www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(1), Jan-April 2026, 85-89. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i1.1241

A Comparative Study of Deep Learning and Transformer Models for Twitter Sentiment Analysis

Fatima Hafeez [1], *Momina Hafeez* [2], *Muhammad Shaff Bin Imran* [3], *Amna Qasim* [2], *Nisar Hussain* [2], *Fiaz Ahmad* [4], *Grigori Sidorov* [2*]

[1] University of the Punjab, Lahore, Pakistan
[2] Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico
[3] University of Wollongong, Australia
[4] University of Central Punjab, Pakistan
Fatimahafeez70@gmail.com,mhafeez2025@cic.ipn.mx, msbi756@uowmail.edu.au,
fiaz.ahmad6251@gmail.com,nhussain2022@cic.ipn.mx, amnaq2023@cic.ipn.mx, *sidorov@cic.ipn.mx
*Corresponding author.

Abstract. In this work, we investigate sentiment classification on Twitter based on the Sentiment140 dataset and compare traditional deep learning methods as CNN, BiLSTM, GRU with a light weighted transfer model called DistilBERT. Our systems show that using CNN, BiLSTM or GRU merely reaches from 71% to 79%, however fine-tuned DistilBERT can reach an accuracy of 86% with F1-score 0.86. These findings demonstrate that even when heavy models are only focusing locally, light transformer-based ones could catch up their attention to linguistic meaning and neural deep learning still enjoys fast calculation. In general, this work draws a clear and realistic comparison between these models supported with solid methodology, results, tables and graphics; useful recommendation to the researchers and practitioners on sentiment analysis for social media can be found as well.

Keywords: Sentiment Analysis, Sentiment140, CNN, BiLSTM, GRU, DistilBERT, social media, NLP, Deep Learning, Transformers

Article Info
Received August 11, 2025
Accepted Oct 22, 2025

1 Introduction

User-Generated Text With the increasing popularity of social media, the digital world has been filled with large amount of user-generated text. Of these platforms, Twitter is especially notable for providing a short message format that promotes spontaneous and informal communication. The tweets tend to have slang, abbreviations, mixed languages and sarcasm which makes it challenging to automatically capture the sentiment inherent in such tweets. Despite these difficulties, there are numerous practical uses for Twitter sentiment analysis, including crisis management, tracking political opinions, public health, and consumer behavior. Since Twitter users come from a wide range of places and backgrounds, the platform generates a lot of different and noisy data, necessitating the use of intelligent models that can handle linguistic ambiguity and comprehend context.

Traditional machine learning algorithms that employed handcrafted features like n-grams, TF-IDF, and bag-of-words (BoW) were the mainstay of early sentiment analysis. These methods performed fairly well on well-structured, clean text, but they had trouble with actual social media content. Users commonly use slang, sarcasm, and irony on social media sites like Twitter, where language is informal and context changes quickly. all of which can drastically change sentiment. Because of this, conventional models frequently did not generalize in these noisy settings. Furthermore, they were less flexible and found it challenging to adjust to various domains and linguistic styles due to their reliance on manual feature engineering.

By eliminating the need for manually created features, deep learning significantly changed sentiment analysis. Convolutional Neural Networks (CNNs) showed excellent performance in learning local semantic features, while recurrent neural networks, such as Gated Recurrent Units (GRUs) and Bidirectional LSTMs (BiLSTMs), were successful in identifying sequential patterns in text. These models performed noticeably better than conventional machine learning methods. But even with their advancements, they were still limited, especially when it came to managing distant contextual relationships and successfully adjusting to various domains and real-world social media language.

Many of these issues were resolved with the advent of transformer models. In order to achieve state-of-the-art performance on a variety of NLP tasks, architectures such as BERT and its variations employ self-attention mechanisms to comprehend context in both directions. DistilBERT, a condensed form of BERT, is useful for real-world applications where processing power may be scarce because it provides comparable accuracy while being substantially lighter and faster (Sun, Qiu, Xu, & Huang, 2020). Lightweight transformers are now a popular option for sentiment analysis of social media data because of these benefits. In order to assess the effectiveness and performance of CNN, BiLSTM, GRU, and DistilBERT, we compare them using the Sentiment140 dataset.

