Received: 09 Jan 2026, Accepted: 27 Jan 2026, Published: 28 Jan 2026
Digital Object Identifier: https://doi.org/10.63503/j.ijaimd.2025.219

Review Article

Comparative Analysis: Traditional Models Vs Transformers in Hate Speech Detection

Poonam Chaudhary[1*], Eshaan Nanda[2], Divyam Garg[3], Harshit Goyal[4]

[1, 2, 3, 4] Department of Computer Science and Engineering, The NorthCap University, Gurugram, India

poonamchaudhary@ncuindia.edu[1]

*Corresponding author: Poonam Chaudhary, poonamchaudhary@ncuindia.edu

ABSTRACT

The detection of hate speech is one of the basic functions in deciphering the contents provided in the internet and it is detectable by the natural language processing. It will be noteworthy to offer automated tools that are computationally effective, robust besides having the capability of establishing and regulating hate speech that is produced online. The study is to compare the classical deep learning methods with transformer-based modelling systems of the detection of hate speech that is being expressed in English language. The results of the different methods were placed on the findings of the classification of hate speech with a labelled dataset in terms of the common evaluation measures, accuracy, precision, recall and F1-score. The focal loss and data augmentation approaches used to overcome the imbalance of classes and improve the generalization ways were also addressed in the given research. As per the findings of the present study, the transformer-based methodologies, more so RoBERTa-large methodology, has a significantly superior performance as compared to the classical deep learning methodologies to identify the subtle and context sensitive hate speech cases. Furthermore, this paper demonstrates that the use of large-scaled pre-trained language models and the application of hybrid models are even more important to increase the functionality and efficiency of automated systems of hate speech detection.

Keywords: *hate speech, Long Short-Term Memory Neural Networks, transformer, RobBERTa-Base, RobBERTa-Large.*

1. Introduction

With the rapid growth of the social media platform usage and online communication, hate speech has become a burning social issue. The spread of harmful content endangers the safety of the digital world, fosters discrimination, and leads to the violent reality. It is important to find and cure hate speech to create a secure and inclusive online environment. Maintaining a balance between the changing rather dynamic character of the offensive language that frequently incorporates sarcasm, implicit hate, and code-mixed phrases became extremely hard with the application of classical methods, including rule-based, and lexicon-based approaches. Recent improvements in deep learning, especially the Transformer type models and sequential approaches including RNN, LSTM and GRU, have significantly improved the hate speech detection accuracy [1][2]. Nevertheless, there are a few problems in the area of hate speech detection which we encounter namely the detection of implicit hate speech, implicit hate speech is offensive language that is hidden and contextual. The other one is the dataset bias where unbalanced datasets may result in bias model predictions, over-detecting specific demographic groups and underperforming on others [4]. Moreover, deep learning models used can fail to generalize, i.e. a model that is trained on one platform (e.g., Twitter) might not work on another (e.g., YouTube or Reddit) because of dissimilarities in the use of language and community standards [5]. This paper seeks to answer the question of the low quality in hate speech accuracy by considering a mix of

deep learning models, such as ANN, RNN, LSTM, GRU, TextCNN and Transformer-based models, such as RoBERTa, as well as hybrid models, such as RoBERTa-base + LSTM to detect hate speech better.

The other interesting area of attention is how model design and imbalance in data can be used to enhance the detection accuracy [6]. Furthermore, this research topic examines the problems of datasets biases and examines the possibilities of reducing the disparate model predictions to increase the accuracy and safety of the hate speech classification models [7]. Most recent studies also highlight the importance of more complex fusion strategies and regularization to enhance hate speech recognition. Transformer-based networks and other mechanisms like "Attentive Fusion" have been investigated in order to fuse complex features [8]. With all these developments, there are still issues when it comes to dealing with implicit hate speech and biases in datasets. Transfer learning has been studied in order to solve these problems [10][11]. The study is very valuable in theory and practice. Theoretically, it helps to increase knowledge base on deep learning techniques to detect hate speech, especially in the context of learning the effects of various architectures on the accuracy of the classification. In practice, the results can be used to assist social media platforms, policymakers, and content moderation units in creating more useful and objective automated systems of hate speech detection and removal. Moreover, it puts emphasis on a literature review of the traditional, ML, and DL-based hate speech detection approaches, aligned with the thorough discussion of the research methodology, such as dataset selection, preprocessing methods, and model architecture. The next part of the work illustrates the outcomes of the experiment, the comparison of the work of various models, and their usefulness. Lastly, the paper will talk about major findings, limitations, and possible directions of the further research, in which the fairness and generalization is discussed as an important issue in hate speech detection models.

