CITE

Journal of Analytical and Applied Computer Sciences
Volume: 2, Issue: 1, 2026 (January - June), Pages: 37-46

Research Article

Comparative Evaluation of OpenAI and Transformer Models for Twitter Sentiment Classification

Babangida Pada Solomon[1*], Abdullahi Musa Yola[2], Nura Muhammad Sani[3]

[1]Department of Computer Science, Federal Polytechnic Kaltungo, Gombe State Nigeria. babangidasolomon3@gmail.com
[2]Department of Computer Science, Federal University of Kashere, Gombe State Nigeria. amyola@ulk.ac.rw
[3]Faculty of Computer Science, Universitas Dian Nuswontoro, Semerang, Indonesia. nuramuhammad07061@gmail.com

[*]**Correspondence**
Babangida Pada Solomon
babangidasolomon3@gmail.com

Received: 10 January 2026 / Accepted: 16 April 2026 / Published: 30 June 2026

This paper presents a comparative analysis of five advanced Natural Language Processing (NLP) models: GPT-4, GPT-3.5, BERT, RoBERTa, and DistilBERT, specifically trained and evaluated for sentiment classification on Twitter data. The study emphasizes the development of these models and assesses their performance using standard metrics, including accuracy, precision, recall, and F1-score. The results indicate that BERT achieved the highest F1-score of 0.8% with a balanced focus on accuracy and efficiency, while DistilBERT delivered a competitive accuracy of 0.86% with significantly reduced inference times. Although GPT-based models excelled in contextual understanding, they exhibited higher latency. These findings highlight the trade-off between predictive accuracy and computational efficiency when deploying AI models for real-time sentiment analysis applications.

Keywords: *Sentiment Analysis, Natural Language Processing, Transformer Models, OpenAI, GPT, Twitter*

Introduction

Social media networks, especially Twitter, produce excessive amounts of user-generated content daily, making such a platform a rich corpus in terms of studying public opinion (Ak & Paroubek, 2010). Sentiment analysis or also called opinion mining, is one of the key areas in Natural Language Processing (NLP) that enables the isolation of subjective data in a written form (Tan & Lim, 2023). The recent advancements in the transformer-based language models, such as BERT, RoBERTa, and XLNet, have strongly increased the accuracy and stability of the sentiment classification problems with more efficient use of the context, polysemantic expressions, and the complexity of the natural language (Lin & Liu, 2024; Widyananda & Fauzi, 2025). However, sentiment analysis, as carried out in the Twitter context, is faced with several unique challenges. Tweets are usually short, informal and highly laden with slang, abbreviations, and emojis, which adds to the lack of semantics. Moreover, sarcasm and irony, which are widespread in other studies (Wiguna et al., 2021; Yunitasari & Sari, 2019), make the detection of the sentiment more difficult. Such phenomena will require specialised methods of recognition, beyond the traditional methods of text processing. Social media platforms such as Twitter serve as powerful sources of public opinion, where millions of short, informal posts provide insights into user sentiment. Sentiment analysis is an essential task in Natural Language Processing (NLP), which aims to determine whether a piece of text expresses a positive, negative, or neutral opinion.Recent advancements in transformer-based models, such as BERT, RoBERTa, and DistilBERT, as well as generative AI models like GPT-3.5 and GPT-4, have significantly improved the performance of sentiment classification systems. However, systematic comparisons between these models in practical deployment contexts remain limited.This study focuses on developing and evaluating five AI

models trained on Twitter data, comparing their sentiment classification performance using performance evaluation metrics such as the confusion matrix, accuracy, precision, recall, and F1-score.

Previous studies (Semary et at., 2023) have shown that transformer models like BERT and RoBERTa outperform traditional machine learning algorithms such as Naïve Bayes and SVM in sentiment classification. Liu (2022) attempted to tackle the issue of implementing the transformer model in the resource-constrained setting through the exploration of the knowledge distillation methods in the sentiment analysis domain. The study established that student models trained on the MiniLM and TinyBERT distilled models made up to 91.6 running of SST 2 and 73.5 running of SemEval datasets only slightly less than the teacher model. Liu also highlighted the efficiency of the distillation frameworks with attention-based internal behaviour and introduced a new Two-Step Distillation in reducing overfitting. This approach contrasts transformer layer modelling with prediction layer learning, resulting in stronger student models. However, Liu has observed that simply scaling down a model does not reduce overfitting, and distillation is not as effective on data of all sizes and models of all complexity. In a related study by Gowda et al. (2025), the performance of various transformer architectures—specifically BERT, RoBERTa, and GPT—was examined in sentiment analysis tasks across four distinct domains: e-commerce, healthcare, finance, and social media. The findings indicated that RoBERTa outperformed the others, achieving an impressive accuracy of up to 96 per cent on domain-specific tasks and demonstrating strong resilience to noisy and informal data. The paper highlighted the critical roles of self-attention and positional encoding in attaining contextual knowledge while also addressing the trade-offs between computational efficiency and performance. However, it noted significant drawbacks, including high resource demands and limited interpretability, which underscored the need for further research into lightweight models and explainable AI methodologies. Similarly, Miranda et al. (2023) examine the evolution of sentiments expressed in Spanish-language tweets related to the pandemic through the fine-tuning of the BERT architecture. The authors demonstrated that the application of BERT can achieve an accuracy of up to 89 percent in a three-class sentiment classification (positive, negative, and neutral). Their analysis of over six million tweets collected in Colombia during 2020 and 2021 yielded F1-scores of 0.911, 0.907, and 0.853, respectively. The study identified key emotional patterns during the COVID-19 pandemic, notably prevalent feelings of anger and fear, and emphasized the necessity to adapt models to the unique characteristics of the Spanish language.

