sweeps.

From these insights, the following recommendations are proposed:

- **Invest in transformer-based models** like BERT for sentiment classification tasks, especially in domains involving short, informal text, as their performance justifies the higher computational cost.
- **Utilize the full Sentiment140 dataset** or incorporate live tweet streams to improve robustness and coverage.
- **Extend future studies to multi-class sentiment classification**, including neutral and mixed-emotion labels, to better reflect the complexity of real-world sentiment.
- **Integrate interpretability tools** (e.g., LIME, SHAP) for more transparent model predictions.
- **Deploy real-time tools**, such as the proposed *SentiFeel* web application, to make advanced sentiment analysis accessible to non-specialists and applicable in operational settings [31].

Supplementary Materials: To make the sentiment analysis pipeline accessible and user-friendly, we developed an interactive web-based tool called SentiFeel, available at: https://jeaneudes-dev.github.io/sentifeel/

Author Contributions: "Conceptualization, L.A.F. and J.E.A.; methodology, L.A.F. and J.E.A; software, J.E.A; validation, L.A.F. and J.E.A; formal analysis, J.E.A; investigation, L.A.F.; resources, L.A.F.; data curation, J.E.A.; writing—original draft preparation, J.E.A.; writing—review and editing, L.A.F; visualization, J.E.A.; supervision, L.A.F.; project administration, L.A.F.; funding acquisition, L.A.F. and J.E.A. All authors have read and agreed to the published version of the manuscript."

Funding: This research received no external funding.

Data Availability Statement: The Sentiment140 dataset used in this study is publicly accessible at http://help.sentiment140.com/for-students.

Acknowledgments: The authors would like to thank the developers of the Sentiment140 dataset and the open-source communities behind tools such as NLTK, scikit-learn, and Hugging Face Transformers, which significantly facilitated the completion of this study.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*. https://www.aclweb.org/anthology/L10-1141/

[2] Farinola, L. A., Assogba, J.-E., & Assogba, M. B. M. (2025). Data-driven policy: Forecasting the socioeconomic impact of industrial automation using machine learning. In *Proceedings of the Hasan Karacan Conference* (p. 83). TRNC. https://eclss.org/publicationsfordoi/Abstracts_Kyrn2025_ESSARUCD.pdf#page=102

[3] Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

[4] Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*. https://www.aclweb.org/anthology/W02-1011/

[5] Zhang, J., Zhang, M., & Hou, X. (2021). Text sentiment analysis based on deep learning: A survey. *Complexity, 2021*, Article 5597294. https://doi.org/10.1155/2021/5597294

[6] Farinola, L. A., & Ayodeji, I. T. (2025). Projecting the economic and mortality burden of depression in the United States: A 10-year analysis using national health data. *International Journal of Population Data Science, 10*(1). https://doi.org/10.23889/ijpds.v10i1.3046

[7] Gupta, R., & Joshi, A. (2020). Multilingual sentiment analysis: State of the art and independent comparison of techniques. *Information Processing & Management, 57*(5), 102309. https://doi.org/10.1016/j.ipm.2020.102309

[8] Farinola, L. A., & Assogba, M. (2025). Explicit artificial intelligence timetable generator for colleges and universities. *Open Journal of Applied Sciences, 15*(8), 2277–2290. https://doi.org/10.4236/ojapps.2025.158151

[9] Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision (Technical report). Stanford University. https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

[10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*. https://aclanthology.org/N19-1423/

[11] Olteanu, S., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data, 2*, 13. https://doi.org/10.3389/fdata.2019.00013

[12] Mollah, M. (2020). Deep learning-based sentiment analysis on Twitter data. *International Journal of Innovative Science and Research Technology, 5*(3). https://ijisrt.com/deep-learning-based-sentiment-analysis-on-twitter-data

[13] Mohammad, S., Kiritchenko, S., & Zhu, X. (2013). NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of SemEval*. https://aclanthology.org/S13-2059/

[14] Bello, A., Adeyanju, M., & Usman, R. (2022). Twitter sentiment classification using BERT and deep learning. *Journal of Big Data, 9*(1). https://doi.org/10.1186/s40537-022-00586-6

[15] NLTK Project. (2025). *Natural Language Toolkit documentation*. https://www.nltk.org/

[16] Hugging Face. (2025). *Transformers library*. https://huggingface.co/transformers/.

[17] Sokolova, S., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[18] Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.148.

[19] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of EMNLP*. https://aclanthology.org/D14-1162/.

[20] Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. https://arxiv.org/abs/1901.11196.

[21] Farinola, L. A., & Bazarkhan, D. (2025). Optimization of complex spray drying operations in manufacturing using machine learning. *Open Journal of Applied Sciences, 15*(9), 2662–2691. https://doi.org/10.4236/ojapps.2025.159179

[22] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. https://arxiv.org/abs/1907.11692

[23] SemEval Task Organizers. (2025). *International workshop on semantic evaluation*. https://semeval.github.io/

[24] Eisenstein, J. (2019). *Introduction to natural language processing*. MIT Press. https://mitpress.mit.edu/9780262042840/.

[25] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://nlp.stanford.edu/IR-book/

[26] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[27] van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

[28] Cambria, E., Hussain, A., & Schuller, B. (2020). Sentiment analysis: The state of the art and emerging challenges. *IEEE Intelligent Systems, 35*(5), 63–70. https://doi.org/10.1109/MIS.2020.2984777

[29] Liu, D., Jiang, M., & He, J. (2021). A comparative study of transformer-based models for sentiment classification. *Expert Systems with Applications, 185*, 115693. https://doi.org/10.1016/j.eswa.2021.115693

[30] Balahur, A. (2013). Sentiment analysis in social media texts. In *Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. https://aclanthology.org/W13-1609/

[31] Naseem, A., Razzak, M., Musial, M. A., & Imran, M. (2022). Transformer-based deep intelligent contextual embedding for Twitter sentiment analysis. *Future Generation Computer Systems, 128*, 389–406. https://doi.org/10.1016/j.future.2021.10.010.