e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

INNOVATIVE FUSION OF LSTM AND BI-GRU NETWORKS FOR ENHANCED HATE SPEECH DETECTION IN SOCIAL MEDIA

Henilsinh Raj*1, Nisharg Nargund*2

*1UG Student, Parul Institute Of Engineering And Technology, Parul University, Vadodara, Gujarat, India.

*2UG Student, Nisharg Nargund, School Of Computer Science And Engineering, Kalinga Institute Of Industrial Technology, Bhubaneswar, Odisha, India.

DOI : https://www.doi.org/10.56726/IRJMETS49147

ABSTRACT

In recent times, the escalating prevalence of hate speech on various social media platforms has emerged as a pressing concern, underscoring the critical need for the development of resilient and effective mechanisms aimed at automated detection. This imperative is not only rooted in the imperative to combat the proliferation of hate speech but also in the broader goal of enhancing the overall user experience within online communities. Against this backdrop, this research endeavors to present an innovative and sophisticated approach to hate speech detection, employing the capabilities of combination of Long Short-Term Memory (LSTM) networks and Bidirectional Gated Recurrent Unit (Bi-GRU)

The assessment of the proposed hate speech detection model is meticulously conducted through the utilization of standard performance metrics, including precision, recall, and F1 score. The outcomes of these analyses reveal a noteworthy enhancement in the model's accuracy in discerning instances of hate speech, thereby attesting to the substantial potential of LSTM-BiGRU networks in effectively addressing the multifaceted challenges posed by the ever-evolving and context-dependent nature of hate speech. This research not only contributes to the technological landscape of automated hate speech detection but also underscores the broader societal significance of fostering online environments that prioritize safety, inclusivity, and a positive user experience.

Keywords: Hatespeech, Detection, LSTM, Machine Learning, NLP, Deep Learning, Natural Language Processing, Bi-GRU, Long Short-Term Memory, Bi-GRU, Bidirectional Gated Recurrent Unit.

I. INTRODUCTION

Modern culture is greatly influenced by social media, which also promotes global connectivity and instantaneous communication. It is an effective instrument for information dissemination that facilitates the quick dissemination of ideas, news, and cultural trends. Social media platforms offer a forum for a range of voices, enabling people to share their thoughts and take part in global conversations. It has completely changed marketing and business by providing a low-cost means for businesses to connect and interact with their target markets. Additionally, social media is essential to activism since it gives movements strength and makes it possible to organize people around political and social issues. It serves as a catalyst for innovation, inspiring people and bringing like-minded people together across boundaries.

While social media has numerous benefits, it also unfortunately serves as a platform for the dissemination of hate speech. Some sites' anonymity can give people the confidence to voice prejudiced opinions, which can aid in the propagation of dangerous ideologies. Social media hate speech is a serious threat to online communities because it creates a hostile atmosphere and occasionally encourages physical violence. Platforms have to struggle to find a way to protect free speech while also preventing hate speech from being amplified.[1] Online abuse and harassment against adults and kids aged 13 to 17 has increased significantly over the past year. Adults reported experiencing online harassment 52% of the time, up from 40% in 2022 and the highest percentage in four years. [2] Victims of the hate material engaged in high levels of online activity. Their attachment to family was weaker, and they were more likely to be unhappy. Online hate is not only in form of sexist, religious, anti-immigrant, racist, verbal attacks based on ethnicity, or political ideology but high-profile individuals are regular targets of trolling and meme cultures. Machine learning techniques such as RNNs can be used to efficiently detect the nature and toxicity level of the content. By leveraging these advanced

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

algorithms, platforms can implement efficient mechanisms to identify and filter out hateful comments. This proactive approach aims to create a more inclusive and user-friendly environment by mitigating the negative impact of toxic content.