In a related effort, Perikos and Diamantopoulos (2024) proposed an explainable aspect-based sentiment analysis (ABSA) framework that fine-tunes various transformer models—including BERT, RoBERTa, DistilBERT, and XLNet—on datasets such as MAMS, SemEval, and Naver. RoBERTa-base emerged as the most effective model, achieving accuracy rates of 89.16 and 97.62 on SemEval and MAMS/Naver, respectively. To mitigate the challenges posed by the black-box nature of transformers, the authors employed five explainability methods: LIME, SHAP, attention visualization, integrated gradients, and Grad-CAM. Their findings highlighted the influence of specific words on aspect-level predictions, underscoring the importance of explainability in enhancing the transparency and reliability of sentiment analysis models. Shafikuzzaman et al. (2025) evaluated the performance of GPT-4 and other pretrained language models for sentiment classification in the software engineering (SE) domain using zero-shot and few-shot settings. Their comprehensive study revealed that GPT-4, even without task-specific fine-tuning, achieved strong results, with an F1-score up to 76.6% on multi-class sentiment classification. It outperformed many traditional models, including some that were fine-tuned on domain-specific data. GPT-4's ability to understand contextual subtleties, such as sarcasm and domain-specific jargon, was attributed to its extensive pretraining on diverse corpora. However, the study also highlighted a critical limitation: the substantial computational resources required by models like GPT-4 make them less feasible for real-time or resource-constrained applications. However, Mathebula et al. (2024) proposed LFEAR, a sentiment analysis framework that combines retrieval-augmented generation (RAG) with fine-tuned auto-regressive language models like GPT-4o Mini and Llama-3. Applied to financial reviews, LFEAR achieved high performance (98.45% precision, 93.85% correctness) and effectively handled nuanced sentiment. However, its reliance on high computational resources and retrieval quality limits real-time use. Future work will enhance scalability and cross-domain adaptability. Similarly, Chumakov et al. (2023) also presented an aspect sentiment triplets extraction (ASTE) classifier, which uses few-shot learning and fine-tuning GPT-3 and RuGPT-3, utilising a generative architecture founded on GPT. The approach gave F1 results of 0.74 on English and 0.52 on Russian, which is 12-percent ahead of state-of-the-art (SOTA) models on both English and Russian metrics. The major weaknesses were connected with reliance on data refinement by hand when it comes to the Russian language and difficulties with the consideration of the format. The upcoming studies aim at extending multilingual assistance and reducing the use of manual annotation. Prova (2025) suggested a multilingual emotion-sensitive recommendation system that employs the GPT-based embeddings, along with the meta-ensemble system with XLM-R, mBERT and BiGRU to categorise the emotions in e-commerce review in their respective cultures. The framework combines content-based filtering, collaborative filtering, and emotion indicators which attain 11.7 per cent recommendation accuracy and 45 per cent conversion rate improvement rates compared to traditional frameworks. Weaknesses are the inability to detect sarcasm and cold-start problems whenever using it by a new user. Future studies will include multimodal data and develop cultural adaptability on low-resource languages. However, Fei et al. (2023) investigated the importance of chain-of-thought (CoT) prompting in improving

