International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

AUTOMATED HATE SPEECH DETECTION USING RECURRENT NEURAL NETWORK

Mr. Atharva Khurpe[1], Mr. Aman Shaikh[2], Mr. Swanand Muley[3], Mr. Imaad Patel[4],
Prof. Rutika Shah[5]

[1,2,3,4] UG Student, [5]Asst. Professor, Dept. of Computer Engineering,

Trinity College of Engineering and Research, Pune, Maharashtra.

Abstract - The spread of online entertainment and information sharing has indeed brought significant advances to humanity. However, it has also created various challenges, such as the spread of derogatory content through means such as voicemail. Addressing this emerging problem requires continued consideration, especially in virtual environments, where innovative approaches combining design methods and artificial intelligence are used to identify and filter such content from various data sets. While efforts are being made to combat this problem, there is still a gap in design strategies for that purpose and a comprehensive overview of artificial intelligence algorithms. This includes the need for transparent evaluation criteria and publicly available datasets to benchmark performance and promote collaboration within the research community.
Keywords — Deep Learning, Recurrent Neural Network(RNN), Hate Speech Detection.

I. INTRODUCTION

The spread of online entertainment and information sharing has indeed brought significant advances to humanity. However, it has also created various challenges, such as the spread of derogatory content through means such as voicemail. Addressing this emerging problem requires continued consideration, especially in virtual environments, where innovative approaches combining design methods and artificial intelligence are used to identify and filter such content from various data sets. While efforts are being made to combat this problem, there is still a gap in design strategies for that purpose and a comprehensive overview of artificial intelligence algorithms. This includes the need for transparent evaluation criteria and publicly available datasets to benchmark performance and promote collaboration within the research community. The next downside, however, is the deepening "can't go on" debate. Disrespect for online entertainment can be communicated through posts on Facebook and other online gatherings, tweets on Twitter, YouTube comments and recordings, etc. Due to the duty of confidentiality, customers can create fake profiles and add to their identity without revealing their identity. Individuals use these records to spread cruelty, create confusing effects online, and deceive others. Digital bullying is also one of the biggest problems in virtual entertainment. While these stages of online entertainment also offer administrative guidelines and rules that can lead to record suspensions, online disdain still exists and seems to be constantly evolving. At these stages, the discovery calculations were made by different AI models, and the different structures are not contradictory. In any case, these calculations can be ignored more often. Thus, there is a growing need to find better answers to this problem. Most persistent contempt detection techniques have focused more on paper data such as posts, comments, or tweets. In any case, individuals can also make derogatory recordings and post them on recording sites. At these stages, the discovery calculations were made by different AI models, and the different structures are not contested. In any case, these calculations can be ignored more often.

A. Need of study:

1. Dealing with Online Bullying: Hate speech and online harassment have become more common on social media platforms and other online forums. Automated detection systems can help reduce the spread of hateful content and protect users from harassment or discrimination.
2. Protecting vulnerable communities: Hate speech often targets marginalized groups based on factors such as race, ethnicity, religion, gender, sexual orientation, or disability. Identifying and removing hate speech from online platforms can help protect vulnerable communities from the harmful effects of discrimination and bigotry.
3. Maintain online civility: Hate speech undermines the quality of online discussions and can contribute to the growth of misinformation and polarization. By automatically detecting and filtering hateful content, we can foster a more civil and respectful online environment that encourages constructive dialogue and the exchange of ideas.

II. RELATED WORK

In recent years, starting around 2019 and continuing into 2020 and beyond, there has been a significant increase in research focusing on the automatic recognition of angry speech using recurrent neural networks (RNN). These studies investigated various aspects of hate speech detection, from model architectures to performance evaluation methods. Researchers investigated the complexity of hate speech detection using RNN-based approaches and compared them with alternative machine-learning techniques. They sought to reveal the effectiveness of RNNs in picking up subtle linguistic nuances indicative of hate speech, which improved our understanding of how computer models can distinguish hateful content from textual data. In addition, research has begun to understand. how different types of hate speech, such as racial, gender-based, or religiously motivated hate speech, affect cognitive processes and performance in hate speech detection systems. This line of research not only sheds light on the various manifestations of hate speech but also emphasizes the importance of designing detection systems that are sensitive to these variations.