Recent discoveries in hate speech detection have paid more attention to transformer-based models because of their better performance in terms of addressing the peculiarities of the offensive language. This dataset has been tested on classical DL architectures such as CNNs, LSTMs, and GRU's to compare them and identify the best performing one [2, 10, 17]. It has been found out that the combination of various feature types, i.e., semantic, sentiment, topical, may enhance the performance dramatically, and such hybrid models as DeepHate and CNN-GRU-BERT prove particularly effective [2, 10]. Transformer-based architecture has turned out to be unorthodox in hate speech detection. BERT, RoBERTa, ELECTRA, and XLM-R have repeatedly been shown to outperform even deep learning models on a number of different datasets [3, 6, 13, 14]. The contextual improvement by fine-tuning these transformers on labeled datasets greatly boosts classification accuracy, especially at detecting subtle or coded hate speech [4, 5, 15]. Multimodal approaches that combine both text and image data have too been encouraging, as they have demonstrated the importance of integrating features of various modalities [5, 11]. Also, a number of studies note the importance of preprocessing methods such as tokenizing, stopwords, TF-IDF, GloVe, and BERT-based embeddings to effectively represent tweets [1, 3, 8, 9, 12]. One of the most prominent trends in current research is the focus on transfer learning and data augmentation to overcome such issues as the impact of the dataset bias, a lack of classes, and the problem of domain generalization [16, 13]. A number of works also indicate the enhancement of efficiency and scalability of transformers to be used in real-life applications [4, 6, 7]. To sum up, it can be concluded that there is a change in literature to the shift of traditional models to strong and context-sensitive architectures, especially in transformers, to detect hate speech. Nevertheless, issues such as multilingual support, interpretability, bias in the data used, and computation strength are still matters to consider in the future [12, 13, 14, 16].

This study is of great importance to the theory and practice. Theoretically, it helps to broaden the knowledge base on the deep learning approaches to hate speech detection, especially the insights into the effects of various architectures on the classification accuracy. In practice, the results can be used by social media companies, policy makers, and human content moderators to develop more helpful and objective automated hate speech identification and removal systems. Moreover, it brings up to the literature review of the traditional, ML, and DL-based hate speech detecting procedures, aligned with the description of the research methodology, such as data set selection, preprocessing, and model structure. Experimental results are described in the following section where various model

performances are compared and their effectiveness is evaluated. Lastly, the paper also presents essential findings, limitations and future research directions by pointing out the significance of fairness and generalization in hate speech detection models.

2. Material and Method

2.1 Dataset: Twitter Hate Speech Dataset is a frequently utilized resource in the study related to categorization of text as negative or positive and identification of hate speech instances in the past. Twitter Hate Speech Dataset is composed of 3 different categories of tweets:

Hate Speech (0): tweets that included hate speech sent to specific people or groups.

Offensive Language (1): tweets have offensive language but will not explicitly provoke additional hatred.

Neither (2): Tweets that should not be regarded as hate speech and offensive language are neutral.

The data itself is represented by six major columns:

- Unnamed: 0: An index column, which does not play a part in analysis.
- hate_speech The number of hate speech indicators per tweet.
- offensive_language: The number of off-language pointers in a tweet.
- neither: The number of neutral words in the tweet.
- class: Last label of all the tweets (Hate Speech: 0, Offensive Language: 1, Neither: 2).
- tweet: The content of the tweet.