implicit sentiment analysis (ISA) by breaking down the reasoning process into three phases (1) identification of implicit aspects, (2) inference with latent opinions, and (3) identification of sentiment polarity. Their Three-hop Reasoning (ThoR) architecture, applied to large language models like Flan-T5 and GPT-3, had seen major performance improvements over traditional models, with Flan-T5 (11B) also improving supervised performance by 6 - unofficial F1 over the baseline and GPT-3 (175B) improving zero-shot performance by more than half. The success of the strategy depends on the size of models with smaller ones only providing slight improvements. Future research would make ThoR more efficient and apply it to other natural language processing related problems, which involve multi-step reasoning. In their sentiment analysis experiment, Kang and Choi (2025) matched FinBERT (another model that is finetuned on the case of financial news) with GPT4o which is optimised using prompt engineering on the case of business, health and technology industries. FinBERT scored 10% lower than GPT-4o which had a peak score of 0.55 accuracy (technology research). Limiting resources were constrained dataset and imbalance of the labels. The research highlighted that GPT-4o was better than the FinBERT in contextual sentiment analysis, whereas FinBERT excelled in financial texts specific to a domain. Future research with market volatility data and testing more large language models will be done as planned. The most recent developments in sentiment analysis have involved the implementation of large language models (LLMs) and chain-of-thought (CoT) prompting methods in order to improve effectiveness in a variety of fields. Fei et al. (2023) made the proposal of Three-hop Reasoning (THOR), which subdivides the implicit sentiment analysis into three stages, namely, to identify implicit features, infer latent opinions, and to estimate the sentiment polarity. Their models showed the most gains with the Flan-T5 (11B) reversing the situation by increasing the F1-score by 6 percent of the traditional models in the supervision and GPT-3 (175B) by more than 50 percent in the zero-shot setting. Kang and Choi (2025) compared GPT-4o optimized by prompt engineering to FinBERT in a range of industries in the financial sector. Their results showed that GPT-4o was as much as 10 percent more accurate than FinBERT when used in contextual sentiment analysis, leading to the conclusion that reasoning-based and non-reasoning models may be effective in zero-shot financial sentiment analysis. In their work, they discovered that GPT-4o when not prompted with CoT correlated more with human-labeled sentiment, which is an indicator that reasoning can bring in overthinking and result in suboptimal predictions in a financial scenario. Qin et al. (2023) suggested that cross-lingual prompting (CLP) can be used to enhance the reasoning of zero-shot CoT across languages. They also attained the state-of-the-art performance in their methodology, involving cross-lingual alignment and task-specific solver prompting, which proves that CLP can be used to improve the process of multilingual sentiment analysis. Turpin et al. (2023) was however, careful to warn that CoT explanations can be misleading in a systematic way. According to their study, the CoT reasoning may also be biased to give plausible but erroneous explanations which is why better mechanisms should be employed to guarantee fidelity of the model generated reasoning. Wu et al. (2025) proposed an optimized sentiment classification model for e-commerce applications, but their work lacked a direct comparison between proprietary large language models (LLMs) such as GPT-4 and open-source transformer models like BERT or RoBERTa. Similarly, Reddy et al. (2025) examined model ensembles for social media sentiment detection, yet they did not evaluate deployment trade-offs, including resource consumption, processing overhead, and scalability in web-based environments.

Methodology

Here we explain the methods and procedures used to conduct our research on Comparative Evaluation of OpenAI and Transformer Models for Twitter Sentiment Classification, as shown in Figure1.

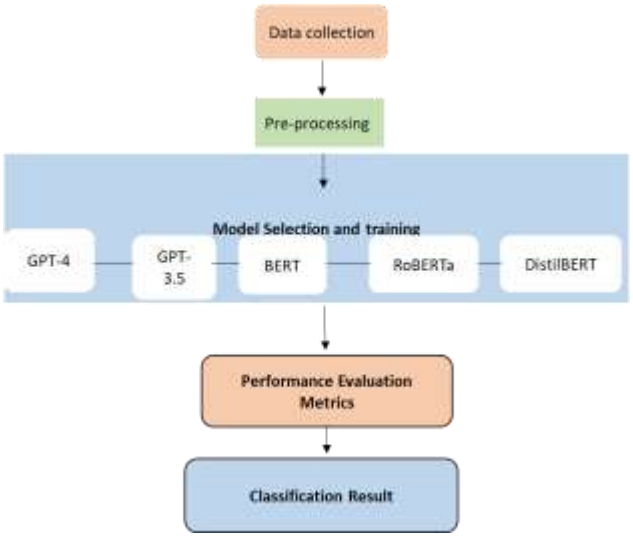


Figure 1. Architecture of the Research Methodology

Data collection

The dataset used for this study is a Twitter-based sentiment analysis dataset consisting of 162,980 records. It classifies sentiments into three categories: positive (1.0), neutral (0.0), and negative (-1.0). Among these, 72,250 tweets are positive, 55,213 are neutral, and 35,510 are negative. This indicates a slightly skewed dataset with a slight tilt towards positive sentiments. The dataset is secondary data obtained from the Kaggle database

Data Pre-processing

Preprocessing was an important stage of data preparation for sentiment classification. We used Cleaning, Lowercasing, Tokenization, Padding and Truncation, and Label Encoding techniques to preprocessor dataset. These steps made the dataset clean, consistent, and compatible with the input structure that is needed by transformer-based models.

Model Selection

Different Natural Language Processing models were chosen for conducting this study based on their common adoption, excellent performance in sentiment analysis studies in the past, and accessibility through open source or API access. These are models of proprietary and open-source transformer architectures, and theypresenta balanced picture of performance-based deployment trade-offs.

Performance Evaluation Metrics

Different categories of evaluation metrics were also utilized in an attempt to have a fair and comprehensive comparison of the sentiment analysis models. These are the metrics that were taken to find out the predictive precision of each model.

Classification Performance Metrics

The following standard metrics were usedin the evaluation of the efficiency of each model to classify sentiment:

Accuracy

Computes the ratio of correct prediction as against the total number of predictions.

