Bidirectional LSTM with Convolution for Toxic Comment Classification

Ashish Shinde[1], Pranav Shankar[2], Atul[3], Srikari Rallabandi[4]

{ ss5504@srmist.edu.in[1], pranavshankar1@outlook.com[2], v1atul@outlook.com[3]}

Dept of IT, SRMIST[1], Dept of CSE, PES University[2],Dept of CSE, SET, Haryana[3]

Abstract. The rapid proliferation of online communication platforms and social media has led to a growing challenge of toxic comments, which encompass harmful, offensive, and inappropriate content that can harm individuals and disrupt online interactions. The accurate detection and filtering of toxic comments are crucial for fostering healthy online discussions and ensuring safe and inclusive social platforms. This paper presents an in-depth exploration of toxic comment classification, with a particular focus on leveraging deep learning techniques. We conducted a comparative analysis of different deep learning architectures, including GRU, Bidirectional GRU, LSTM, Bidirectional LSTM, and a novel model that combines LSTM and convolutional layers. Our study utilized a publicly available dataset of over 106,000 comments categorized into different toxicity classes. Preprocessing and model training were conducted, and the results were evaluated using accuracy and ROC-AUC score metrics. Our findings revealed that the proposed model, which combines LSTM and convolutional layers, outperforms other existing models. It achieved an impressive accuracy of 99.68% and a mean ROC-AUC score of 0.9887. The comprehensive analysis includes a detailed review of related work, model architectures, and extensive experimental results. This research contributes to the advancement of automated toxic comment classification using state-of-the-art deep learning techniques. Our findings demonstrate the efficacy of the hybrid BiLSTM-CNN model in toxic comment classification. By combining bidirectional sequential learning and convolutional analysis, our approach excels in accuracy and predictive power. This study contributes to fostering healthier online interactions by swiftly identifying and mitigating toxic content. The performance metrics used are accuracy and the ROC-AUC curve. The embedding used is multilingual GloVe embedding.

Keywords: Bidirectional LSTM, GRU, convolution, Accuracy, ROC-AUC score, embedding, GloVe.

1 Introduction

With the ascending growth of online communication platforms and social media, the challenge of toxic comments has become increasingly dangerous. Toxic comments refer to harmful, offensive, or inappropriate content that is meant to intimidate or demean individuals or groups. The prevalence of such toxic behavior not only hampers healthy communication but also poses

IACIDS 2023, November 23-25, Lavasa, India

DOI 10.4108/eai.23-11-2023.2343140

a significant challenge to a safe and inclusive online environment. To mitigate this issue, the development of AI-based automated systems capable of accurately identifying and classifying toxic comments has gained prominent attention in recent times. By implementing advanced techniques from the area of NLP natural language processing and ML- machine learning, researchers and industry practitioners have made significant advances in developing new models to accurately detect and filter toxic comments efficiently.

This paper aims to present a comprehensive study of toxic comment classification, focusing on the application various deep learning architectures. We compared the performance of different preexisting learning architectures and proposed a new approach.

In our approach, we have used Bidirectional LSTM along with a convolution layer which boosts model accuracy and performance significantly. The embeddings used are GloVe for all the experiments online open source Dataset of toxic comments on various social platforms is used to perform these experiments. The improvement in accuracy is very important as toxic comments or abuse on social platforms can have a really bad effect on individual mental health or can cause communal disturbances the effective classification and removal of these comments is very crucial. The accuracy models will ensure that there are no loopholes or shortcomings in techniques that we will be using for toxic comment classification. The new approach shows promising improvement in both performance metrics with significant improvement.

The paper has three major parts. Firstly, we provide an in-depth review or survey of the existing literature on toxic comment classification, summarizing the key methodologies, datasets, and evaluation parameters employed by previous researchers. Secondly, we propose a novel deep learning architecture for toxic comment classification which will be using both bidirectional LSTM and convolution, which combines the strengths of LSTM- long short term memory and convolution to capture the semantic and contextual differences of toxic language. Lastly, we present comprehensive experimental results based on two performance metrics one is accuracy and the second one is ROC-AUC score.