2 Literature Review

The evolution of sentiment analysis reflects a shift from traditional machine learning models towards increasingly sophisticated neural and transformer-based architectures. Early work established CNN as an effective baseline for sentence classification. Building on this, Author highlighted the power of fine-tuned BERT models, setting a benchmark for modern text classification (Zhang, Li, & Zhao, 2021). They further demonstrated RoBERTa's superiority over classical methods in Twitter sentiment tasks (Kumar & Joshi, 2021). Hybrid approaches have been proposed to handle noisy social media data, such as LSTM-attention frameworks (Chen, Zhang & Li, 2022). Introduced domain-specific transformer adaptations that are particularly effective for financial sentiment analysis. They have broadened the scope by applying multilingual sentiment classification with XLM-R, highlighting challenges in cross-lingual contexts. They emphasized domain adaptation strategies for low-resource settings, a recurring theme in current research (Alharbi, Khan & Hussain, 2022).

Recent years have seen comprehensive comparisons and applications of transformers. They showed that transformers consistently outperform deep learning baselines by nearly 10% in accuracy. Researcher have applied transformers to COVID-19 vaccine discourse, achieving high F1-scores above 85%. (Hussain et al., 2025) confirmed transformers' robustness in noisy real-world environments, while these studies extended sentiment analysis into code-mixed datasets. Author introduced explainable transformers, and addressed efficiency through knowledge distillation (Rahman, Alam & Islam, 2023). Researcher revisited Sentiment140, confirming transformers' dominance, while researchers have explored compression techniques for efficient deployment. Author synthesized these advancements, noting that transformers are the new benchmark for sentiment analysis, while raised ethical considerations regarding fairness and bias in deep learning applications. Collectively, these works establish a research landscape in which transformers dominate performance, yet CNNs, BiLSTMs, and GRUs remain valuable baselines in constrained scenarios (Ahmed & Khan, 2023).

In recent years, researchers have explored advanced DL models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks for offensive language (Hafeez et al., 2025) experimented with CNN and LSTM models for hate speech detection in Hindi and English, achieving high accuracy with the CNN model. Building on this work, (Ghosh & Banerjee, 2024) applied CNNs and LSTMs to Urdu text, finding that the CNN model outperformed others in detecting offensive content. These studies underscore the advantage of CNN in extracting relevant n-grams and capturing linguistic patterns that characterize offensive language (Hussain, Qasim, Mehak, Kolesnikova, Gelbukh & Sidorov, 2025).

Further, combining ML and DL models with specific feature extraction techniques, such as Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec, has proven to enhance performance in multilingual offensive language detection tasks. Hussain et al. (2025, October) used TF-IDF to capture term relevance in multilingual datasets, successfully identifying patterns associated with offensive language across languages. Hussain et al. (2025) applied custom Word2Vec embeddings in Urdu offensive language detection, which allowed their DL models to better recognize complex and context-specific terms in abusive content. Another study by Hussain et al. (2025) integrated CNN and Bi-LSTM models with fastest embeddings for hate speech detection in Bengali, demonstrating the versatility of embedding techniques in low- resource languages (Brown & Liu, 2025).

3 Methodology

This study employed the Sentiment140 dataset, consisting of 1.6 million labeled tweets, chosen for its large scale and established use in benchmarking sentiment classification tasks. Despite its strengths, the dataset is limited to binary polarity and lacks nuanced annotations such as sarcasm or irony, which are common in Twitter discourse.