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

Precision

Estimates the proportion of actual positive predictions that are obtained from all of the positive predictions made by the model.

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

Recall

Gives the ratio of the positive cases that the model could identify correctly.

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

F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-Score: Provides aharmonic mean of precision and recall, trying to balance between the metrics within a holistic score.

Results and Discussion

In this section, we present the experimental results of our comparative analysis of OpenAI and Transformer models for Twitter sentiment classification.

Understanding the dataset

It is an important fundamental step in the data analysis process, as it helps to providecontext and understanding of the data before more complex analyses are performed. Italso enables us to identify potential data quality issues, explore relationships between variables, and generate hypotheses for further investigation. The dataset that was used for this study is a Twitter-based sentiment analysis dataset that consists a total of 162,980 records. It has two major columns. clean text that will contain the preprocessed content of tweets and category that will indicate the sentiment label of every tweet. There are three categories for the sentiment classes, which are: positive (1.0), neutral (0.0), and negative (-1.0), with 72,250 tweets taking the positive category, 55,213 as neutral, and 35,510 as negative. This points to a barely skewed dataset, with some more positive sentiments as compared to negative ones.

1. Sentiment Label Distribution

Figure 4.1 depicts how the different levels of sentiment occur in the dataset. The graph reveals that most of the tweets are labelled with positive emotions, then followed by those with neutral sentiment. The fewest tweets express negative feelings. As a result, there are relatively more positive tweets than neutral or negative ones in the dataset. Nonetheless, the dataset still covers each sentiment category evenly enough for training a reliable sentiment classifier. The chart was generated using **(seaborn Count Plot)** function and displays the number of tweets in each sentiment category along the y-axis.

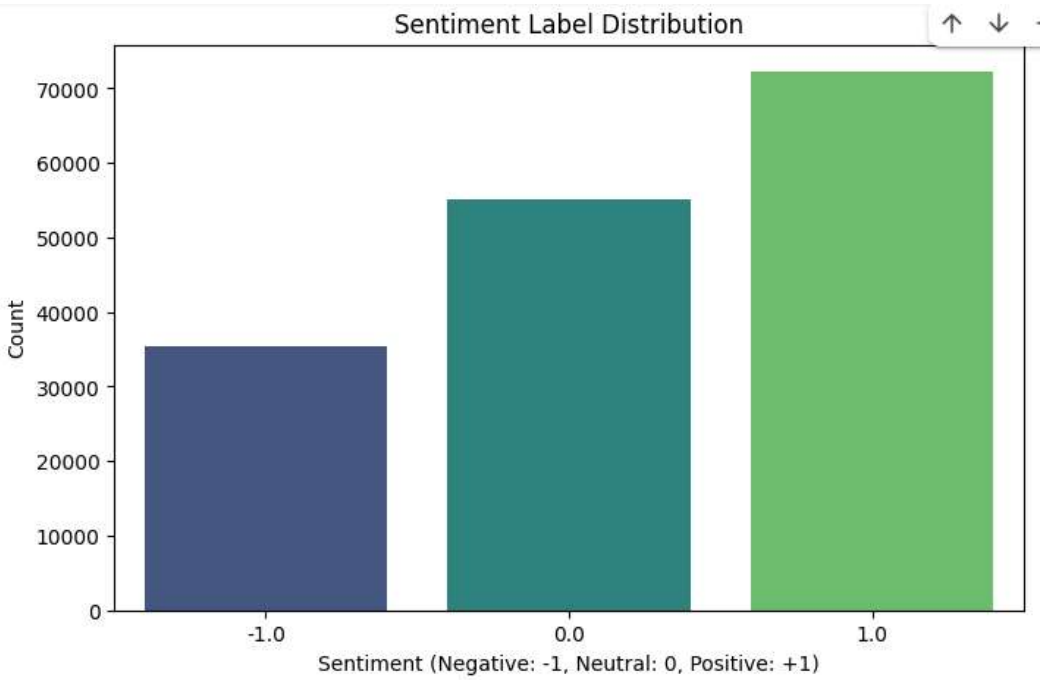


Figure 1. Sentiment Label Distribution

2. Text Length Analysis

Figure 2 bellow, present information about the distribution of tweet lengths in the dataset, according to the number of characters used. 162,980 tweets make up the dataset, averaging about 124 characters in length each. It shows that there is some variation in the length of tweets, with values tending to cluster around a specific boundary. The shortest tweet consists of no characters and the longest tweet is up to 274 characters. The lengths range from 0 to 274 characters. This shows that the majority of tweets are pretty short, with half of them falling between 66 and 183 characters in length. These statistics are significant for setting appropriate sequence lengths during tokenization and how to process both short and long tweets during the classification task.

```
--- Text Length Statistics ---
count    162980.000000
mean        124.173445
std          67.925237
min           0.000000
25%          66.000000
50%         114.000000
75%         183.000000
max         274.000000
Name: text_length, dtype: float64
```