This data is in good format and lacks any missing values so that it is fit for analysis. The data can be accessed on https://www.kaggle.com/datasets/yashdogra/toxic-tweets.

2.2 Methodology

Regarding the comparative analysis, the data was preprocessed with the dataset. To resolve the problem of class imbalance and enhance the process of generalization, such tools as focal loss and data augmentation were utilized. Subsequently other classical deep learning architectures like Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM) Neural Network, Gated Recurrent Unit (GRU) and Text Convolutional Neural Network were used. These models were contrasted with transformer-based architectures such as RoBERTa-Base and LSTM and RoBERTa-Large. In this paper, the performance of both models is determined in terms of such evaluation measures as accuracy, precision, recall and F1-score. According to our results, the models based on transformer, especially RoBERTa-large, achieve a much better result in the localization of subtle and context-specific examples of hate speech in comparison with the traditional architectures. Emerging criticality of large pre-trained language models and hybrid solutions to improving automated systems of hate speech detection are emphasized in this research. Figure 1 offers the research pipeline according to the suggested comparative analysis.

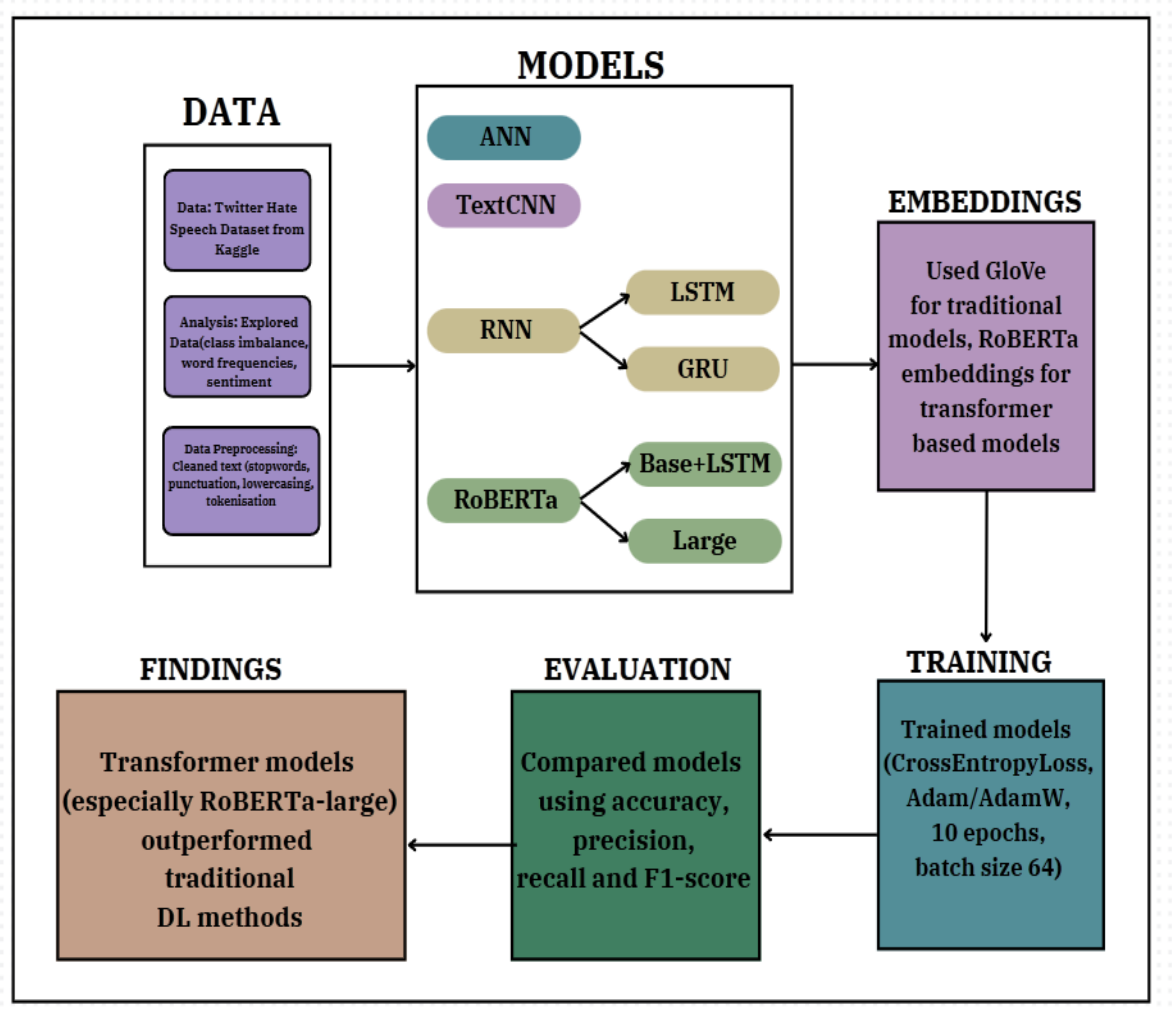
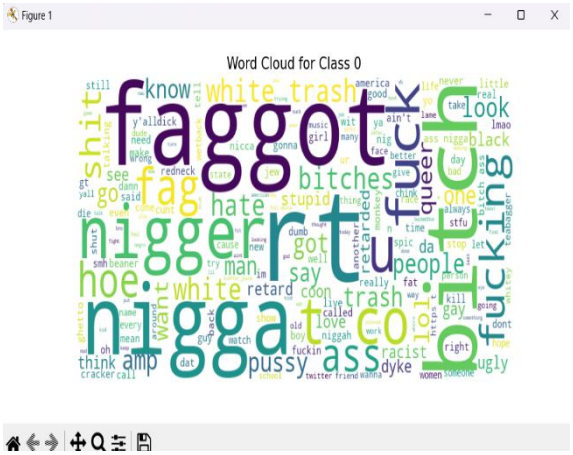


