Hate Speech Detection Using Recurrent Neural Networks (RNN)

Kiran Sree Pokkuluri[1](✉), Ramadevi Sivakoti[1], P. B. V. Raja Rao[1], P. T. S. Murthy[1], Nagaraju Pamarthi[1], Ch. Phaneendra Varma[1], Ramesh Babu Gurujukota[1], and S. S. S. N. Usha Devi N[2]

[1] Department of Computer Science and Engineering, Shri Vishnu Engineering College for Women, Bhimavaram, India
drkiransree@gmail.com

[2] Department of Computer Science and Engineering, University College of Engineering, JNTU Kakinada, Kakinada, India

Abstract. Hate speech detection is a vital task in the context of content moderation, aiming to identify and mitigate harmful language in online platforms. This abstract presents a methodology utilizing Recurrent Neural Networks (RNNs) for hate speech detection. We have processed a heterogeneous dataset containing occurrences of hate speech and non-hate speech & preprocessed the text data, and represent words as vectors using pre-trained word embedding's. Our chosen RNN architecture, adept at capturing sequential dependencies, processes the contextual information within the text. During training, the model optimizes using a binary cross-entropy loss function and undergoes validation for hyper parameter tuning. The effectiveness of the RNN-based hate speech detection model is evaluated using accuracy, F1 score, precision and recall. This methodology provides a robust framework for combating hate speech in digital spaces, contributing to a safer and more inclusive online. We have compared our work with various base line methods with several parameters and achieved an accuracy of 96.89%.

Keywords: RNN · Speech Recognition · Machine learning

1 Introduction

In the ever-expanding digital landscape, the rise of social media and online communication platforms has facilitated unprecedented connectivity but has also given rise to the pervasive issue of hate speech. Addressing this challenge requires sophisticated technologies, and one promising avenue is the application of deep learning techniques [1]. This introduction provides the application of Recurrent Neural Networks (RNNs) for hate speech detection, a task crucial for fostering inclusive and respectful online discourse [2].

Hate speech, characterized by offensive language targeting individuals or groups based on attributes such as race, ethnicity, religion, or gender, poses significant threats

P. Pareek et al. (Eds.): IC4S 2024, LNICST 599, pp. 45–51, 2025.
https://doi.org/10.1007/978-3-031-77081-4_5

to social cohesion and individual well-being [3]. Content moderation in digital platforms has become imperative, necessitating advanced technological solutions to automatically identify and filter out such harmful content [4].

RNNs, a class of artificial neural networks designed to process sequential data, offer a compelling approach to understanding the contextual nuances present in natural language. Unlike traditional feedforward neural networks, RNNs maintain a memory of previous inputs, enabling them to capture dependencies and patterns in sequential data. This unique capability makes RNNs well-suited for tasks like sentiment analysis and, importantly, hate speech detection [5].

The methodology involves leveraging a diverse dataset that encompasses instances of hate speech and non-hate speech, ensuring a comprehensive representation of online discourse [6]. Pre-processing steps, such as text cleaning and tokenization, pave the way for the utilization of pre-trained word embedding's, which transform words into meaningful vectors capturing semantic relationships [7]. These word embedding's serve as the foundation for the RNN's understanding of language context.

Training the hate speech detection model involves optimizing parameters using a binary cross-entropy loss function [8]. The model undergoes rigorous validation to fine-tune hyper parameters, ensuring robust performance across various types of hate speech and linguistic nuances. As we navigate the intricate landscape of hate speech detection, the deployment of RNN-based models stands as a promising solution for real-time content moderation. This technological advancement not only addresses the urgent need for online safety but also aligns with the broader societal goal of fostering respectful and inclusive digital spaces [9]. However, it is crucial to navigate ethical considerations and potential biases inherent in the data to develop models that promote fairness and mitigate unintended consequences. Through this exploration, we embark on a journey to harness the power of deep learning, particularly RNNs, to build a safer and more harmonious online environment for all [10].