Figure 2. Descriptive Statistics of Tweet Text Lengths in the Dataset

3. Distribution of Tweet Length

Figure 3 below, is a graph illustrates how tweet lengths vary across the dataset, indicating which length ranges are the most common. Tweets are sorted by the number of characters each one contains, from the shortest to the longest of 280 characters and the frequency with which different tweet lengths occur is displayed on the y-axis. There are two distinct peaks in the distribution: The two modal points at 60–70 characters and 220–240 characters reflect the tendency of many users to favour either brief or lengthier tweets, respectively. The distribution suggests that the majority of tweets are either brief or full-length while relatively few fall in between the two extremes. The blue curve (Kernel Density Estimate) clearly illustrates the distribution of tweet lengths. Understanding of this distribution is essential for activities such as text pre-processing and model training, as it guides the choice of input length management techniques prior to applying the models to categorise tweets.

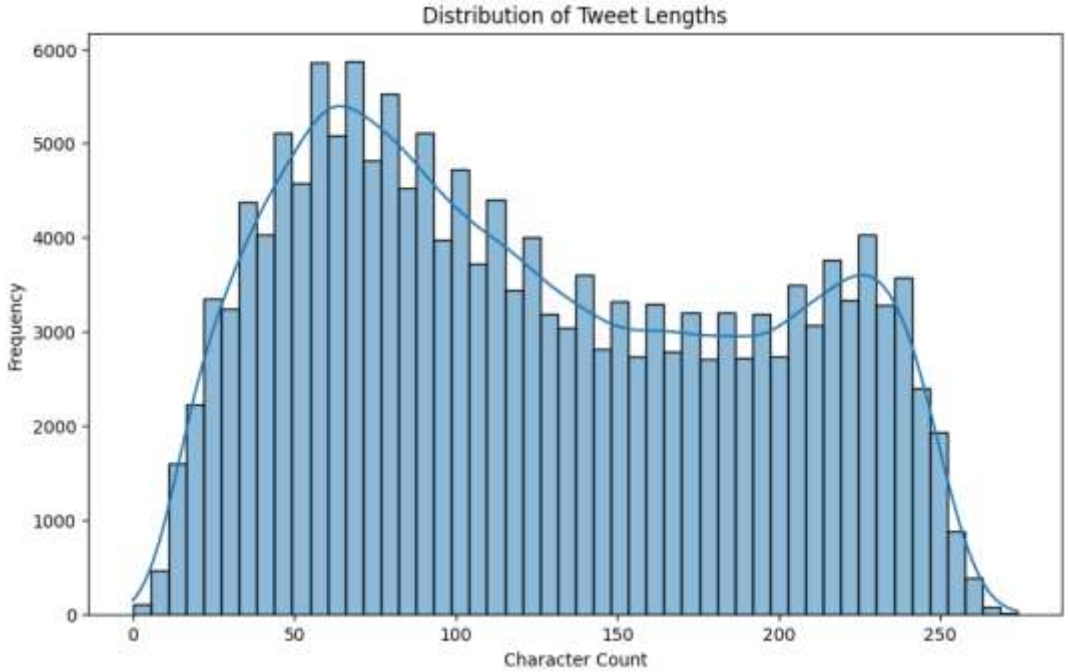


Figure 3. Tweet Length Distribution Based on Character Count

Confusion Matrix

Confusion metrics used to evaluate the performance of the models, it gives a detailed understanding of the accuracy and errors occurred. From the confusion matrix obtained, various performance metrics such as accuracy, precision, recall, and F1 score were computed. Figures 1, 2, 3, 3 and 4 below are performance evaluation results obtained from BERT, RoBERTa, DistilBERT, GPT-4, and GPT 3.5 Models**.**

In detail, the BERT model was able to distinguish 94 tweets as negative but mistakenly labelled 9 neutral and 22 positive tweets as negative. It was able to classify 187 tweets accurately as neutral while incorrectly labelling 12 negative and 15 positive tweets as neutral. The model correctly identified 223 tweets with positive sentiment; nevertheless, it also mistook 17 tweets classified as negative and 21 categorised as neutral as positive. Similarly, the RoBERTa model correctly identified 71 tweets as negative, but it incorrectly classified 22 neutral and 30 positive tweets as negative. Twenty-nine (29) tweets were identified as negative by the model, despite being neutral. It also correctly identified 217 positive tweets, but labelled 27 negative tweets and 16 neutral tweets as positive. The DistilBERT model was correct in determining 186 tweets as neutral, but incorrectly considered 10 negative tweets and 21 positive tweets as neutral. It also predicted that 225 tweets were positive; however, it misclassified 19 negative and 16 neutral tweets as positive. The GPT-4 model can handle identifying tweets as negative, neutral or positive in their sentiment. Out of the negative tweets, the model classified 38 correctly, but it classified 47 as neutral and 38 as positive, which shows it's not always precise at detecting negativity. GPT-4 identified the vast majority of neutral tweets correctly, seeing 199 as neutral and just 18 as falsely positive. Twitter conversations marked as positive were correctly predicted 207 times and 46 were classified incorrectly as neutral, and another 7 were predicted as negative.