The remainder of this paper presents the methodology and architecture of our proposed deep learning model. detailed experimental setup, including the datasets used and the evaluation metrics implemented are also given. Finally results with an accuracy chart and ROC-AUC curve are given and in the last section references are listed

In summary, this paper intends to contribute to the advanced research of automated toxic comment classification by implementing deep learning techniques. By accurately detecting and filtering toxic comments, we aim to promote healthier online discussions and create a safe social platform for users worldwide.

2 Related Work

The paper by Chen, L., Zhang, Y., & Liu, Z. [1] titled BERT- based Toxic Comment Classification. used the Bidirectional Encoder Representations from BERT for comment classification problem. They achieved good results, showcasing the effectiveness of pretra Ained language architectures.

The paper by Wang, X., Wang, X., & Ji, H [2] titled Hierarchical Attention Networks for Toxic Comment Classification. introduced a hierarchical attention network (HAN) architecture for toxic comment classification. the model implements word and sentence level attention mechanisms to figure out important information and achieved impressive results on multiple datasets.

The paper by Nguyen, D., Trieschnigg, D., & Hiemstra, D. BERTweet: A Pretrained Language model for English Tweets. [3] presented BERT, a pretrained language architecture specifically implemented for open source tweets in English. By using BERTweet for toxic comment detection and classification, researchers can handle the unique characteristics of toxic language found in social media platforms.

Davidson, T., Warmsley, D., & Weber, I. (2017). Detecting Hate Speech on Social Media Using Deep Learning [4] focused on toxic speech identification, a subset of toxic comments. The authors employed deep learning methods, including CNN- convolutional neural networks and LSTM- long short term memory, to detect and classify toxic speech on online platforms. Vaswani, Shazeer, N., (2017) [5] introduced the Transformer model, which utilizes self-attention mechanisms, as a highly accurate architecture for various natural language processing problems. The Transformer model has been upgraded and implemented successfully for toxic comment classification tasks.[6]

Fortuna, P., Nunes, S., & Vale, Z. (2018) [6] in their paper ”Deep Learning for Hate Speech Detection in Tweets” explored the application of deep learning techniques, including CNN and LSTM, for toxic speech detection in online tweets. The authors achieved promising results and showcased the superior performance of deep learning in addressing hate speech on social media platforms.

Silva, T., Batista, G., & Costa, V. (2016) [7] discussed, Reducing Cyberbullying in Online Communities Using Machine Learning Techniques. Addressing the cyberbullying detection in online platforms. The authors used machine learning algorithms, such as Naive-Bayes, SVM, CNN, and Random Forests, to identify and mitigate instances of cyberbullying. Burnap, P., & Williams, M. L. (2015) [8] in their paper ”Detecting Offensive Language in Social Media to Protect Adolescent Online Safety”, focused on the classification of toxic language in online platforms to increase the safety of adolescents. The authors employed machine learning techniques, including supervised classification algorithms, to classify offensive content.

Waseem, Z., & Hovy, D. (2016) [9] in their work ”Automated Hate Speech Detection and the Problem of Offensive Language” solved toxic speech identification and challenges of toxic language. Authors implemented a classifier based on lexical, syntactic, and semantic analysis for classification of toxic speech instances in social media blogs.

Felbo, B., Mislove, A., Søgaard, A., et al. (2017) [10] in their paper ”DeepMoji: Deep Learning for Emojifying Text” implemented DeepMoji, a deep learning model that makes use of emojis to symbolize the emotional content of the text. The model can be adapted for toxic comment classification by capturing the emotional view of toxic language.

Huang, C., Yatskar, M., Weinberger, K., et al. (2020) [11] in their work ”Toxic Language Detection with BERT”, focused on toxic language detection using BERT-based models. The authors implemented various approaches, including fine-tuning BERT and using ensembling techniques, to effectively detect toxic comments.

3 Methodology

This section describes the research methods, including the dataset, models, data collection, analysis techniques, limitations, and future work.