According to the evolving linguistic landscape of online discourse, researchers have tried to tune RNN models to fit multilingual or code-switching environments. This adaptability reflects the recognition of the different linguistic contexts in which hate speech can occur and the need to identify the functionality of systems across language boundaries.

In addition, efforts have been made to mitigate the challenges of data imbalance and generated noise. according to the comments. hate speech in datasets. To address these issues, new data processing and augmentation techniques have been proposed to improve the robustness and generalizability of hate speech detection models.

At the same time, progress has been made in improving evaluation methods to comprehensively evaluate hate speech detection performance. Systems metrics such as precision, recall, F1 scores, and area under the receiver operating curve (AUC-ROC) were used to provide a nuanced understanding of system performance on various dimensions.

Finally, researchers began to weigh in critically. ethical implications of automatic anger detection. Concerns about impartiality, fairness, and censorship have been explored, emphasizing the need to develop open and accountable identification systems that adhere to principles of justice and fairness.

In general, increasing research in this area highlights a growing interest and principles of fairness. The versatility of using RNNs for automatic hate detection, as well as the interdisciplinary nature of addressing the complex challenges of hate detection in digital environments.

III. ALGORITHM

Recurrent Neural Network (RNN):

RNN architectures such as Long Short Term Memory - LSTM or Gated Recurrent Unit - GRU can convey information while managing vanishing/bursting gradients. Because it is typically a bidirectional RNN that responds with a 1-second delay, it allows the model to distinguish past and future dependencies at a specific point in the audio, as shown in Figure.It is ready to handle the RNN detection problem. RNN or Repetitive Neural Network is a technique used for careful memory.

Suppose you remember what happened in the previous scene while reading a book or watching a movie. According to how RNN works, they collect past data and use it to manage current data. A disadvantage of RNNs is that they cannot remember long-term situations due to vanishing bias. The RNN is specifically designed to avoid long-term dependency problems.

The first segment specifies whether the previoustimestamp data is returned or is irrelevant and can be ignored. The cell will try to transfer fresh data to that cell in the next section. The cell sends the updated data from the current timestamp to the next timestamp of the third and final part. These three components of an RNN cell are called input paths. The first part is called carelessness, the second is the door of knowledge, and the last is the door of result.

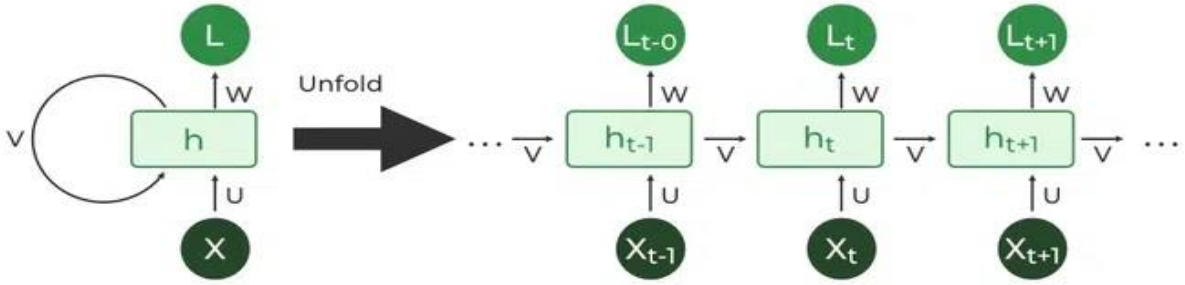
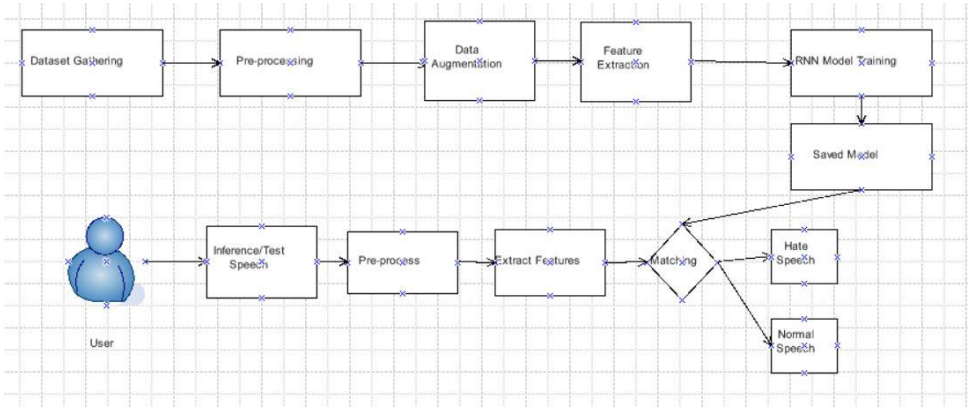