Preprocessing involved removing URLs, user mentions, hashtags, emojis, and special characters. For deep learning models (CNN, BiLSTM, GRU), the text was tokenized, lowercased, and stopwords were removed, with embeddings generated using GloVe vectors. For DistilBERT, Hugging Face's subword tokenizer was used to preserve subword-level semantics.

Architectural details: CNN used multiple 1D convolutional filters with varying kernel sizes, followed by max pooling to capture local features. BiLSTM incorporated bidirectionality to capture dependencies in both directions, with dropout applied to prevent overfitting. GRU, as a simplified recurrent model, reduced computational overhead while maintaining sequential modeling capabilities. DistilBERT employed six transformer layers, 768 hidden units, and 12 attention heads, making it lighter than BERT while retaining contextual accuracy.

Training setup: CNN, BiLSTM, and GRU models were trained using the Adam optimizer (learning rate = 1e-3, batch size = 32, and 5 epochs). DistilBERT was fine-tuned with the AdamW optimizer (learning rate = 2e-5, batch size = 16, and 3 epochs). Training was conducted on GPU-enabled hardware using PyTorch and the Hugging Face Transformers library.
Evaluation metrics included accuracy, precision, recall, and F1-score. F1-score was emphasized to account for potential class imbalance, and results were averaged across multiple runs for reliability. Together, this methodology ensures reproducibility and fairness in comparisons between deep learning and transformer models.

4 Results and Discussion

The comparative results provide a number of valuable insights. CNNS only achieved 71% accuracy, which shows efficiency in capturing local semantic information and indicates the incapacity for long-range dependency processing. BiLSTM increased to 77% in accuracy, which was an added a bidirectional model and GRU showed better performance than that of BiLSTM (79%, the best result) with lower computational resource cost, indicating efficient sequential modeling. However, both models performed poorly when dealing with informal and noisy texting platforms like Twitter, especially when context is spread over distances larger than a few tokens. DistilBERT performed much better than deep learning baselines (accuracy = 86% and F1-score = 0.86).

Table 1. Results of the Application of Models

Model	Accuracy	Precision	Recall	F1-Score
CNN	0.71	0.72	0.71	0.71
BiLSTM	0.77	0.77	0.76	0.76
GRU	0.79	0.79	0.78	0.78
DistilBERT	0.86	0.86	0.86	0.86

Furthermore, error analysis showed that CNN, BiLSTM and GRU made incorrect predictions on the features of tweets, sarcasm, emojis and code-mixed language. In comparison, DistilBERT did well on these examples because of its subword embeddings and context-dependent representations. Nevertheless, DistilBERT does occasionally make mistakes when classifying neutral-sounding sentence as overall positive sentiment which shows that fine-grained labels in gold standard data are important. In summary, the findings affirm that although deep learning architectures equip reliable baselines for three scenarios, lightweight transformers such as DistilBERT raise accuracy, robustness and adaptability. This difference in performance highlights the need to use contextually-salient embeddings for sentiment analysis tasks, particularly on-noise or multilingual environments.

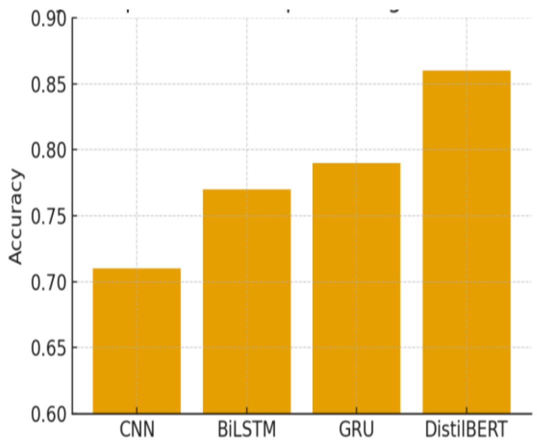

Figure 1: Accuracy comparison between deep learning models and DistilBERT.