The GPT-3.5 model categorises tweets into negative, neutral and positive groups. The model correctly found 94 negative tweets, but labelled 13 of them as neutral and 16 of them as positive. Out of the neutral tweets, the model identified 166 as neutral, but mislabelled 35 as negative and 16 as positive. For the positive tweets, 194 were identified correctly, but 45 were thought to be negative, and the remaining 21 for neutral. GPT-3.5 performs very well in finding positive emotions, but it is less reliable than GPT-4 when it comes to telling apart negative and neutral sentiments

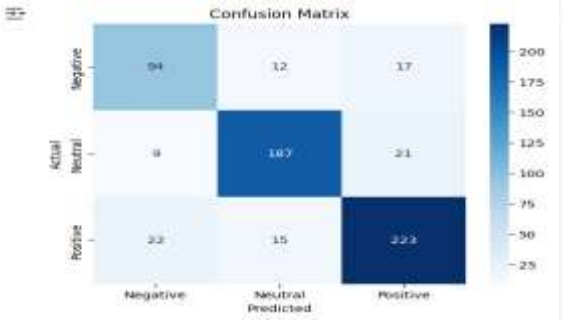


Figure 4. Confusion Matrix for BERT Model.

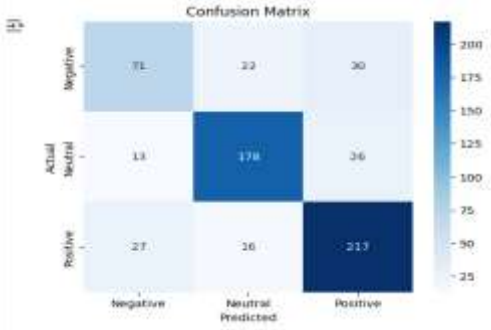
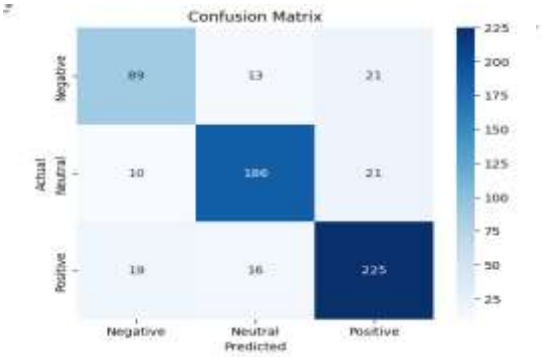


Figure 5. Confusion Matrix for RoBERTa Model



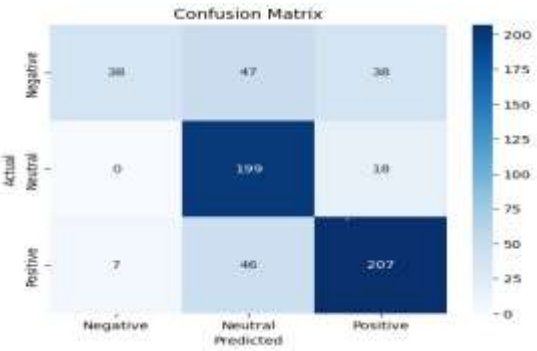


Figure 6. Confusion Matrix for DistilBERT Model. Figure 7 Confusion Matrix for GPT-4 Model

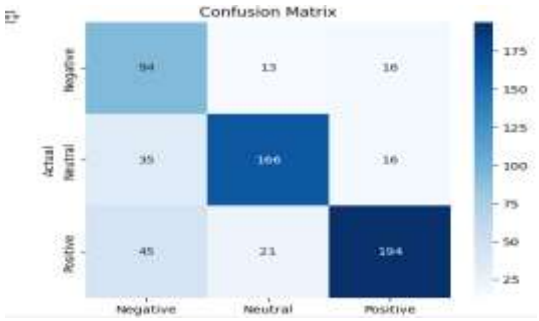


Figure 8. Confusion Matrix for GPT-3.5 Model

Performance Comparison of Transformer-Based Models

The classification performance metrics revealed significant variations across the five models. Table 1 summarizes the overall accuracy, precision, recall, and F1-scores

Table 1. Performance Comparison of Transformer-Based Models

Model	Accuracy	Precision	Recall	F1-Score
BERT	0.84	0.87	0.86	0.87
DistilBERT	0.83	0.87	0.86	0.86
RoBERTa	0.78	0.82	0.82	0.82
GPT-3.5	0.76	0.83	0.76	0.80
GPT-4	0.74	0.68	0.92	0.78