Fig. 1: Research pipeline used for the proposed comparative analysis for Hate Speech detection in English

2.3 Data Pre-processing: Figure 2 provides insight into the dataset including class distribution, word cloud for class 0, class 1, class 2. Figure 3 shows the boxplot for tweet lengths by class and distribution for tweet lengths. Tweet lengths are important for checking the most common lengths and overall variability of the content.



(a) (b)

(c) (d)

Fig. 2 *(a) Class Distribution, (b) Word Cloud for class 0, (c)Word Cloud for class 1, (d) Word Cloud for class 2*

Class imbalance is one of the vital challenges of this dataset. It is evident from the distribution of tweets across the categories:

- Hate Speech [0]: 1,430 tweets (~5.7%)
- Offensive Language [1]: 19,190 tweets (~76.3%)
- Neither [2]: 4,163 tweets (~16.5%)

The Offensive Language category dominates, which may cause models to develop a bias toward this class. To address this issue, techniques such as Synonym Replacement have been applied [9].

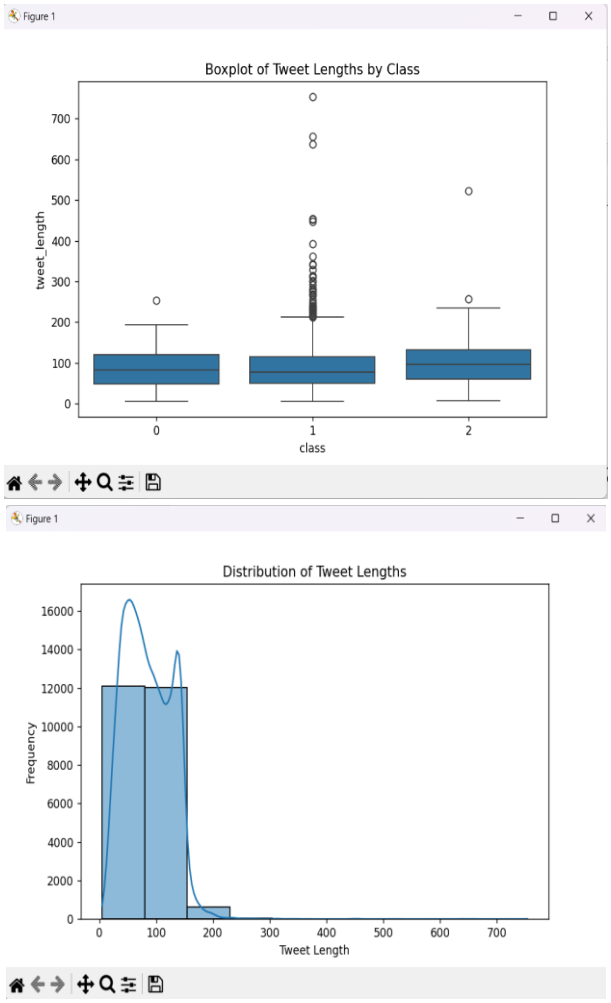


Fig. 3: *(a) Boxplot for Tweet Lengths by Class (b) Frequency Distribution for Tweet Lengths*

Further, several preprocessing steps were applied for preparing the data for model training and to clean the tweets:

- Stop word Removal: Removed ordinary words (for instance, "the", "is", "and") that do not serve any meaningful information.
- Punctuation and Special Character Removal: Cleaned the text by removing unnecessary characters.
- Tokenization: Divide each tweet into tokens.
- Lowercasing: Converted the tweet into lowercase to make sure uniformity across words.

The most recurring words in the dataset include terms like "rt", "bitch", "bitches", "ass", "fuck", "shit", which reflect the appearance of aggressive and offensive language, especially in tweets categorized as hate speech and offensive language.

N-Gram analysis revealed common two-word (bigram) and three-word (trigram) phrases such as:

- Bigram: ("bitch", "ass"), ("hoe", "bitch")
- Trigram: ("hoe", "ass", "bitch"), ("fuck", "bitch", "ass")

These n-grams indicate the use of slang and offensive language, supporting the need for contextual NLP models that can understand patterns beyond simple keyword matching [7].

Sentiment analysis was performed using the TextBlob library [19], and the average sentiment scores across the classes were:

- Hate Speech (0): -0.071 (Most negative sentiment)
- Offensive Language (1): -0.007 (Neutral to slightly negative sentiment)
- Neither (2): +0.081 (Slightly positive sentiment)

This analysis indicates that hate speech tweets are the most negative in sentiment, while neutral tweets tend to be more positive, reflecting general non-aggressive discourse.

Preprocessing and Embedding Setup

As transformer-based architecture requires distinct preprocessing steps,tokenization for the first five models (ANN to TextCNN) was performed using NLTK [26]. A fixed-length sequences of 50 tokens with word embeddings initialized using pre-trained 100-dimensional GloVe vectors [21]. Whereas for RoBERTa-base + LSTM, tokenization and input encoding were handled using Hugging Face's Roberta Tokenizer [28]. We extracted sequence embeddings via the roberta-base transformer (with hidden states enabled) and processed them through a bi-directional LSTM to capture dependencies. The resulting states were pooled and classified using fully connected layers. Further, for RoBERTa-large, tokenization and input encoding were handled using HuggingFace's Roberta Tokenizer. Sequence embeddings were extracted using the roberta-large [26] transformer with output hidden states enabled. The last hidden states were pooled and passed through fully connected layers for classification.

2.4 Model Training and Testing

We compared various DL architectures for hate speech detection using a consistent pipeline. The dataset comprises tweets labeled as Hate Speech, Offensive Language, or Neither. We employed a stratified 80/20 split for training and testing. Given the class imbalance, the weighted F1-score [20] served as the primary performance metric.