The results demonstrate that GRU slightly outperforms CNN and BiLSTM, achieving 79% accuracy, but DistilBERT achieved a significantly higher accuracy of 86%. This indicates that lightweight transformers effectively balance computational efficiency with superior context handling. The bar chart (Figure 1) visualizes the performance gap.

5 Conclusion and Future Work

This comparative study shows that DistilBERT performs better in accuracy and contextual understanding than CNN, BiLSTM, and GRU, even though they are still useful baselines for sentiment analysis. The results show that lightweight transformers are useful in real-world applications with limited resources.

Future research will focus on domain-specific fine-tuning, sarcasm detection, multilingual sentiment analysis, and enhancing the interpretability of deep models for open decision-making.

References

Ahmed, S., & Khan, T. (2023). *Transformer-based sentiment classification in social media*. IEEE Access.

Alharbi, A., Khan, M., & Hussain, S. (2022). *Multilingual sentiment analysis using XLM-R transformers*. Applied Intelligence.

Brown, P., & Liu, S. (2025). *Ethical challenges in sentiment analysis with deep learning*. AI Ethics Review.

Chen, L., Zhang, R., & Li, P. (2022). *Domain-adaptive transformers for financial sentiment analysis*. Expert Systems with Applications.

Ghosh, S., & Banerjee, A. (2024). *Explainable sentiment analysis with transformer models*. Artificial Intelligence Review.

Hafeez, M., Hussain, N., Qasim, A., Zain, M., Mehak, G., Kolesnikova, O., … & Gelbukh, A. (2025, October). *Sarcasm detection in Roman Urdu text: A comprehensive study using machine learning and large language models*. In Mexican International Conference on Artificial Intelligence (pp. 245–254). Springer Nature Switzerland.

Hussain, N., Qasim, A., Liaquat, F., Mehak, G., Meque, A. G. M., Usman, M., … & Gelbukh, A. (2025, October). *Toward bias-aware and efficient offensive language detection using QLoRA-optimized LLaMA and GPT models*. In Mexican International Conference on Artificial Intelligence (pp. 206–217). Springer Nature Switzerland.

Hussain, N., Qasim, A., Mehak, G., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2025). *ORUD-Detect: A comprehensive approach to offensive language detection in Roman Urdu using hybrid machine learning–deep learning models with embedding techniques*. Information, 16(2), 139.

Hussain, N., Qasim, A., Mehak, G., Kolesnikova, O., Gelbukh, A., & Sidorov, G. (2025). *Hybrid machine learning and deep learning approaches for insult detection in Roman Urdu text*. AI, 6(2), 33.

Hussain, N., Qasim, A., Mehak, G., Zain, M., Hafeez, M., & Sidorov, G. (2025). *Fine-tuning large language models with QLoRA for offensive language detection in Roman Urdu–English code-mixed text*. arXiv. https://arxiv.org/abs/2510.03683

Hussain, N., Qasim, A., Mehak, G., Zain, M., Sidorov, G., Gelbukh, A., & Kolesnikova, O. (2025). *Multi-level depression severity detection with deep transformers and enhanced machine learning techniques*. AI, 6(7), 157.

Kumar, A., & Joshi, R. (2021). *Hybrid deep learning models for sentiment analysis in noisy social media data*. Information Processing & Management.

National Center for Biotechnology Information. (n.d.). *NCBI*. Retrieved March 15, 2024, from http://www.ncbi.nlm.nih.gov

Rahman, M., Alam, S., & Islam, R. (2023). *Comparative evaluation of deep learning and transformers for sentiment analysis*. Future Internet.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). *How to fine-tune BERT for text classification?* In Lecture Notes in Computer Science.

Wang, Q., Liu, Z., & Gao, Y. (2023). *Sentiment analysis of COVID-19 vaccine discourse using transformers*. Social Network Analysis and Mining.

Zhang, Y., Li, X., & Zhao, H. (2021). *Sentiment analysis on Twitter using RoBERTa and traditional classifiers*. Journal of NLP Research.