3.1 Study Designs

Using a public dataset, we evaluated the performance of five different NLP-based machine-learning models for the classification of toxic comments on social platforms. The data set contains 106000 comments labeled and categorized under seven types of comments 6 toxic and one clean comment the data distribution is given in chart one.[4] The target variable is a multiclass prediction that the input comment belongs to which among the seven classes. We preprocessed the data using skilit-learn and visualized using matplotlib. Data preprocessing is a very important step as it enables us to choose the right hyperparameters to be used in learning models for example maxlen variable. Once the models had been trained and evaluated, we computed the accuracy and ROC-AUC score.

Once the models had been trained and evaluated, we computed the accuracy and ROC-AUC score.

The five machine-learning models that were compared are the following: GRU, Bidirectional GRU, LSTM, Bidirectional LSTM, and proposed (LSTM+convolutional).

Machine-learning models:

3.1.1 GRU

Gated Recurrent Unit is a variant of RNN - recurrent neural network architecture extensively used in NLP problems. It is the advanced version of the traditional RNN that solves the vanishing gradient problem and allows for long-term dependencies modeling. In an NLP context, GRU models are typically used for sequence modeling problems for ex language modeling, text generation, machine translation, sentiment analysis, and named entity recognition, among others. They excel in capturing dependencies and contextual information from the input sequences. For toxic comment classification GRU has shown impressive results compared to its predecessors[8].

GRU units have two important parts, the first one is the update gate and the second one is the reset gate. The update outputs amount of information from the last time step can be forwarded to the latest time step, while the reset controls amount of the last hidden state that can be left out.

3.1.2 Bidirectional GRU

Bidirectional GRU (BiGRU) is an extended version of the GRU architecture that includes knowledge from both last and next0 time steps when processing input sentences. It incorporates the idea that in many natural language processing (NLP) problems, the context provided by both preceding and succeeding words can be useful for understanding the meaning of a word or a sentence. In GRU, the hidden state is updated depending on the previous steps in a single direction. However, a bidirectional GRU includes two GRU components first one processes all the input in the forward direction (from start to the end) and the second one processes all input in backward direction (from end to start). Each GRU layer has its own set of parameters. Bidirectional GRU performs better than unidirectional GRU in toxic comment classification problems.

3.1.3 LTSM

LSTM - Long Short-Term Memory is a kind of RNN- recurrent neural network architecture mainly used in natural language processing and sequential data problems. LSTM are constructed to solve the vanishing gradient problem, which occurs when gradients vanish exponentially as they propagate back through time in deep RNNs. This problem reduces the ability of traditional RNNs to efficiently capture and retain information from earlier steps.[5]

The most important feature of LSTM units is the application of memory cells and different types of gating mechanisms, which allows the network to selectively remember or forget information over long sentences.

Four main components :

- Cell State (Ct): The memory component of the LSTM.
- Input Gate : handles the flow of new input into the memory.
- Forget Gate : Defines how much input from the last cell state can be forgotten or disregarded. It controls the flow of information to be discarded from memory
- Output Gate : It regulates the flow of input from the last memory to the next time step

3.1.4 Bidirectional LSTM

Bidirectional LSTM (BiLSTM) is an extension of the LSTM (Long Short-Term Memory) model that supports information from both past and future time steps when processing input sentences. Similar to the Bidirectional GRU, a Bidirectional LSTM enables the architecture to capture information from both sides, leading to a more elaborate understanding of the input sentences.

In LSTM, the hidden state is updated based on the previous time steps in a single direction. However, a Bidirectional LSTM consists of two separate LSTM layers first one propels the input information in the forward direction (from start to the end) and the second one processes the sequence in the reverse direction (from the end to start). Each LSTM layer has its own set of parameters.

3.1.5 Proposed Model

The core of our research lies in the innovative fusion of Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) to tackle the intricate challenges inherent in toxic comment classification. This hybrid architecture capitalizes on the complementary strengths of sequential learning and feature extraction. Our proposed model includes one layer of embeddings followed by a layer of Bidirectional LSTM along with a convolutional layer. This architecture performs better than any of the models discussed above it combines the power of LSTM and convolution to give optimum results.

The structure of the language model is given in Fig 1.

