An In-Depth Research of Hate SpeechDetection on Twitter Using Machine Learning

Ram Totkar[1], Sneha Gujar[2], Ankleshwar Vishwakarma[3], Niharika Solanki[4], Pratik Upadhyay[5], Nikhil Pandhare[6]

Department of Computer Engineering, Genba Sopanrao Moze College of Engineering Balewadi, Pune-411045, Maharashtra, India

ABSTRACT

Hate speech is now a major issue that poses a threat to individuals and communities. One possible option for addressing this issue is utilizing machine learning algorithms to automatically identify and alert to hate speech in text data. Training a machine learning model for hate speech detection requires using a dataset with labelled examples of hate speech and non-hate speech. Different aspects like specific words/phrases, grammar, and syntax are taken from the text data, and the model is trained to differentiate between hate speech and non-hate speech using these aspects. The model that has been trained can be utilized to categorize fresh text data as either hate speech or non-hate speech. Nevertheless, it is crucial to acknowledge that utilizing machine learning for hate speech detection is not flawless and may be influenced by biases present in the training data or the algorithm itself. Current studies are concentrated on enhancing the precision and impartiality of algorithms designed to detect hate speech. In general, the utilization of machine learning for identifying hate speech could be beneficial in combating hate speech, but it is crucial to acknowledge its restrictions and prejudices.

Keywords: Hate Speech, Machine Learning, Dataset, Text Analysis.

INTRODUCTION

The detection of hate speech through machine learning is a significant and relevant subject in the current world, as hate speech and online harassment are increasing. Hate speech is defined as words or actions that show bias or discrimination towards a specific group due to their race, ethnicity, gender, religion, sexual orientation, or other personal traits. It is crucial to create techniques and strategies to identify and reduce the harmful effects of hate speech on people, communities, and the overall society.

Machine learning is effective for detecting hate speech because it can process vast amounts of data and identify patterns and characteristics to categorize text as hate speech or not. Machine learning algorithms can learn from labeled datasets of hate speech to recognize important characteristics and trends that can be utilized to automatically categorize new textas hate speech or not. This article will examine different methods for identifying hate speech using machine learning, such as supervised and unsupervised learning, feature engineering, deep learning, and natural language processing. We will also address the obstacles and constraints of identifying hate speech through machine learning, including the absenceof annotated datasets, the challenges in defining and detecting hate speech, and the risk of bias in machine learning algorithms.

In general, the goal of this paper is to give a summary of the current status of identifying hate speech through machine learning and to emphasize the possibilities and obstacles for future studies in this crucial and rapidly changing area.

LITERATURE SURVEY

"Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network" by Gao, W., et al. (2020). This paper proposes a deep learning approach for hate speech detection on Twitter. The model uses a combination of convolutional and GRU layers for feature extraction and classification.

"Deep Learning for Hate Speech Detection: A Comparative Analysis" by Mishra, P., et al. (2019). This paper presents a comparative analysis of various deep-learning approaches for hate speech detection. The authors experiment with several models, including CNNs, LSTMs, and GRUs, and evaluate their performance on multiple datasets.

"Combating Hate Speech on Social Media with Unsupervised Text Style Transfer" by Li, J., et al. (2018). This paper proposes an unsupervised text-style transfer approach for combating hate speech on social media. The authors use a neural network model to transform hate speech into non-offensive language while preserving the meaning of the original text.

"Deep Learning for Hate Speech Detection in Tweets" by Badjatiya, P., et al. (2017). This paper presents a deep learning approach for hate speech detection on Twitter. The model uses a combination of convolutional and LSTM layers for feature extraction and classification.

"Hate Speech Detection with Comment Embeddings and LSTM Networks" by Wulczyn, E., et al. (2017). This paper proposes a hate speech detection model that uses LSTM networks and comment embeddings. The authors use a large dataset of comments from online forums and social media platforms to train the model.

"Automated Hate Speech Detection and the Problem of Offensive Language" by Davidson, T., et al. (2017). This paper presents a study on the problem of automated hate speech detection. The authors create a dataset of Twitter posts labelled as hate speech or not, and experiment with various machine learning techniques for classification.

"Hate Speech Detection on Twitter: A Comparative Study" by Djuric, N., et al. (2015). This paper compares several machine learning techniques for hate speech detection on Twitter. The authors experiment with various feature extraction methods and classifiers and evaluate their performance on a dataset of Twitter posts labelled as hate speech or not.

SYSTEM ARCHITECTURE

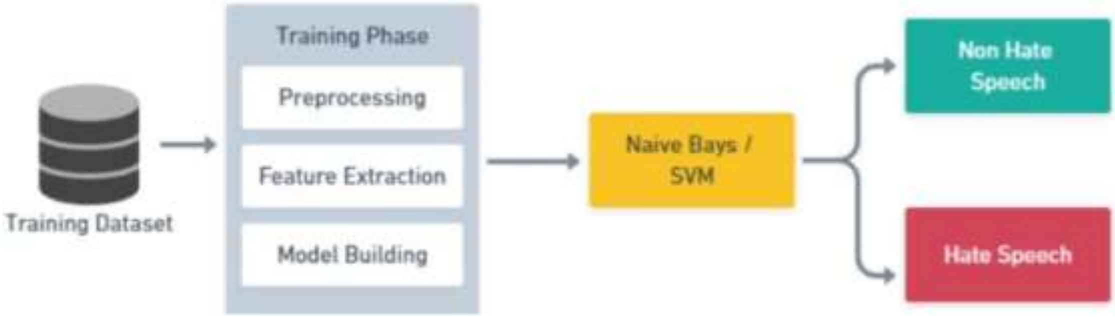


Figure 1 – System Architecture

METHODOLOGY

Identifying offensive language with the help of machine learning models like Support Vector Machines (SVM) and Naive Bayes is a popular method in the field of natural language processing. These algorithms can be used in creating a hate speech detection system by following these steps -