Fig: Recurrent Neural Network

IV. PROPOSED WORK

1. Dataset preparation: Collect diverse datasets that contain examples of hate speech and ensure representation of different forms of hate speech (e.g. racial, sexual, religious).
2. Preprocessing: Clean and preprocess data, including text normalization, tagging, and noise reduction. . or irrelevant information.
3. Model architecture selection: Choose appropriate RNN architectures such as Long Short Term Memory (LSTM) or Gated Recurrent Unit (GRU), which are known for efficiently capturing sequential patterns.
4. Feature extraction: Feature extraction from text data to capture. semantic information corpora or other representational learning techniques.
5. Training models: Train the RNN model on a prepared dataset, and fine-tune hyperparameters such as learning rate and set size to optimize performance.
6. Evaluation metrics: Evaluate the model. performance using standard metrics such as precision, recall, F1 score, and area under the ROC curve (AUC-ROC).

7. Cross-validation: Perform cross-validation to ensure model reliability and generalization across different subsets of the dataset.
8. Hyper-parameter tuning: Perform model performance optimization to further optimize tuning hyperparameters to calculate factors such as the number of levels, hidden units and dropout rates.
9. De-policing and fairness: Apply techniques to reduce bias and ensure the fairness of incitement detection while considering the ethical implications. algorithmic decisions.
10. Deployment and Monitoring: Deploy the trained model in real-world applications, continuously monitor its performance, and update it as needed to adapt to changing language trends and emerging forms of hate speech..timestamp data is returned or is irrelevant and can be ignored. The cell will try to transfer fresh data to that cell in the next section. The cell sends the updated data from the current timestamp to the next timestamp of the third and final part. These three components of an RNN cell are called input paths. The first part is called carelessness, the second is the door of knowledge, and the last is the door of result.



V. ADVANTAGES AND DISADVANTAGES

A. Advantages:

- Promotes online safety by identifying and reducing harmful content.
- It helps protect vulnerable communities from discrimination and harassment.
- Supports content monitoring by automating the detection process.
- It helps promote a more inclusive and respectful online environment.

B. Disadvantages:

- Risk of false positives, leading to censorship of legitimate speech.
- Difficulties accurately capturing nuances of language and context, leading to errors in detection.
- Algorithmic biases can cause biased or discriminatory results.
- Ethical privacy concerns. and monitoring, especially e-mail related to monitoring.

VI. ACCURACY

Many techniques have been used to tackle hate speech identification, such as Convolutional Neural Networks (CNN), Support Vector Machines (SVM), and Term Frequency-Inverse Document Frequency (TF-IDF) representations. CNN models have demonstrated increased accuracy by automatically learning hierarchical text representations, whilst SVM models provide respectable performance with high interpretability. TF-IDF, on the other hand, can offer a solid baseline for hate speech identification jobs even though it is a straightforward algorithm. But in comparison to more advanced models like CNNs, its performance might be constrained. The effectiveness of each of these techniques varies, and the quality of the dataset, feature representations, and model architecture are some of the variables that affect how accurate each method is at detecting hate speech.