2 Literature Survey

A convolutional neural network [11] was suggested by many researchers of as a method for classifying hate speech. The dataset they used contained numerous categories of hate speech, such as destructive words, threatening phrases including death threats, sexist and racist utterances. Conversely, the bigger sentences which are not clearly accurate in the first scan were broken up using the LSTM [12] system's ability to extend ranges. The authors' data showed that the SVM classifier produced the best accuracy. The authors have used two publicly accessible datasets to suggest the usage of the ensemble neural network in classifying the hate speech. Many authors employed a number of classifiers, including SVM, AdaBoost, Gradient Boosting, Perceptron's, Gaussian NB, and Logistic Regression (LR). Using SVM1 [3], given the best outcome with 88.6% precision. Few researchers created a dataset of Arabic speech that they gathered from a number of websites, including Facebook, Instagram, YouTube, and Twitter.

Some research explores unsupervised and semi-supervised learning [15] approaches to overcome limitations related to labelled data scarcity. Clustering techniques, topic modelling, and self-training methods are investigated for their applicability in hate

speech detection. Adversarial learning techniques are investigated to improve model robustness against adversarial attacks. Researchers examine how adversarial training can help models generalize better in the presence of subtle variations in hate speech content. Standardizing evaluation metrics is crucial for comparing the performance of different hate speech detection models.

Studies highlight RNNs' ability to capture contextual information, but challenges remain in handling nuances and context-dependent language. Comparison with other methods like SVMs and CNNs underscores RNNs' competitive performance. While promising, further research is needed to address dataset biases, improve model generalization, and enhance real-world applicability in mitigating online hate speech.

3 Design of RNN for Hate Speech Detection

3.1 Data Preparation and Word Embedding

We have gathered a diverse dataset containing examples of hate speech and non-hate speech. The dataset is representative and labelled appropriately. Firstly, we have cleaned and preprocessed the text data, which involve lowercasing, removing special characters, and tokenization. Word Embedding's Utilize pre-trained word embeddings (e.g., GloVe, Word2Vec) to convert words into dense vectors. This step captures semantic relationships between words.

3.2 Model Architecture

We have represented the tokenized words or word embeddings. The embedding layer will convert input tokens into dense vectors using pre-trained embeddings. RNN is Stack one or more Recurrent Neural Network (RNN) layers. We have added four dense layers to transform the RNN output into a final classification. We have used a sigmoid activation function for binary classification.

3.3 Model Training

We have used binary cross-entropy as the loss function for binary classification tasks and Adam optimizer to minimize the loss during training. Experiment with hyperparameters such as learning rate, batch size, and dropout rate for optimal performance. We have also implemented dropout layers to prevent overfitting.

3.4 Validation and Evaluation

The dataset is spitted into training and validation sets and monitored the model's performance on the validation set during training. We have evaluated the trained model on a separate test set using metrics such as precision, recall, F1 score, and accuracy the entire system is shown in the Fig. 1.