The utilization of RNNs in content moderation fosters healthier discussions by automatically detecting and addressing harmful language or behaviour. This enables platforms to prioritize and promote discussions that are solution-oriented, positive, and conducive to a constructive exchange of ideas. As a result, users can engage in online conversations without the hindrance of offensive or inappropriate content, contributing to a more positive and welcoming digital community.

II. RELATED WORKS

In the realm of hate speech detection on social media, numerous research endeavors have been undertaken, employing diverse approaches and methodologies to attain optimal solutions. One notable study by Akanksha et al. [3] harnessed the power of word embeddings coupled with Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) models to identify hate speech on Twitter, achieving an impressive 86% accuracy. On a parallel note, Tin Van Huynh et al. [4] delved into a deep learning methodology anchored in Bidirectional Gated Recurrent Unit (Bi-GRU), LSTM, and Convolutional Neural Network (CNN) architectures, resulting in a commendable 70.5% F1-Score on Vietnamese text. Building on this linguistic context, Hang Thi-Thuy et al. [5] also focused on Vietnamese text, employing Bidirectional Long Short-Term (Bi-LSTM) models and attaining a noteworthy accuracy rate of 71%. In a different stride, Pradeep Kumar Roy et al. [6] introduced a Deep Convolutional Neural Network (DCNN) model, leveraging GloVe embedding vectors to capture tweet semantics through convolution operations. This model showcased remarkable performance, with precision, recall, and F1-score values of 0.97, 0.88, and 0.92, respectively, in the best-case scenario, outperforming existing models.

Extending the exploration, Raza Ali et al. [7] conducted experiments with four distinct variants of BERT, exploiting transfer learning. Their study demonstrated that BERT, xlm-roberta, and distil-Bert yielded encouraging F1-scores of 0.68, 0.68, and 0.69, respectively, in a multi-class classification task. This multifaceted landscape of research showcases the dynamic and evolving nature of hate speech detection methodologies, each contributing valuable insights to the overarching pursuit of fostering safer online environments. [8] The study addresses the pervasive issue of unfiltered content on social media by proposing a multi-aspect hate speech detection approach. Leveraging pre-trained BERT models and combining them with Deep Learning models, including Bidirectional LSTM and Bidirectional GRU on GloVe and FastText word embeddings, the researchers achieve a notable 98.63% ROC-AUC score in enhancing hate speech detection on social media, demonstrating the efficacy of their comprehensive ensemble learning strategy.

III. METHODOLOGY

Proposed Approach

To address the challenges of Hate speech detection, we present a novel approach that leverages recurrent neural networks (RNNs) to capture temporal dependencies within the data. Our proposed model combines the strengths of Long Short-Term Memory (LSTM) and Bidirectional Gated Recurrent Unit (Bi-GRU) layers, aiming to achieve optimal performance.

In the initial stages, the input data undergoes text vectorization, converting textual information into a format suitable for neural network processing. Following this, an embedding layer is applied to capture semantic relationships between words. The LSTM layer, with return sequences enabled, is employed to capture long-range dependencies in the sequential data.

A key innovation lies in the utilization of Bidirectional Gated Recurrent Unit (Bi-GRU) layers. By processing input data in both forward and backward directions simultaneously, the model gains a richer understanding of the contextual information present in the sequences. This bidirectional approach enhances the model's ability to capture nuanced patterns and dependencies within the data.

To mitigate the risk of overfitting, a dropout layer with a dropout rate of 0.5 is introduced. This regularization technique enhances the generalization capabilities of the model, ensuring robust performance on unseen data.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

The output layer employs a softmax activation function, facilitating the classification of input sequences into one of three categories.

Getting One with Data

[9] For the research purpose we have used '**Hate Speech and Offensive Language Dataset'** dataset by ANDRII SAMOSHYN. Further we have divided the dataset into training and testing 70% of data is reserved for training purpose and 30% of dataset is reserved as testing(validation) dataset. The dataset is well shuffled for avoiding any biasness toward any class.

Model Architecture

The architecture of our proposed model is summarized as follows:

Input Layer: Accepts sequential input data.