A simple feedforward Artificial Neural Network (ANN) model that flattens GloVe-embedded tokens and passes through two fully connected layers. ANN was suitable as baseline but struggles with capturing sequential information. To capture sequential information, multiple Recurrent Neural Networks (RNN, LSTM, GRU) were implemented. These models utilize word embeddings followed by respective recurrent layers to capture sequential dependencies. LSTM and GRU performed better than vanilla RNNs due to their gating mechanisms, which help with long-term dependency modeling [23,24,25]. A convolutional model, TextCNN, that uses multiple filter sizes (3, 4, 5) to pull out n-gram level features from word embeddings, succeeded by max-pooling and classification layers. TextCNN performed competitively for detecting offensive content [22].

In the framework of this hybrid model LSTM+RoBERTa-base, the input material is passed through a pretrained RoBERTa-base model [26] to generate contextualised rich embeddings. These embeddings can encode the fine semantic and syntactic meaning of the text. The RoBERTa output embeddings are then fed into an LSTM layer [24] to obtain sequential dependencies and learn additional information about the way information circulates between the tokens. Finally, a softmax activation dense layer is used to perform classification. This combination combines the strength of transformer-based feature extraction and the sequential modeling capability of LSTM that leads to a superior performance than single models. The most effective model is RoBERTa-large that classifies with the help of boundary token embeddings generated by a very powerful transformer-based encoder (roberta-large) [26]. The

final hidden states were accumulated and transmitted through fully connected layer so as to perform classification. The benefit of this model was that it had rich contextual token embeddings and the top performance on any class.

3. Result and Discussion

CrossEntropyLoss and Adam [29] were the optimizers employed to train all the models with 10 epochs. The RoBERTa-large model incorporated AdamW [30] and was able to converge in 3-5 epochs.

RoBERTa-based models had a batch size of 64 and early stopping of overfitting. The monitoring of the performance of the classes was performed by the classification report of Scikit-learn. In order to assess the efficiency of the deployed models, we compared their performance in terms of the weighted F1-score, accuracy and the class-wise precision and recall [20].

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

$$\text{Macro F1} = \frac{1}{N}\sum_{i=1}^{N} \text{F1}_i$$

$$\text{Weighted F1} = \sum_{i=1}^{N} \left(\frac{n_i}{\sum_{j=1}^{N} n_j} \times \text{F1}_i \right)$$

Table 1: Comparative analysis of the algorithms

Model	***Accuracy***	***Macro Avg F1-Score***	***Weighted Avg F1-Score***
ANN	0.82	0.54	0.80
RNN	0.77	0.29	0.68
LSTM	0.77	0.29	0.68
GRU	0.90	0.70	0.89
TextCNN	0.89	0.70	0.88
LSTM + RoBERTa-base	0.89	0.70	0.88
RoBERTa-large	0.90	0.82	0.90

Table 1 provides performance of all the above-mentioned DL and Transformer based models. **ANN** served as a basic baseline model but struggled with identifying hate speech specifically, showing low recall for that class. **RNN** and **LSTM [23,24]** models showed poor macro F1-scores, mainly because they failed to identify hate speech and "neither" class effectively (both models predicted mainly

"offensive" instances correctly). **GRU** [25] outperformed RNNs and LSTM by a significant margin, achieving a 0.90 accuracy and a much higher macro F1-score. **TextCNN[22]** achieved comparable performance to GRU, showing its effectiveness in extracting local n-gram features. **LSTM + RoBERTa-base [24,26]** further improved results by leveraging contextual embeddings from transformers. **RoBERTa-large [26]** displayed the best results overall, with a weighted F1-score and accuracy of 0.90 and a macro average F1 of 0.82. It effectively balanced precision and recall across all classes, including the minority hate speech class. Figure 4 model comparison on hate speech detection and Fig. 5 shows the Performance metrics for RoBERTa base+LSTM and RoBERTa large.