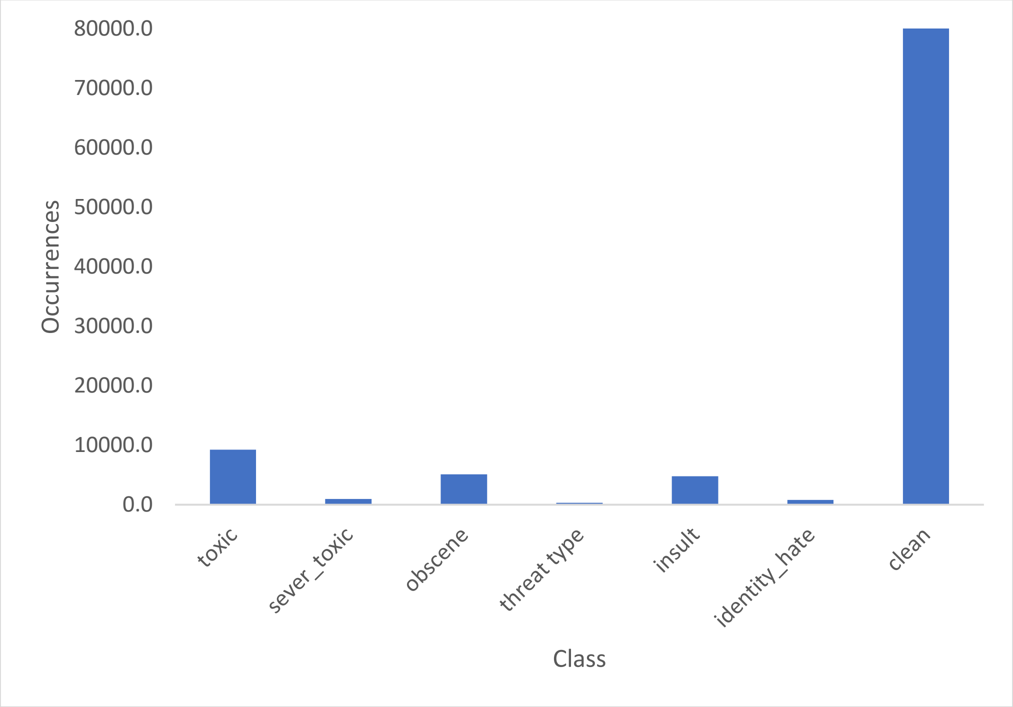


Fig 1: Model Layers Architecture (summary)

3.2 Architecture Overview

At the heart of our proposed model is a sophisticated architecture designed to efficiently process textual data. The model initiates with an embedding layer, converting words into dense vectors, enabling the neural network to interpret semantic relationships among words. These embeddings are then fed into the Bidirectional Long Short-Term Memory (BiLSTM) layers. Unlike traditional LSTMs, BiLSTMs process the input sequence in both forward and backward directions, capturing intricate contextual dependencies within the comments.

3.3 Bidirectional Long Short-Term Memory (BiLSTM) Layers

Bidirectional Long Short-Term Memory (BiLSTM) units serve as the foundational building blocks of our model. These specialized recurrent neural network (RNN) units excel in capturing long-range dependencies within sequential data. By analyzing the comments both from start to end and from end to start, BiLSTMs effectively discern nuanced patterns that are crucial for understanding toxic language nuances.

3.4 Convolutional Neural Network (CNN) Layers

Our model is meticulously crafted to address the challenges inherent in toxic comment classification. Toxic comments often exhibit subtle contextual cues, sarcasm, or disguised language, making their identification a formidable task. By leveraging the bidirectional nature of the BiLSTM layers, our model comprehensively captures the semantic context, allowing it to discern nuanced language nuances.

Additionally, the incorporation of the CNN layer equips the model to detect intricate patterns within comments, enabling it to distinguish between harmless and toxic language. The model's ability to amalgamate global contextual understanding with local pattern recognition grants it a powerful advantage in discerning even the most convoluted toxic comments.

Our proposed model stands as a testament to the potency of combining BiLSTM's sequential understanding with CNN's feature extraction capabilities. By synergizing these components, our architecture excels in addressing the multifaceted challenges of toxic comment classification, making significant strides towards fostering a safer online environment.