Text Vectorization: Converts textual information into a format suitable for neural network processing.

Embedding: Captures semantic relationships between words.

LSTM Layer: Processes input sequences and captures long-range dependencies.

Bidirectional GRU Layer: Enhances contextual understanding by processing data in both forward and backward directions.

Dropout Layer: Mitigates overfitting by randomly dropping a fraction of connections during training.

Output Layer: Produces classification results using softmax activation.

In our approach, we utilize the sparse categorical cross-entropy as the loss function, suitable for multi-class classification (0 - hate speech 1 - offensive language 2 – neither) tasks where the target labels are integers. The Adam optimizer is chosen for its adaptive learning rate capabilities, which often leads to efficient convergence during training. For model evaluation, accuracy is selected as the metric of interest, providing insights into the model's overall classification performance.

These choices aim to strike a balance between model expressiveness, training stability, and ease of interpretation, contributing to the robustness of our proposed methodology.

The training process consists of five epochs, during which the model learns to map input sequences to their corresponding labels. The training dataset (**train_sentences**) and labels (**train_labels**) are utilized for this purpose. Additionally, model performance is monitored on the validation dataset (**val_sentences** and **val_labels**) during each epoch, providing insights into the model's generalization capabilities.

IV. RESULTS AND DISCUSSION

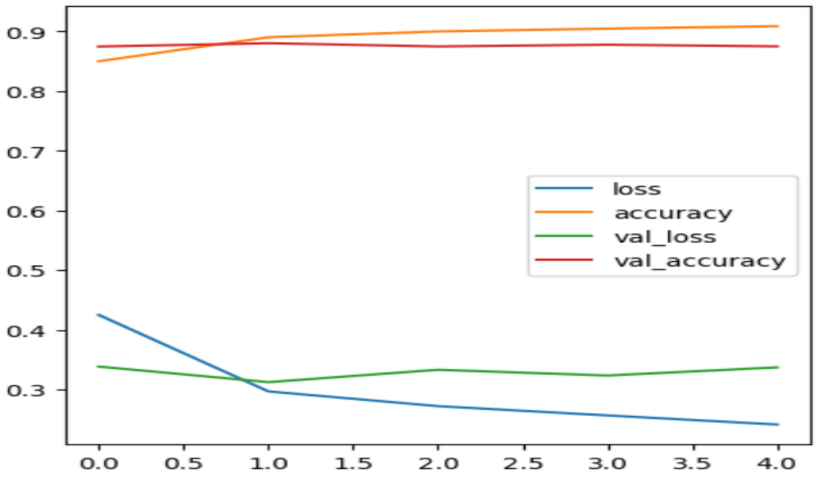

Fig 1: Loss curves

Above graph indicates the high accuracy of the proposed model on both training and validation datasets we can also observe the gradual downfall of losses with each passing epoch.

Accuracy = TP / ((TP + FN)) Training accuracy refers to the accuracy of the model on the training data during the training process. As the model is trained on the training data, the training accuracy increases, indicating that the model is getting better at fitting the training data Validation accuracy, on the other hand, refers to the

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:02/February-2024 Impact Factor- 7.868 www.irjmets.com

accuracy of the model on a validation set of data that is not used for training. The validation accuracy gives an estimate of how well the model will generalize to new, unseen data

Table 1. Metrics

SN.	Metrics	Results
1	Accuracy	90.50
2	precision	86.49
3	recall	87.50
4	F1-Score	86.51

Accuracy is the overall correct predictions made by the model divided by the total number of predictions. It is calculated using the formula Accuracy = (True Positives + True Negatives) / (True Positives + True Negatives + False Positives + False Negatives)

Precision = True Positives / (True Positives + False Positives)

The recall represents the ability of the model to correctly identify positive instances out of all actual positive instances. It is calculated using the formula:

Recall = True Positives / (True Positives + False Negatives)

