Recognize Hate Speech on Twitter Using Machine Learning

Yashaswini Kini[1] , Chitra K[2*], Harilakshmi V.M.[3]

[1,2,3]Dayananda Sagar Academy of Technology and Management, Karnataka, India

***Email:** chitra-mca@dsatm.edu.in*

Abstract

Convolutional Neural Network (CNN) is a frequent-deep learning algorithm that is powerful in classifying image and text data, the system analyses individual tweets in order to determine if it contains hate speech. The occurrence of offensive speech in online forums poses significant challenges to maintaining a safe and inclusive digital environment. This study addresses these challenges by developing a hate speech recognition system ML methods, specifically CNN algorithms aimed primarily at analysing hate speech in tweets, attempting to increased resource efficiency and accuracy, its system analyses textual content in the tweet and produces and indicates whether it contains hate speech and determines the percentage of intolerance speech present in the tweet. The results of this study highlight the power of CNN-based strategies in preventing cyberbullying and promoting healthy digital discourse.

Keywords

Hate-Speech Detection, Machine Learning, Convolutional Neural Network (CNN), Social Networks Analysis, Text Classification

Introduction

Sentiment analysis of Twitter data through a hybrid machine learning model is explored. Emphasis is placed on classifying sentiment across multiple levels, providing a nuanced approach to understanding and categorizing sentiment in tweets. The methodology, model architecture, and results of employing this hybrid machine learning approach for sentiment analysis on Twitter are discussed by the authors (Bacha et al., 2023). The proposed study focuses on combating the problem of hate speech in CNNs through the use of machine learning techniques. This system is intended for the real-time processing of tweets and defining the amount of offensive words included in them. Employing CNN for more profound text analysis and using 0 to 100 percent to represent the probable intensity levels of hate present in the given tweet, the system also displays the percentage of offensive speech in a given tweet. The results that are obtained alongside the entered tweeting messages are further saved in a history page for further use or analysis. Firebase is used for securely storing data and for authenticating the users; Flask is used as the backend of the site whereas HTML, CSS, and JavaScript serves as the frontend of the website. This makes for efficient and holistic user experience which adds up to the greater course of eliminating hate speech in the various social media applications.

Submission: 13 October 2024; **Acceptance:** 1 November 2024

Problem Statement

The high levels of hatred on social media give rise to various risks and dangers affecting the online communities inclusive of discrimination, violence and heightened social tension. Preventing and combating such negative content is a significant issue regarding amount of material that is produced every day and context sensitivity of hate speech. Existing practices of moderate the hate speech are inadequate and ineffective for interventional action, there additional demand for the automated systems to detect and measure the offensiveness in the context of users' posts. Another problem is that the hate speech is very varied in its linguistic manifestations, and it is constantly developing so that methods of the detection of such content must be effective and versatile.

Literature

Abro compared automatic hate speech detection using machine learning techniques (Abro et al., 2020). They applied many feature engineering approaches and classifiers on a sample of hate speech tweets. From their conclusion, they discovered that SVM with bigram and TFIDF as features, had the highest accuracy rate of 79 percent. The study described the working of feature engineering and played the crucial role in discriminating texts class and mentioned the shortcomings of their investigation and perspectives on its further development.

Bache proposed an application of Deep Learning model to detect the offensive text in the various social platforms using YOLO, originally designed for the object's detection (Bacha et al., 2023). They pointed out that when it comes to memes with small text, the model did not work optimally, meaning that including more memes with such text might boost the results. Also, the model was limited to identifying only multi-lingually offensive text in English and classified inputs into only two classes. The future works can be focused on increasing the number of classes and enhancing the support for detecting the offensive text in other languages.

Glazkova compared logistic regression, random forest, linear support vector classifier, convolutional neural network, BERT, and RoBERTa on four Twitter benchmarks (Glazkova, 2023). Consequently, the results showed that, for the given models, some preprocessing techniques increased accuracy but decreased efficiency at the same time, and the other way around.

Concerning the recognition of hate-speech using NLP and Deep Learning, (Jahan et al., 2023) systematically reviewed the available literature. They stressed on the absence of high-quality projects accessible to the public, scarce comparative research, and inadequate funding for the non-English experiments. However, the study demonstrates that automated hate-speech identification is a crucial area for society, and there are many research avenues for this domain.

Mahajan later forward a model which uses Bi-directional Long Short-Term Memory (BiLSTM), Bi-directional Gated Recurrent Unit (Bi-GRU), CNN, and Long Short-Term Memory (LSTM) to increase the rate of the performance (Mahajan et al., 2024). On nine actual social datasets in different languages, they compared it to other current approaches and proved better performance in hate speech and cyberbully classification. Based their results the scores increased at least by 4. 44% in F1 scores .