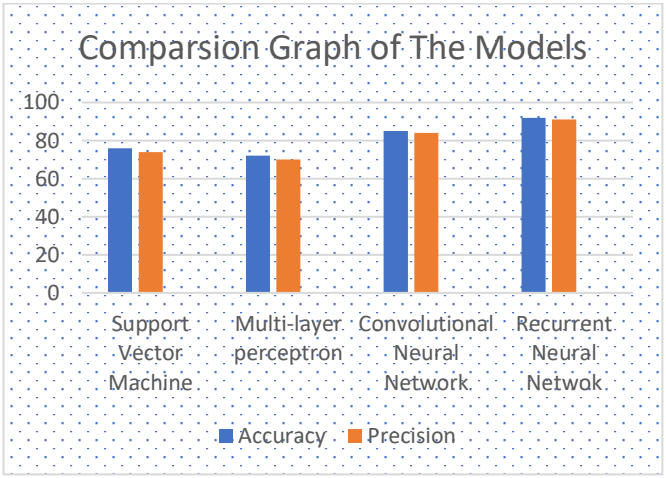
Numerous algorithms, such as Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and TF-IDF, have been used in the field of hate speech identification. The accuracy of each method in detecting hate speech in text data has varied. Traditional machine learning algorithms like Support Vector Machines (SVM) have shown some accuracy in identifying hate speech, especially when paired with well designed characteristics. In contrast, CNNs have demonstrated encouraging performance in identifying patterns and local dependencies in text, outperforming SVMs in this regard. Furthermore, even though they are less complex, methods like TF-IDF have shown successful in identifying the significance of individual words in a document, which improves the accuracy of hate speech identification. Nevertheless, a number of variables, such as feature engineering, dataset quality, and complexity, affect final performance.
So in conclusion SVM gives the lowest accuracy while RNN gives the highest till date