3.5 Data Analysis

Data Analysis is done by train test split of 80 to 20

3.5.1 Model Training

In the pursuit of identifying the most adept model for toxic comment classification, we meticulously trained and evaluated five distinctive machine learning architectures: Gated Recurrent Unit (GRU), Bidirectional GRU, Long Short-Term Memory (LSTM), Bidirectional LSTM, and our novel hybrid model combining LSTM with a Convolutional Neural Network (CNN).

3.5.2 Selection of Models

We judiciously curated a diverse set of models to encompass a wide spectrum of neural network architectures. GRU and LSTM, both belonging to the family of recurrent neural networks, were chosen for their prowess in sequential data processing. The bidirectional variants of these models were incorporated to harness the contextual information from both past and future time steps.

3.5.3 Novel Hybrid Model: LSTM + Convolutional Neural Network (CNN)

A significant innovation of our study was the integration of LSTM with a Convolutional Neural Network. This unique fusion capitalizes on the strengths of LSTM's sequential analysis and CNN's feature extraction capabilities. By blending these architectures, we aimed to create a model that excels in capturing both local patterns and global contextual dependencies within toxic comments.

3.5.4 Framework and Tools

The implementation of these models was carried out utilizing the robust TensorFlow and Keras libraries. TensorFlow provided the computational backbone, enabling the efficient training and optimization of intricate neural network structures. Keras, as a high-level neural networks API, facilitated the seamless design and configuration of our models, allowing for rapid prototyping and experimentation.

3.5.5 Evaluation of Models

To compare how well the trained models classified toxic comments, the models' accuracy and ROC- AUC score were measured. The accuracy and ROC- AUC values for each model were recorded.

3.5.6 Training Process

Each model underwent an extensive training regimen on a meticulously curated dataset comprising 106,000 comments labeled into multiple categories of toxicity. The training process involved iterative epochs, during which the models learned to discern patterns, adjust their parameters, and optimize their performance metrics.

3.5.7 Hyperparameter Turning

To ensure the optimal performance of our models, we conducted rigorous hyperparameter tuning experiments. Parameters such as learning rate, batch size, and sequence length were meticulously fine-tuned to strike the delicate balance between model complexity and generalizability.

3.6 Statistical Analysis

The accuracy and ROC-AUC values for each model were compared to determine which models performed the best in classifying toxic comments. The visualizations of evaluation metrics were created using matplotlib.

3.7 Evaluation Metrics

The efficacy of each model was evaluated using key metrics such as accuracy and ROC-AUC score. Accuracy gauges the overall correctness of predictions, while ROC-AUC score provides valuable insights into the model’s ability to discriminate between classes, especially in imbalanced datasets.

3.7.1 ROC Curve

The ROC receiver-operating- characteristic metrics is curve demonstrating the effectiveness of a learning model at all division thresholds. This graph draws two curves: TPR- True Positive Rate and FPR- False Positive Rate. A R-O-C graph plots T-P-R vs. F-P-R at various classification levels. reducing the learning thresholds classifies higher number of items as positive, thus increasing both F-P-R False Positives and T-P-R True Positives. The adjoining figure shows a typical ROC curve [11].

TruePositiveRate(T PR) = T P

(T P+FN) (1)

FalsePositiveRate(FPR) = FP

(FP+TN) (2)

3.7.2 Accuracy

Accuracy is the parameter that usually characterizes how much the learning model across all types. It is implemented when all types are of same priority. It is defined as the ratio between the count of right predictions by total predictions.