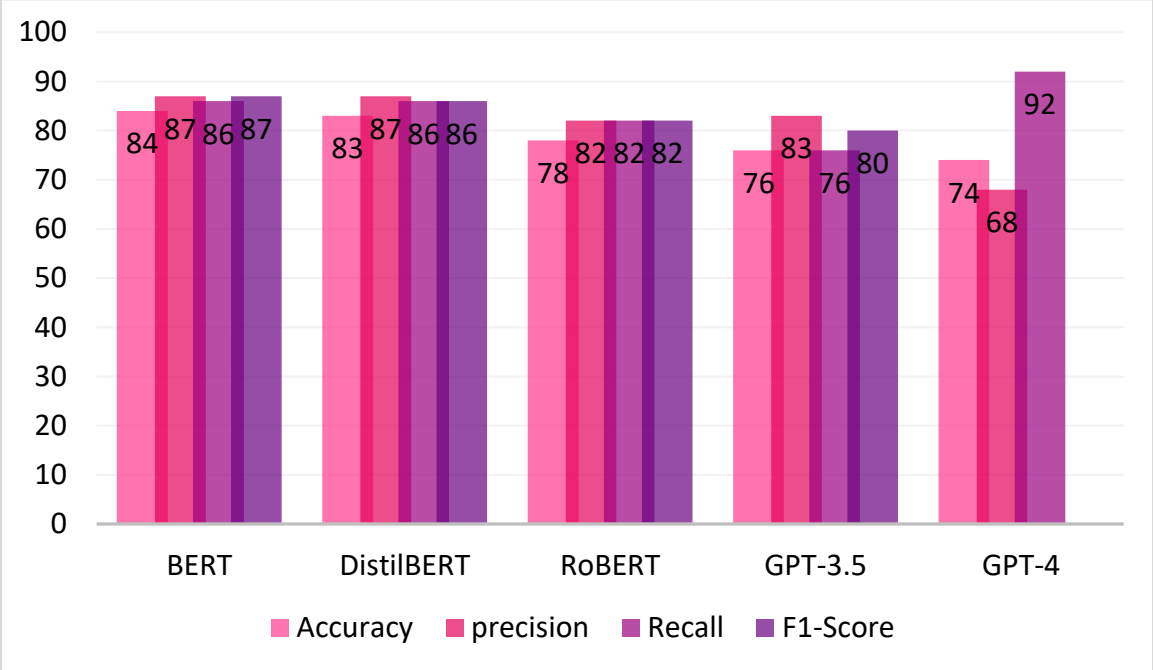


Figure 4. Performance Comparison of Transformer-Based Models

BERT model is provingto be the best performance model among all other models, it achievedthe highest accuracy of 84% and F1-score of 87%, followed by model, making it a more efficient substitute with almost equal effectiveness. The RoBERTa performs moderately, with all resultsaround 82%, showing it is less effective than BERT-based models. GPT-3.5 shows a reasonable precision 83% but a lower recall of76%, leading to an F1-score of 80%. GPT-4, while achieving the highest recall 92% has the lowest precision 68%, suggesting it identifies most relevant cases but also produces more false positives, resulting in the lowest F1-score among the models 78%. It wasdiscovered that BERT and DistilBERT got the most balanced and effective performance for this classification task, while GPT-based models may require further fine-tuning to reach similar levels.

In the present research, we compare various models involving the use of transformers, namely: GPT-4, GPT-3.5, BERT, RoBERTa, and DistilBERT, in the context of sentiment analysis of Twitter data, which is often informal and contains slang, abbreviations, and emojis. Contrary to the existing literature, which aims to maximize the performance of a single model, our method compares the performance of the models based on the various performance metrics, including accuracy, precision, recall, and F1-score. We find that GPT-based models demonstrate high contextual understanding but poor precision, so they are not appropriate to be use in real-time. Conversely, it is clear that BERT and DistilBERT are more accurate and generate faster processing times and hence aremore viable in real-time sentiment analysis. This is an illustration of the performance and efficiency versus computational costs of models. We further discuss the particular issues of Twitter data and provide how these models deal with non-formal language. Our study can be used to make valuable contributions to the process of choosing the most suitable model on the basis of what the task requires, and it is necessary to focus on the aspect of balancing between the understanding of context and efficiency. The next generation of

research will be devoted to the enhancement of the accuracy of GPT-based models and investigating hybrid models that would integrate the advantages of other methods.

Conclusion

In this research,we compared five NLP models, which include GPT-4, GPT-3.5, BERT, RoBERTa, and DistilBERT models for Twitter sentiment classification. Results showed that the BERT model achieved the best performance result with an F1-score of 0.87%, while DistilBERT offered similar accuracy with lower computation cost. GPT-4 achieved high recall but lower precision, indicating its strong contextual understanding yet weaker balance in prediction accuracy.The BERT-based models remain the most efficient and balanced for sentiment classification, while GPT models need further fine-tuning for real-time use. Future work will focus on improving GPT-based model precision, exploring hybrid ensemble methods, and testing model scalability in real-time applications.

Author contributions

All authors are contributed equally for designing the research plan to development of this manuscript.

Funding

This research received no specific grant from any funding agency.

Conflict of interest

The author declares no conflict of interest. The manuscript has not been submitted for publication in other journal.

Ethics approval

Not applicable

AI tool usage declaration

The authors not used any AI and related tools to write this manuscript.