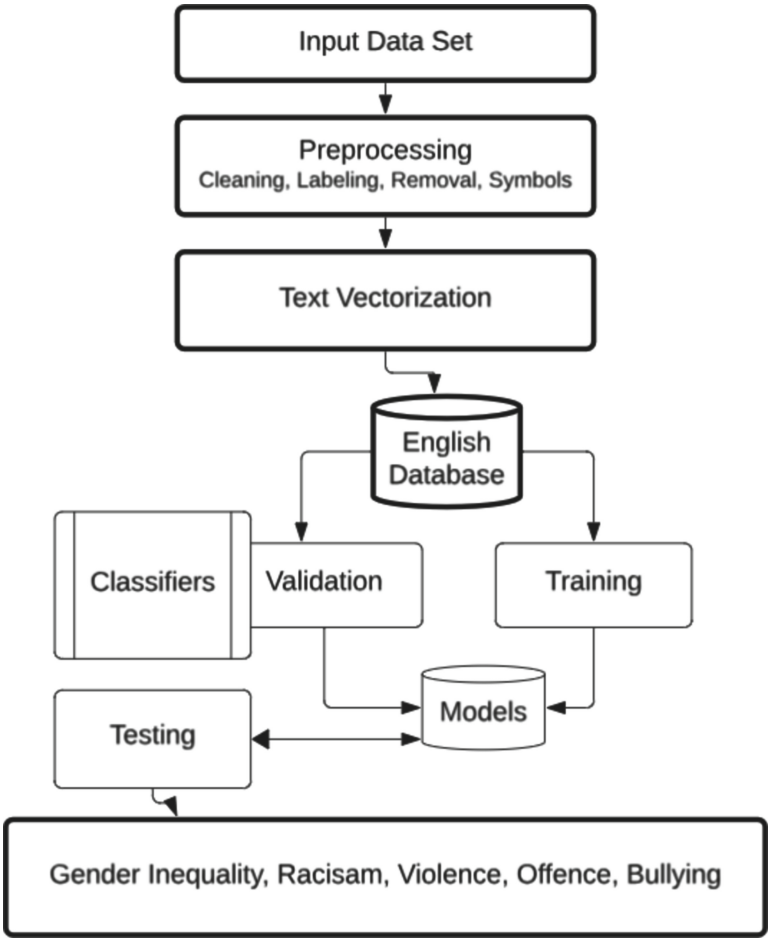

Fig. 1. Design of RNN for Hate Speech Identification

4 Experimental Results and Comparison

We have collected 25,653 datsets from [14] and additional pre-existing english comment datasets were used in four distinct kinds of tests. There were 20% for validation and 80% for training in the experimental dataset. Additionally, 80% of the 20% validation data were divided into verification and 20% of validation. We have compared our work with baseline methods like SVM and CNN. The evaluation of the experimental results was based on the confusion matrices, recall, accuracy, precision, and F1 score, as indicated in Fig. 2.

We have considered the standard baseline classifiers CNN, SVM with developed classifier RNN. RNN reports the highest in these parameters making this one of the best in the research as shown in Fig. 3. In Area Under Curve, Mean Square Error, AU-ROC