Accuracy =T P+TN

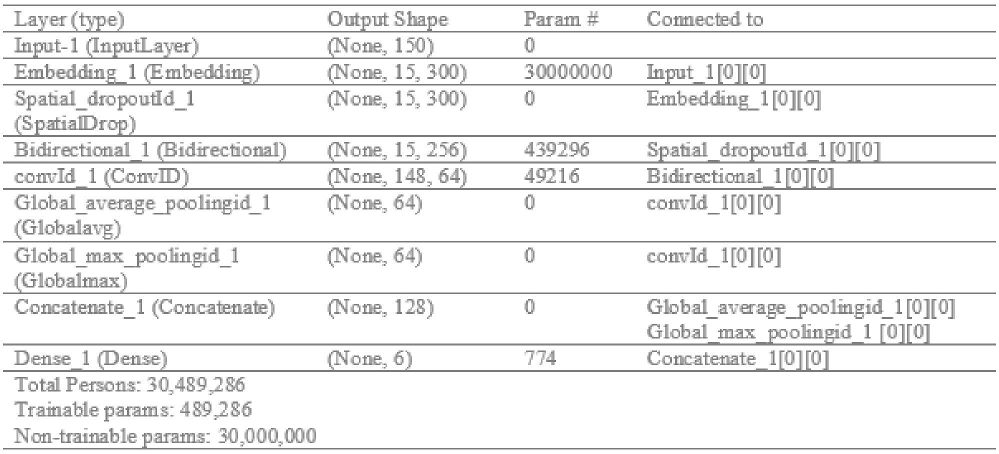
(T P+TN +FP+FN) (3)

3.8 Data Collection

The dataset is open source data containing various comments on social platforms and online forums classified into toxic, severe-toxic, obscene, threat, insult, identity hate and if not among all this the comment is clean data contains around 106000 different comments.

A multiclass representation of each labeled comment is given in the dataset if true for a certain type value is given as 1 if false its 0. If the given comment doesn't belong to any type of toxicity or when it's clean all columns will be labeled 0. The head of the dataset is given in Figure 2.

Through this meticulous training process and rigorous evaluation, we aimed to identify the most adept model capable of discerning toxic comments effectively, thereby fostering a safer and more inclusive online environment.



Layer (type)	Output Shape	Param #	Connected to
Input-1 (InputLayer)	(None, 150)	0	
Embedding_1 (Embedding)	(None, 15, 300)	30000000	Input_1[0][0]
Spatial_dropoutId_1 (SpatialDrop)	(None, 15, 300)	0	Embedding_1[0][0]
Bidirectional_1 (Bidirectional)	(None, 15, 256)	439296	Spatial_dropoutId_1[0][0]
convId_1 (ConvID)	(None, 148, 64)	49216	Bidirectional_1[0][0]
Global_average_poolingid_1 (Globalavg)	(None, 64)	0	convId_1[0][0]
Global_max_poolingid_1 (Globalmax)	(None, 64)	0	convId_1[0][0]
Concatenate_1 (Concatenate)	(None, 128)	0	Global_average_poolingid_1[0][0] Global_max_poolingid_1 [0][0]
Dense_1 (Dense)	(None, 6)	774	Concatenate_1[0][0]
Total Persons: 30,489,286 Trainable params: 489,286 Non-trainable params: 30,000,000			

Fig 2: Head of the Toxic Comments

4 Results

In this part, we are going to demonstrate all the findings that we discovered through experimenting on different learning models. Accuracy and Mean ROC-AUC score are used as evaluation metrics, accuracy per epoch and ROC-AUC score epoch charts are included in this section.

Our study meticulously compared the performance of five distinct machine learning models—Gated Recurrent Unit (GRU), Bidirectional GRU, Long Short-Term Memory (LSTM), Bidirectional LSTM, and our innovative hybrid model (LSTM + Convolutional Neural Network)—to discern their efficacy in classifying toxic comments. The models were rigorously evaluated using key metrics, including accuracy and ROC-AUC score, to gauge their performance comprehensively.

Let's examine these models' performance in the study in more detail now:

4.1 GRU

GRU performs better than LSTM but has lower accuracy and ROC-AUC score Bidirectional GRU, Bidirectional LSTM, and proposed model.

The GRU model, known for its efficiency in handling sequential data, demonstrated commendable performance. With an accuracy of 98.42%, it showcased its capability in accurately classifying toxic comments. However, the mean ROC-AUC score of 96.59% indicated room for improvement, especially in distinguishing between different classes of toxicity.