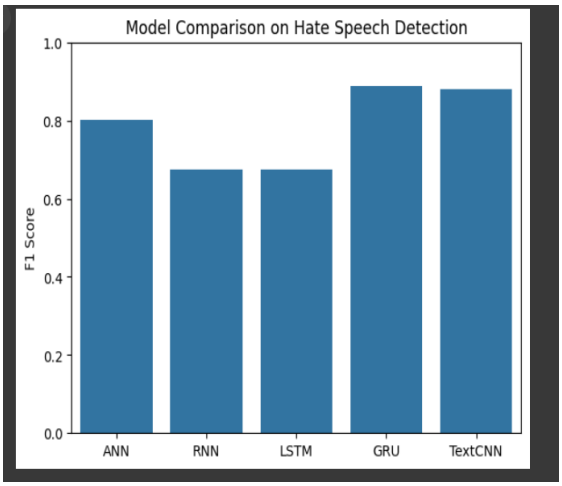


Fig 4 *Traditional DL Model Comparison*

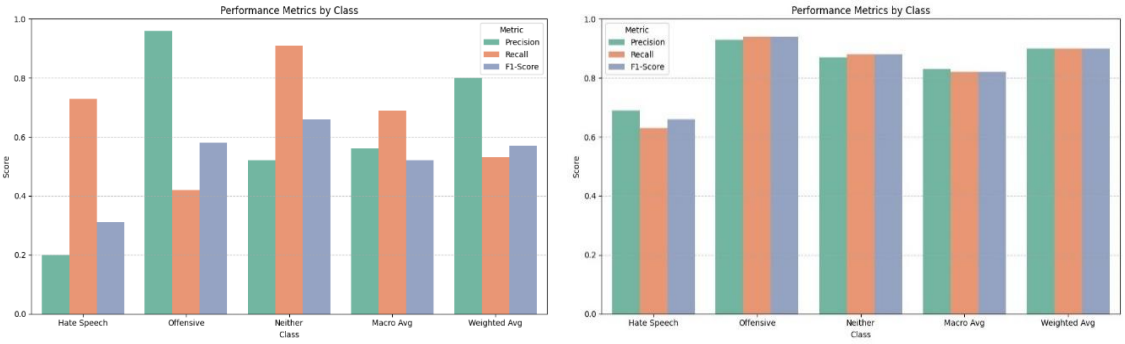


Fig 5 *(i) Performance metrics for RoBERTa base+LSTM, (ii) RoBERTa large*

4. Conclusion

The results demonstrate that transformer-based models, particularly RoBERTa-large, significantly outperform traditional architectures (ANN, RNN, and LSTM). This underscores the critical role of contextualized word representations and large-scale pretraining in hate speech detection. Traditional models struggled with low recall, often biasing predictions toward the dominant 'Offensive' class while misclassifying 'Hate Speech' and 'Neither.' However, while highly effective, RoBERTa-large incurs a higher computational cost compared to simpler architectures. Future research should address this trade-off by exploring lightweight transformers like DistilBERT or ALBERT. Additionally, extending the system to include multimodal data (images, metadata), multilingual support, and Explainable AI (XAI) would enhance robustness and trust in diverse social media environments.

Funding source

None.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] S. G. Roy, U. Narayan, T. Raha, Z. Abid, and V. Varma, "Leveraging multilingual transformers for hate speech detection," *arXiv preprint*, arXiv:2101.03207, 2021.

[2] J. S. Malik, H. Qiao, G. Pang, and A. van den Hengel, "Deep learning for hate speech detection: A comparative study," *International Journal of Data Science and Analytics*, pp. 1–16, 2024.

[3] D. Liu, M. Wang, and A. G. Catlin, "Detecting anti-semitic hate speech using transformer-based large language models," *arXiv preprint*, arXiv:2405.03794, 2024.

[4] V. Dwivedy and P. K. Roy, "Deep feature fusion for hate speech detection: A transfer learning approach," *Multimedia Tools and Applications*, vol. 82, no. 23, pp. 36279–36301, 2023.