The article under discussion by (Glazkova, 2023) described a research focused on the implementation of the Sentiment Analysis of the textual data of a social network – Twitter, a hybrid ML model was used. It was centred on of sentiments at different levels and was intended to give more elaborate categorization of sentiment in tweets. The study featured the discussions in methodology, model architecture, and results derived from the utilization of the machine learning for conducting sentiment analysis on Twitter. As for text classification, several text preprocessing strategies have been investigated for hate and other types of offense speech detection on the platform .

Pitsilis used recurrent neural networks RNNs. They used RNNs because they are proved to be effective in handling sequence data in the text format (Pitsilis et al., 2018). Specifically, it concerned methodologies, outcomes and findings with regard to tackling issues concerning the detection of hatred on the Twitter .

Sinyangwe used data CNN and Long Short-Term Memory (LSTM) for image and data identification and assessed their effectivity (Sinyangwe et al., 2023). Besides, they employed Random Forest Classifier for video classification and got a better accuracy rate. they discovered that a Naive Bayes classifier is 62% accurate according to their study. A feature-based approach was 75% accurate while a neural algorithm was 87% accurate on the same task of identification of hateful comments on social media.

William suggests that there is a need for automatic hate speech recognition, and, therefore, looked at using support vector machines for this purpose in social networking (William et al., 2022). To classify hate speech successfully, the authors of the study specialized on text to speech conversion and word recognition.

Zhang proposed deep neural network incorporating CNN and LSTM, for hate speech detection on social media (Zhang et al., 2018). They showed that the model true learns the features automatically and outperforms the traditional feature engineering methods. Accordingly, the study provided intensified focus by questioning the previous perceptions about feature engineering as a crucial step to increase the effectiveness of the machine learning algorithm.

Methodology

Previous work mainly focuses on classifying the tweet into hate speech and not hate speech. Existing system uses many types of algorithms including CNN. But none of the existing method predicts the percentage of hate present in the tweet. The proposed method overcomes the disadvantages of the existing system. It not only classifies the tweet into hate speech and not hate speech, if hate speech is detected it analyses and predicts the percentage of hate present in the tweet. For the proposed system of hate speech recognition under the category of ML, improvement will be made on the effectiveness and reliability of the detection models. Combining advanced NLP fundamentals like deep learning and CNN with the proposed system plans to obtain better results in the semantic analysis of tweets on hate-speech. Flask improvements in the backend will however improve scalability as well as response to user interactions better to the effect of accommodating the increased interactions. Firebase will remain the database of the project where the detected tweets are updated in real-time and the user login system. The frontend interface will be developed with focus on the usability and probably involve data visualization to display trends in hate speech. The main objective of this comprehensive approach is to offer a long-lasting and highly functional profile dedicated to searching for and fighting hate speech in social networks.

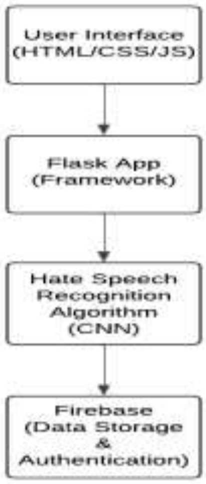
User Interface (HTML/CSS/JS)
Flask App (Framework)
Hate Speech Recognition Algorithm (CNN)
Firebase (Data Storage & Authentication)

Figure 1. Context diagram of the planned system

Results and Discussions

The developed system allows for the identification of hate speech and its intensity in tweets, thus offering a solution to the coverage of hostile information. The algorithm is verified to be most effective one in identifying hate speech, this is due to sound testing and validation processes. The employment of CNN helps to be more accurate in analysing the matter in the tweet, so the given results are credible. This integration of Firebase helps in managing the data more securely and for the authentication of the user that makes the application more functional and user friendly. Flask as the framework and the interaction of html, css, and javascript as the front end enable the users to easily interact through it. The history page feature provides information on the predicted tweets that would be useful once it is implemented, along with traceability, showing that the system is a total package for the hate speech detection.

Activity Diagram

The hate speech recognition system using ML is depicted in the activity diagram. It starts with a tweet input, which is then continued by a comparison to identify if the tweet is a hate speech tweet. In case of identification of hate speech, the percentage of hate in the tweet is determined and decided upon. In any case, the functions of the history page contain the results of the analysis, be it with the presence of hate speech or its absence. After that, all the information inclusive of the tweet and its analysis is saved in Firebase database. The system also includes the signing in and signing out of the users hence the user authentication is also the unit of the system. It also provides an efficient approach of identifying and inputting hate speech data into the system hence improving the functionality of the system and the satisfaction of the users.