References

Chumakov, S., Kovantsev, A., & Surikov, A. (2023). Generative approach to aspect-based sentiment analysis with GPT language models. Procedia Computer Science, 229, 284–293. https://doi.org/10.1016/j.procs.2023.12.030

Fei, Hao, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. "Reasoning implicit sentiment with chain-of-thought prompting." *arXiv preprint arXiv:2305.11255* (2023).

Gowda, K. P., Porwal, R., Ramesh, C., Tiwari, S. S., Srivastava, K., Rambabu, R., & Govinda Rao, S. (2025). Transformers in sentiment analysis: A paradigm shift in deep learning research.*Journal of Information Systems Engineering and Management, 10*(5s), 262–280.https://doi.org/10.52783/jisem.v10i5s.612:contentReference[oaicite:13]{index=13}

https://doi.org/10.1109/SMART50582.2020.9337081

Kang, J.-W., & Choi, S.-Y. (2025). Comparative investigation of GPT and FinBERT's sentiment analysis performance in news across different sectors. *Electronics, 14*(6), 1090. https://doi.org/10.3390/electronics14061090

Kong, Y., Xu, Z., & Mei, M. (2023). Cross-domain sentiment analysis based on feature projection and multi-source attention in IoT. *Sensors, 23*(16), 7282. https://doi.org/10.3390/s23167282 Wu, Q., Xia, C., & Tian, S. (2025). AI-driven sentiment analytics: Unlocking business value in the e-commerce landscape. *arXiv*. https://arxiv.org/abs/2504.08738

Lin, Y., & Liu, T. (2024). *Impact of NLP Algorithms on Sentiment Analysis Efficiency and Accuracy.* Journal of Information Systems and Informatics.

Liu, Y., et al. (2022). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692.

Mathebula, M., Modupe, A., &Marivate, V. (2024). *Fine-tuning retrieval-augmented generation with an auto-regressive language model for sentiment analysis in financial reviews*. Applied Sciences, 14(23), 10782. https://doi.org/10.3390/app142310782

Miranda, C. H., Sanchez-Torres, G., &Salcedo, D. (2023). Exploring the evolution of sentiment in Spanish pandemic tweets: A data analysis based on a fine-tuned BERT architecture. *Data*, *8*(6), 96.

Pak, A., &Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc* (Vol. 10, No. 2010, pp. 1320-1326).

Perikos, I., & Diamantopoulos, A. (2024). Explainable Aspect-Based Sentiment Analysis Using Transformer Models. *Big Data and Cognitive Computing, 8*(11), 141. https://doi.org/10.3390/bdcc8110141

Prova, N. (2025). Multilingual Emotion Classification in E-Commerce Customer Reviews Using GPT and Deep Learning-Based Meta-Ensemble Model. *Available at SSRN 5161505*.

Qin, L., Chen, Q., Wei, F., Huang, S., & Che, W. (2023). Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://aclanthology.org/2023.emnlp-main.163/

Reddy, S., Kumar, A., & Sharma, P. (2025). Leveraging sentiment analysis in the digital era: Uncovering insights from unstructured data for enhanced customer engagement. *Journal of Modern Technology*, 21(3), 212–219Fei, H., Li, B., Liu, Q., Bing, L., Li, F., &Chua, T.-S. (2023). Reasoning implicit sentiment with chain-of-thought prompting. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://arxiv.org/abs/2305.11255

Semary, N. A., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2023). Improving sentiment classification using a RoBERTa-based hybrid model. *Frontiers in human neuroscience*, *17*, 1292010.

Shafikuzzaman, M., Islam, M. R., Rolli, A. C., Akhter, S., &Seliya, N. (2025). *An empirical evaluation of the zero-shot, few-shot, and traditional fine-tuning based pretrained language models for sentiment analysis in software engineering*. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3439450

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, *13*(7), 4550.

Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Proceedings of the 2023 Conference on Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/2305.04388

Vamvourellis, D., & Mehta, D. (2025). Reasoning or overthinking: Evaluating large language models on financial sentiment analysis. *arXiv*. https://arxiv.org/abs/2506.04574

Widyananda, W., Maskur, & Fauzi, A. (2025). *Machine Learning and Transformer-based Model for Sentiment Analysis of Indonesian E-Commerce Reviews.* The Indonesian Journal of Computer Science, 14(4). (ai-soc.org)

Wiguna, B. S., Hudiyanti, C. V., Rausanfita, A., & Arifin, A. Z. (2021). *Sarcasm Detection Engine for Twitter Sentiment Analysis using Textual and Emoji Features.* Jurnal Ilmu Komputer dan Informasi, 14(1). (jiki.cs.ui.ac.id)

Yunitasari, Y., Musdholifah, A., & Sari, A. K. (2019). *Sarcasm Detection For Sentiment Analysis in Indonesian Tweets.* Indonesian Journal of Computing and Cybernetics Systems, 13(1). (Jurnal Universitas Gadjah Mada)