[5] G. Ramos, F. Batista, R. Ribeiro, P. Fialho, S. Moro, A. Fonseca, and C. Silva, "A comprehensive review on automatic hate speech detection in the age of the transformer," *Social Network Analysis and Mining*, vol. 14, no. 1, Art. no. 204, 2024.

[6] Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] D. Putra and H.-C. Wang, "Advanced BERT-CNN for hate speech detection," *Procedia Computer Science*, vol. 234, pp. 239–246, 2024.

[8] P. K. Roy, A. K. Tripathy, T. K. Das, and X. Z. Gao, "A framework for hate speech detection using deep convolutional neural network," *IEEE Access*, vol. 8, pp. 204951–204962, 2020.

[9] M. Madhavi, S. Agal, N. D. Odedra, H. Chowdhary, T. S. Ruprah, V. A. Vuyyuru, and Y. A. B. El-Ebiary, "Elevating offensive language detection: CNN-GRU and BERT for enhanced hate speech identification," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 5, 2024.

[10] Mandal, G. Roy, A. Barman, I. Dutta, and S. K. Naskar, "Attentive fusion: A transformer-based approach to multimodal hate speech detection," *arXiv preprint*, arXiv:2401.10653, 2024.

[11] M. Subramanian, et al., "A survey on hate speech detection and sentiment analysis using machine learning and deep learning models," *Alexandria Engineering Journal*, vol. 80, pp. 110–121, 2023.

[12] H. Saleh, A. Alhothali, and K. Moria, "Detection of hate speech using BERT and hate speech word embedding with deep model," *Applied Artificial Intelligence*, vol. 37, no. 1, Art. no. 2166719, 2023.

[13] L. Yuan, et al., "Transfer learning for hate speech detection in social media," *Journal of Computational Social Science*, vol. 6, no. 2, pp. 1081–1101, 2023.

[14] O. E. Ojo, et al., "Automatic hate speech detection using deep neural networks and word embedding," *Computación y Sistemas*, vol. 26, no. 2, pp. 1007–1013, 2022.

[15] S. Abro, S. Shaikh, Z. H. Khand, A. Zafar, S. Khan, and G. Mujtaba, "Automatic hate speech detection using machine learning: A comparative study," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, 2020.

[16] R. Cao, R. K. W. Lee, and T. A. Hoang, "DeepHate: Hate speech detection via multi-faceted text representations," in *Proceedings of the 12th ACM Conference on Web Science*, Jul. 2020.

[17] S. Loria, "TextBlob documentation," *Release 0.15*, p. 269, 2018.

[18] F. M. Miranda, N. Köhnecke, and B. Y. Renard, "HiClass: A Python library for local hierarchical classification compatible with scikit-learn," *Journal of Machine Learning Research*, vol. 24, no. 29, pp. 1–17, 2023.

[19] S. Anjali Devi and S. Sivakumar, "An efficient contextual GloVe feature extraction model on large textual databases," *International Journal of Speech Technology*, vol. 25, no. 4, pp. 793–802, 2022.
[20] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint*, arXiv:1408.5882, 2014.
[21] Y. Yu, et al., "A review of recurrent neural networks: LSTM cells and network architectures," *Neural Computation*, vol. 31, no. 7, pp. 1235–1270, 2019.
[22] J. Chung, et al., "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint*, arXiv:1412.3555, 2014.
[23] Y. Liu, et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint*, arXiv:1907.11692, 2019.
[24] H. U. Khan, et al., "Analyzing student mental health with RoBERTa-Large: A sentiment analysis and data analytics approach," *Frontiers in Big Data*, vol. 8, Art. no. 1615788, 2025.
[25] P. Timsina, *Building Transformer Models with PyTorch 2.0: NLP, Computer Vision, and Speech Processing with PyTorch and Hugging Face*. New Delhi, India: BPB Publications, 2024.
[26] Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint*, arXiv:1711.05101, 2017.