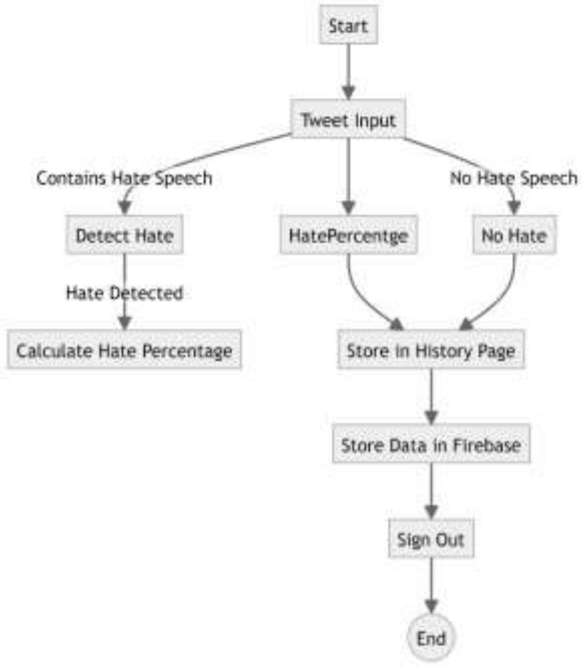

Figure 2. Activity diagram of the planned system

Conclusion

The applied scheme enables one to detect and measure hate speech in the tweets with the help of such an adequate solution to the issue of monitoring the content of the social network. Focusing on the detection of hate-speech and analysing the system output as the percentage of hate speech, the system proves to be efficient and effective, as it works based on the CNN algorithm. Thus, by means of the framework called Flask and employing the Firebase for data storage and authentication, the project provides a smooth experience for a user. The frontend is developed by HTML, CSS, JavaScript; this allows users to get forecasts and view older data easily. The ability to type in the predicted tweets and be able to see all the predicted tweets in one history page also adds functionality to the system especially for people in the procedure of analysing trends in hate speech conversations. The preservation of this functionality, along with the system's accurate detection and scoring by percentage, make this tool useful in combating and preventing hate speech on the internet. The combination of today's technologies and a friendly human interface guarantees the usage of the proposed system by showing that it meets and even goes beyond the criteria for hate speech identification and controlling.

Acknowledgement

The researcher did not receive any funding for this study, and the results have not been published in any other sources.

Reference

Abro, S., Shaikh, S., Khand, Z. H., Zafar, A., Khan, S., & Mujtaba, G. (2020). Automatic hate speech detection using machine learning: A comparative study. International Journal of Advanced Computer Science and Applications (IJACSA), 11(8). https://dx.doi.org/10.14569/IJACSA.2020.0110861

Bacha, J., Ullah, F., Khan, J., Sardar, A. W., & Lee, S. (2023). A deep learning-based framework for offensive text detection in unstructured data for heterogeneous social media. IEEE Access, 11, 124484-124498. https://doi.org/10.1109/ACCESS.2023.3330081

Glazkova, A. (2023). A comparison of text preprocessing techniques for hate and offensive speech detection in Twitter. Social Network Analysis and Mining, 13, Article 155. https://doi.org/10.1007/s13278-023-0155-8

Jahan, M. S., & Oussalah, M. (2023). A systematic review of hate speech automatic detection using natural language processing. Neurocomputing, 546, Article 126232. https://doi.org/10.1016/j.neucom.2023.126232

Mahajan, E., Mahajan, H., & Kumar, S. (2024). EnsMulHateCyb: Multilingual hate speech and cyberbully detection in online social media. Expert Systems with Applications, 236, Article 121228. https://doi.org/10.1016/j.eswa.2023.121228

Ojha, A. C. ., Shah, P. K. ., Gupta, S. ., & Sharma, S. . (2023). Classifying Twitter Sentiment on Multi- Levels using A Hybrid Machine Learning Model. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 328–333. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3711

Pitsilis, G.K., Ramampiaro, H. & Langseth, H. Effective hate-speech detection in Twitter data using recurrent neural networks. (2018). *Appl Intell* 48, 4730–4742 https://doi.org/10.1007/s10489-018-1242-y

Sinyangwe, C., Kunda, D., & Abwino, W. P. (2023), Detecting Hate Speech and Offensive Language using Machine Learning in Published Online Content, Zambia ICT Journal, Vol. 7, Iss. 1, pp. 79-84. https://doi.org/10.33260/zictjournal.v7i1.143

William, P., Gade, R., esh Chaudhari, R., Pawar, A. B., & Jawale, M. A. (2022). Machine Learning based Automatic Hate Speech Recognition System. *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India, pp. 315-318, https://doi.org/10.1109/ICSCDS53736.2022.9760959

Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on twitter using a convolution-gru based deep neural network. In The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15 (pp. 745-760). Springer International Publishing. https://doi.org/10.1007/978-3-319-93417-4_48