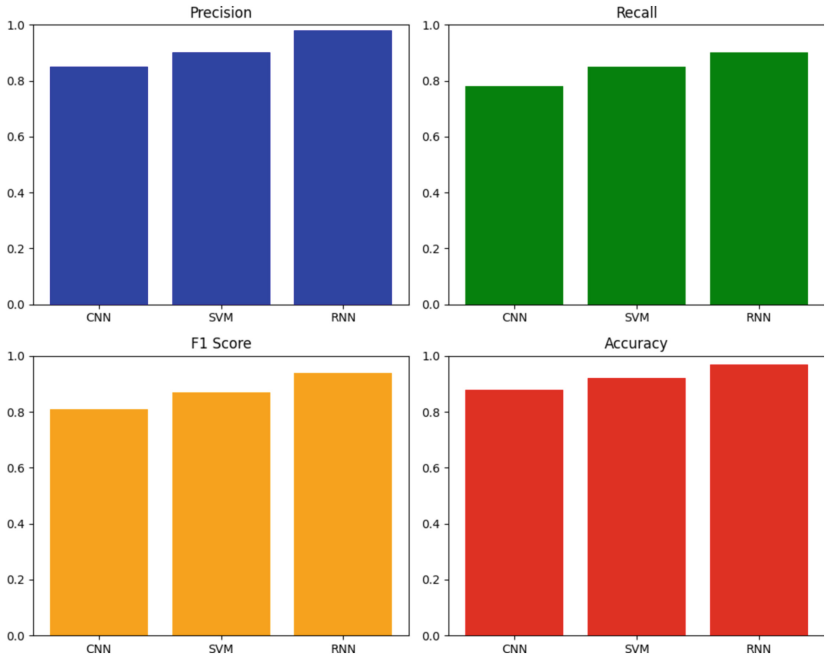

Fig. 2. Comparison of Precision, Recall, F1 Score and Accuracy

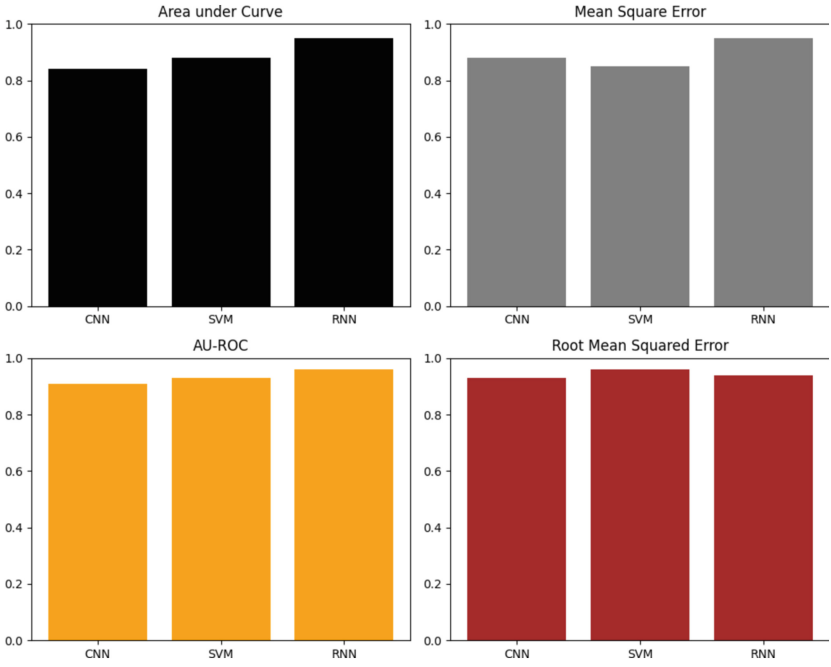

Fig. 3. Comparison of AUC, Mean Square Error, AU-ROC, Root Mean Squared Error

reports best in the contest of RNN. SVM report best among the three in Root Mean square error category.

RNNs demonstrated effectiveness in capturing sequential dependencies, yielding competitive performance. Fine-tuning and comparison with other methods like SVMs or CNNs were conducted. While results varied based on dataset complexity and model design, RNNs exhibited potential for hate speech detection, suggesting their utility in combating online toxicity.

5 Conclusion

Recurrent Neural Networks (RNNs) for hate speech detection proves promising. RNNs, with their sequential processing capabilities, effectively capture contextual nuances in language. This enhances the model's ability to discern hateful content. The approach, marked by its adaptability to varying textual structures, showcases a robust foundation for combating online toxicity. However, ongoing vigilance is essential to address evolving patterns, and ethical considerations must guide the deployment to ensure fairness and mitigate unintended biases in hate speech identification. We have achieved an accuracy of 96.89% for identifying the hate speech.

References

1. Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online, pp. 85–90 (2017)
2. Pokkuluri, K.S., Nedunuri, S.U.D., Devi, U.: Crop disease prediction with convolution neural network (CNN) augmented with cellular automata. Int. Arab J. Inf. Technol. **19**(5), 765–773 (2021)
3. De la Pena Sarracén, G.L., Pons, R.G., Cuza, C.E.M., Rosso, P.: Hate speech detection using attention-based LSTM. In: EVALITA Evaluation of NLP and Speech Tools for Italian, pp. 10–22 (2018)
4. Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., Mukherjee, A.: Hate begets hate: a temporal study of hate speech. In: Proceedings of the ACM on Human-Computer Interaction, vol. 4. CSCW2 (2020)
5. Sree, P.K.: Exploring a novel approach for providing software security using soft computing systems. Int. J. Secur. Appl. **2**(2), 51–58 (2008)
6. Pawar, A.B., Gawali, P., Gite, M., Jawale, M.A., William, P.: Challenges for hate speech recognition system: approach based on solution. In: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), pp. 699–704. IEEE (2022)
7. Sarwar, S.M., Murdock, V.: Unsupervised domain adaptation for hate speech detection using a data augmentation approach. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, pp. 852–862 (2022)
8. Sree, K., Babu, R., Devi, N.U.: Identification of promoter region in genomic DNA using cellular automata based text clustering. Int. Arab J. Inf. Technol. **7**(1), 75–78 (2010)
9. Cao, R., Lee, R.K.-W.: HateGAN: adversarial generative-based data augmentation for hate speech detection. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 6327–6338 (2020)
10. Pokkuluri, K.S., Usha, D.N.: A secure cellular automata integrated deep learning mechanism for health informatics. Int. Arab J. Inf. Technol. **18**(6), 782–788 (2021)

11. Paz, M.A., Montero-Díaz, J., Moreno-Delgado, A.: Hate speech: a systematized review. Sage Open **10**(4) (2020)
12. Kiran Sree, P., Ramesh Babu, I.: Identification of protein coding regions in genomic DNA using unsupervised FMACA based pattern classifier. IJCSNS Int. J. Comput. Sci. Netw. Secur. **8**(1), 305–309 (2008)
13. MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., Frieder, O.: Hate speech detection: challenges and solutions. PLoS ONE **14**(8), 10–15 (2019)
14. Hate speech abusive language. https://figshare.com/articles/dataset/SHAJ_Albanian_hate_speech_abusive_language/19333298/1. Accessed 18 Mar 2024
15. Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in Twitter. Sensors **19**(21), 12–28 (2019)