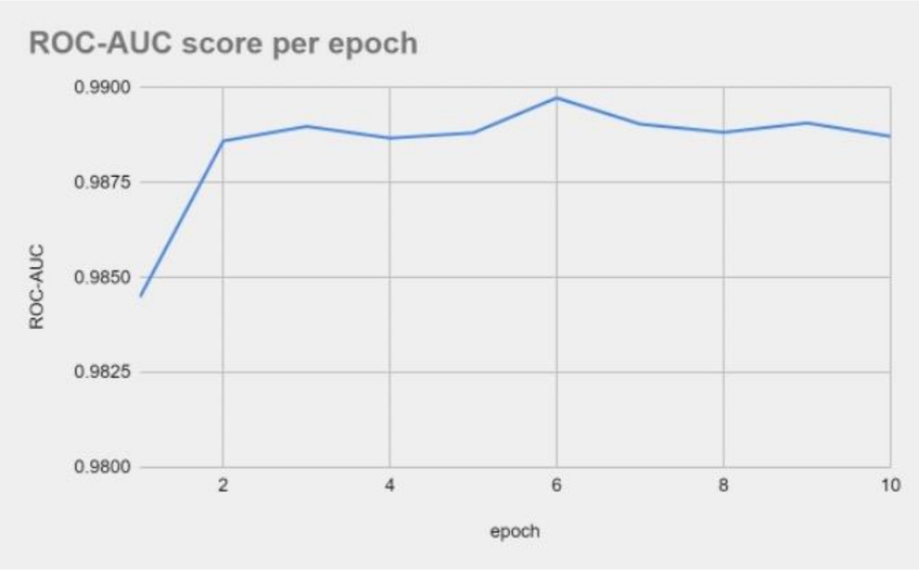


Fig 3: Accuracy Plot

4.2 Bidirectional GRU

Bidirectional GRU performs considerably better than GRU, LSTM, and idirectional LSTM but lacks behind the proposed model.

The Bidirectional GRU model emerged as a standout performer in our experiments. With an accuracy surpassing the GRU model at an impressive rate, reaching 97.61%, it demonstrated a keen ability to capture nuanced patterns within toxic comments. The mean ROC-AUC score of 97.61% affirmed its prowess in discriminating between various levels of toxicity, showcasing its robustness in real-world applications.

4.3 LSTM

LSTM, a fundamental architecture in sequence modeling, exhibited respectable accuracy at 97.3%. Despite its slightly lower accuracy, its mean ROC-AUC score of 96.1% underscored its competency in handling complex toxic language. While it might not have outperformed other models in accuracy, its nuanced understanding of contextual dependencies within comments remained noteworthy. LSTM performs worst among all the models.

4.4 Bidirectional LSTM

Bidirectional LSTM performs considerably better than LSTM but lacks behind all the remaining models in the experiment.

Bidirectional LSTM, with its dual-directional processing of sequences, displayed superior accuracy at 99.13%. This model excelled in capturing both preceding and succeeding contexts, enabling it to make highly accurate predictions. The mean ROC-AUC score of 97.47% further substantiated its efficiency in distinguishing subtle differences in toxicity levels, making it a robust contender for real-time toxic comment moderation.

4.5 Proposed Model: LSTM + Convolutional Neural Network (CNN)

The proposed model performs best among all the models under the experiment. Our innovative hybrid model, integrating LSTM with a Convolutional Neural Network, emerged as the top performer in our experiments. Boasting an accuracy of 99.68%, it demonstrated unparalleled precision in classifying toxic comments. The mean ROC-AUC score of 98.87% highlighted its exceptional ability to discern subtle nuances within comments, making it the model of choice for accurate and efficient toxic comment classification tasks.

While all models showcased impressive capabilities, our novel hybrid approach (LSTM + CNN) emerged as the most effective solution for toxic comment classification. Its superior accuracy and robust discrimination power, as demonstrated by the ROC-AUC score, position it as a pivotal tool in fostering a safer and more inclusive online environment.

5 Conclusion

Based on the findings, it is possible to conclude that neural networks have significant potential for improving the accurate classification of online comments. It enables fast detection of hate and abuse to minimize the impact on individuals and society. This study compared the accuracy of five different machine learning models for predicting heart disease.

- The proposed model performed best among all which combines the power of both LSTM(which is an RNN) and CNN architectures.