- Support Vector Machine(SVM):-
  Accuracy - 76%
  Precision -74%

- Multi-layer perceptron:-
  Accuracy -72%
  Percision -70%

- Convolutional Neural Network(CNN):-
  Accuracy - 85%
  Precision - 84%

- Recurrent Neural Network(RNN):-
  Accuracy - 92%
  Precision - 91%



VII. RESULT

Fig 6.1 Output 1

Fig 6.2 Output 2

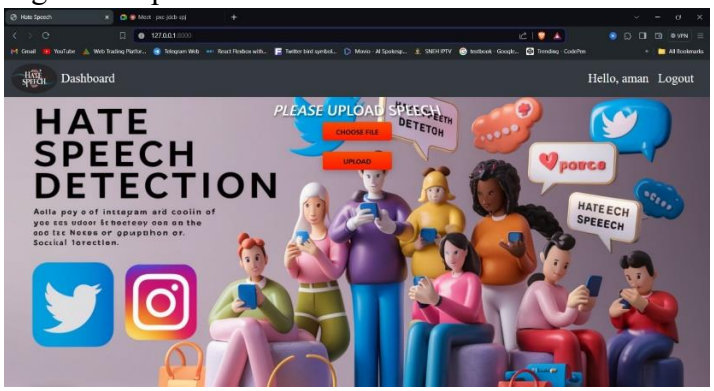


Fig 6.3 Output 3

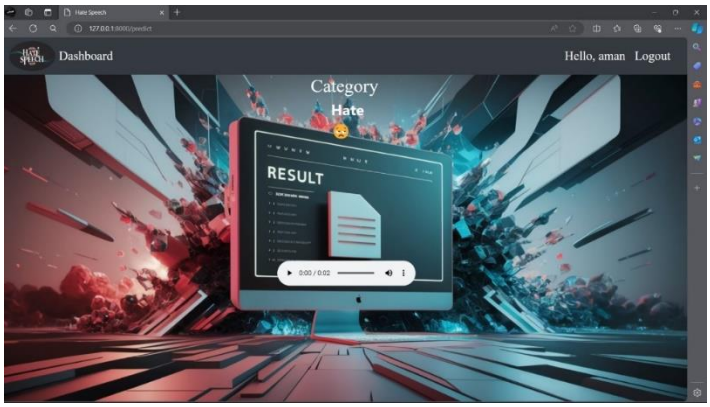


Fig 6.4 Output 4

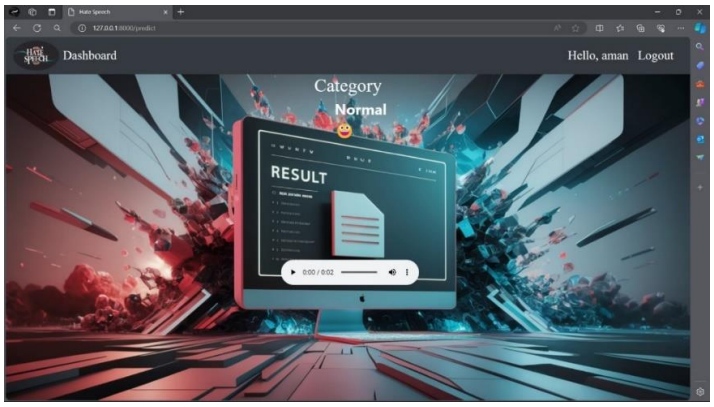


Fig 6.5 Output 5

VIII.CONCLUSION

This work aimed to establish a web enhancement technique and constantly coordinate with simulated intelligence estimations. We employed MFCC for feature extraction and RNN, a deep learning architecture, for classification in a published study. At 300 epochs, we obtain an accuracy of 87.31%. Acknowledgment and dislike to scorn talk was the final goal.The outcomes showed an increased level of classifier accuracy in addition to congruency of the AI estimations in web apps. It has been demonstrated that the brain network estimates communicate frequently astonishing results, even when tested in a basic feed-forward network library.

Future Scope:-

The potential applications of Recurrent Neural Networks (RNNs) for hate speech detection are enormous and bright. In order to better capture the intricate subtleties of hate speech in online communication, RNN designs like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) are constantly being enhanced as technology improves. A primary focus for improvement is augmenting RNNs' comprehension of context and sarcasm, which are prevalent in hate speech. Furthermore, including multimodal input into RNN-based models—such as text, images, and audio—could greatly increase detection accuracy. Moreover, the implementation of RNNs in real-time monitoring systems for online forums and social media platforms can help with the prompt identification and removal of hate speech, promoting safer online communities.

REFERENCES

[1] Waseem, Z. and Hovy, D. "Hateful individuals or hateful symbols? The Association for Computational Linguistics published a paper in the Proceedings of NAACLHLT in 2016 that included predictive variables for hate speech identification on Twitter.
[2] Z. Waseem: "Am I seeing things, or are you a racist? Annotator influence on Twitter hate speech identification, in Proceedings of the First Workshop on Computational Social Science and Natural Language Processing, pp. 138–142, 2016.
[3] Ex machina: Personal attacks visible at scale, E. Wulczyn, N. Thain, and L. Dixon, arXiv preprint arXiv:1610.08914, 2016.
[4] "Measuring the reliability of hate speech annotations: The case of the European refugee crisis," B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki, arXiv preprint arXiv:1701.08118, 201,7.
[5] Y. Mehdad, C. Nobata, J. Tetreault, A. Thomas, "Abusive language detection in online user content," Y. Mehdad and Y. Chang, Proceedings of the 25th.International Conference on the World Wide Web, International Conferences Steering Committee, 2016, pp. 145–153.
[6] "Abusive language detection in online user content," C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, Proceedings of the 25th International Conference on World Wide Web, pp. 145–153, International World Wide Web Conferences Steering Committee, 2016.
[7] "Long short-term memory," S. Hochreiter and J. Schmidhuber, Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
[8] "Hate speech detection with comment embeddings," N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, Proceedings of the 24th International Conference on World Wide Web, p. 29–30, ACM,
[9] "Profanity use in online communities," S. Sood, J. Antin, and E. Churchill, Proceedings of the SIGGCHI Conference on Human Factors inComputing Systems, ACM, 2012, pp. 1481–1490.
[10] In the Proceedings of the Internet, Politics, and Policy conference, Policy and Internet, 2014, P. Burnap and M. L. Williams, "Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making."