The F1-score is a measure of a model's accuracy that takes into account both precision and recall. It is calculated using the following formula F1-score = 2 * (precision * recall) / (precision + recall) The precision represents the ability of the model to correctly identify positive instances out of all instances it predicted as positive

V. CONCLUSION

Our proposed model achieved commendable results, with an overall accuracy of 90.51% on the training set and a robust validation accuracy of 87.51%. The efficacy of our hate speech detection algorithm is further demonstrated by the precision, recall, and F1-score metrics. The model's accuracy in identifying hate speech is demonstrated by its 86.49% precision, which minimizes false positives. Concurrently, the 87.51% recall rate indicates that the model is capable of accurately identifying a considerable proportion of real hate speech incidents, reducing false negatives. The harmonic mean of recall and precision, or the F1-score, is an impressive 86.51%. However, as with any machine learning model, there are areas for future exploration and refinement. Ongoing efforts could focus on expanding the dataset to enhance the model's exposure to diverse hate speech patterns. Additionally, fine-tuning hyperparameters and exploring more advanced architectures might further optimize the model's performance.

ACKNOWLEDGEMENT

We thank Jayveersinh Raj for reviewing early versions of this paper and for helpful feedback on this work.

VI. REFERENCES

[1] Online Hate and Harassment: The American Experience 2023, Anti-defamation League

[2] Oksanen, Atte & Hawdon, James & Holkeri, Emma & Näsi, Matti & Räsänen, Pekka. (2014). Exposure to Online Hate among Young Social Media Users. 10.1108/S1537-466120140000018021. Technology: C Software & Data Engineering, Volume 20, Issue 2, No. 2020, pp 12-20

[3] Bisht, A., Singh, A., Bhadauria, H.S., Virmani, J., Kriti (2020). Detection of Hate Speech and Offensive Language in Twitter Data Using LSTM Model. In: Jain, S., Paul, S. (eds) Recent Trends in Image and Signal Processing in Computer Vision. Advances in Intelligent Systems and Computing, vol 1124. Springer, Singapore. https://doi.org/10.1007/978-981-15-2740-1_17 Gyusoo Kim and Seulgi Lee, "2014 Payment Research", Bank of Korea, Vol. 2015, No. 1, Jan. 2015.

[4] Tin Van Huynh, Duc-Vu Nguyen, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen University of Information Technology, Vietnam National University Ho Chi Minh City 16521827@gm.uit.edu.vn,{vund, kietnv, ngannlt, anhngt}@uit.edu.vn

[5] Hang Thi-Thuy Do, Huy Duc Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen and Anh Gia-Tuan Nguyen University of Information Technology, VNU-HCM {16520339, 16520508}@gm.uit.edu.vn, {kietnv, ngannlt, anhngt}@uit.edu.vn

[6] P. K. Roy, A. K. Tripathy, T. K. Das and X. -Z. Gao, "A Framework for Hate Speech Detection Using Deep Convolutional Neural Network," in IEEE Access, vol. 8, pp. 204951-204962, 2020, doi: 10.1109/ACCESS.2020.3037073. keywords: {Feature extraction; Social networking (online); Blogs; Data models; Convolutional neural networks; Voice activity detection; Logistics; Convolutional neural network; hate speech; LSTM; Tf-Idf; Twitter},

[7] Ali, Raza & Farooq, Umar & Arshad, Muhammad & Shahzad, Waseem & Beg, Mirza. (2022). Hate speech detection on Twitter using transfer learning. Computer Speech & Language. 74. 101365. 10.1016/j.csl.2022.101365.

[8] Mazari, A.C., Boudoukhani, N. & Djeffal, A. BERT-based ensemble learning for multi-aspect hate speech detection. Cluster Comput (2023). https://doi.org/10.1007/s10586-022-03956-x

[9] Hate Speech and Offensive Language Dataset ANDRII SAMOSHYN ; Kaggle.