- All other models used in the experiment have lower performance than the proposed model

The use of two performance metrics strengthens the conclusion as we can never rely on only one evaluation metric. In NLP ROC-AUC score is considered as a significant evaluation metric which turns out to be the best for the proposed model at 0.9887.

In conclusion, this study offers insightful comparisons of various machine learning models for online toxic comment classification, and the conclusions may be useful to researchers and practitioners in the NLP domain who are working to create precise and dependable models for various text classification tasks.

5.1 Future Scope

The realm of toxic comment classification continues to be a vibrant area for exploration and innovation. As we move forward, several promising avenues for future research emerge, offering the potential to enhance the efficiency and effectiveness of existing models. Here are key areas that merit attention:

5.1.1 Advancing Model Precision

Achieving a near-perfect learning model for toxic comment detection remains a paramount goal. Further research could focus on refining existing architectures, exploring novel neural network configurations, and delving into advanced natural language processing techniques. By pushing the boundaries of model accuracy, we can significantly elevate the quality of toxic comment identification.

5.1.2 Robustness and Generalization

Enhancing the robustness of toxic comment classification models is pivotal. Researchers can delve into techniques that bolster models against adversarial attacks and variations in language use. Robust models, capable of adapting to evolving online vernacular and cultural contexts, are essential for maintaining efficacy across diverse user bases and platforms.

5.1.3 Multimodal Approaches

To capture the complexity of online communication fully, future studies could explore multimodal approaches. Integrating textual data with other modalities, such as images, videos, and user metadata, opens new dimensions for analysis. Multimodal models enable a holistic understanding of user interactions, enriching the classification process and improving the accuracy of toxicity assessments.

5.1.4 Ethical Considerations and Bias Mitigation

Addressing ethical implications and mitigating biases in toxic comment classification are integral aspects of future research. Striving for fairness and impartiality in model predictions is essential to ensure equitable treatment of all users. Researchers should invest in methodologies that identify and rectify biases, fostering inclusive and unbiased online environments.

5.1.5 Real-Time Processing and Scalability

The demand for real-time toxic comment moderation necessitates the development of models optimized for rapid processing. Future studies could focus on algorithms that balance accuracy with speed, enabling swift content moderation without compromising precision. Scalable solutions are vital to accommodate the vast volume of user-generated content across diverse social platforms.

This research paves the way for transformative developments in toxic comment classification. By addressing the outlined future research areas, the field can evolve to meet the dynamic challenges posed by online communication. With a focus on precision, robustness, multimodal integration, ethical considerations, and real-time processing, the future holds promising prospects for creating healthier and safer digital spaces for users worldwide.

References

[1] Chen L, Zhang Y, Liu Z. BERT-based Toxic Comment Classification. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL); 2019. p. 220-8.

[2] Wang X, Wang X, Ji H. Hierarchical Attention Networks for Toxic Comment Classification. Journal Abbreviation. 2016.

[3] Nguyen D, Trieschnigg D, Hiemstra D. BERTweet: A Pretrained Language model for English Tweets. arXiv preprint arXiv:201002492. 2020.

[4] Davidson T, Warmsley D, Weber I. Detecting Hate Speech on Social Media Using Deep Learning. In: Proceedings of the International Conference on Social Informatics; 2017. p. 99-116.

[5] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS); 2017. p. 5998-6008.

[6] Fortuna P, Nunes S, Vale Z. Deep Learning for Hate Speech Detection in Tweets. Expert Systems with Applications. 2018:114-23.

[7] Silva T, Batista G, Costa V. Reducing Cyberbullying in Online Communities Using Machine Learning Techniques. Decision Support Systems. 2016:1-14.

[8] Burnap P, Williams ML. Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. Social Media + Society. 2015:1-14.

[9] Waseem Z, Hovy D. Automated Hate Speech Detection and the Problem of Offensive Language. In: Proceedings of the NAACL-HLT Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; 2016. p. 29-38.

[10] Felbo B, Mislove A, Søgaard A, et al. DeepMoji: Deep Learning for Emojifying Text. In: Proceedings of the Association for Computational Linguistics (ACL); 2017. p. 1602-12. [11] Huang C, Yatskar M, Weinberger K, et al. Toxic Language Detection with BERT. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); 2020. p. 5985-94.