1. **Collect a hate speech dataset:** A dataset containing labeled instances of hate speech and non-hate speech is required. Numerous datasets are accessible for this purpose, including the Hate Speech and Offensive Language dataset and the Twitter Hate Speech dataset.

2. **Pre-processing the data:** Pre-processing includes preparing and converting the unprocessed text data into a structure suitable for the machine learning algorithm to utilize. A few typical pre-processing actions consist of tokenization, eliminating stop words, and stemming.

3. **Feature extraction:** This stage includes identifying important characteristics from the text that has been prepared beforehand. Techniques like bag of words, TF-IDF, or word embeddings can be employed to generate features forthe machine learning algorithm.

4. **Train the model:** Split your dataset into training and validation subsets. Utilize the training data to conduct the training process for your machine learning model. SVM and Naive Bayes are commonly used for identifying hate speech due to their ease of implementation and effectiveness with sparse feature vectors in high dimensions.

5. **Evaluate the model:** Utilize the validation set to assess the effectiveness of your model. Precision, recall, F1 score, and accuracy are typical assessment measures. Implement the model: After training and assessing the model, you can implement it to categorize fresh text as hate speech or non-hate speech.

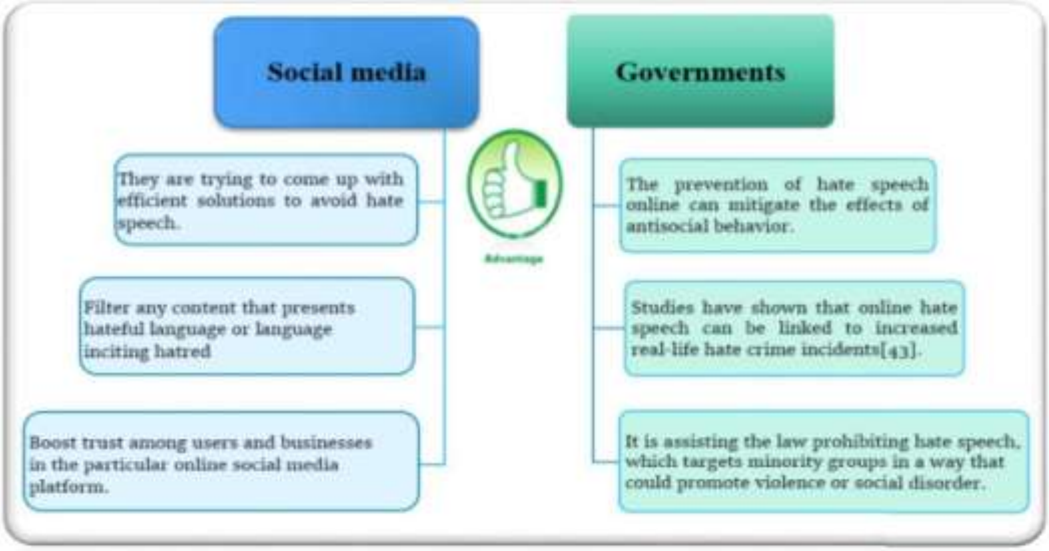

Social media
Governments
Advantage
They are trying to come up with efficient solutions to avoid hate speech.
Filter any content that presents hateful language or language inciting hatred
Boost trust among users and businesses in the particular online social media platform.
The prevention of hate speech online can mitigate the effects of antisocial behavior.
Studies have shown that online hate speech can be linked to increased real-life hate crime incidents[43].
It is assisting the law prohibiting hate speech, which targets minority groups in a way that could promote violence or social disorder.

Figure 2 - Benefits of the study and analysis of hate speech detection methods.

In a recent review article by [15], the authors employed machine learning techniques to categorize hate speech on Twitter, involving generic metadata de-signs, threshold configurations, and divergences. They also discussed the benefits and weaknesses of individual and integrated machine learning algorithms for the classification process. In addition, they displayed the hate speech benchmark dataset for testing the implementation of the classification paradigm. Even thoughsome surveys and reviews are available on this topic, significant limitations exist.

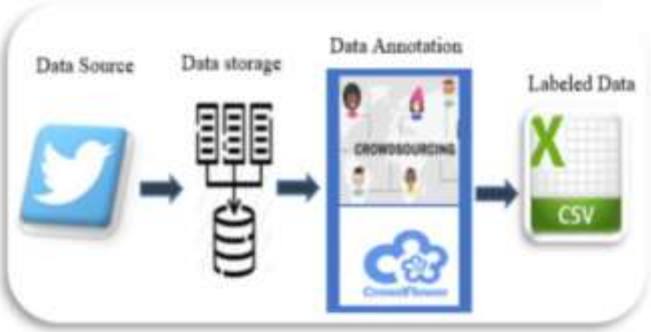
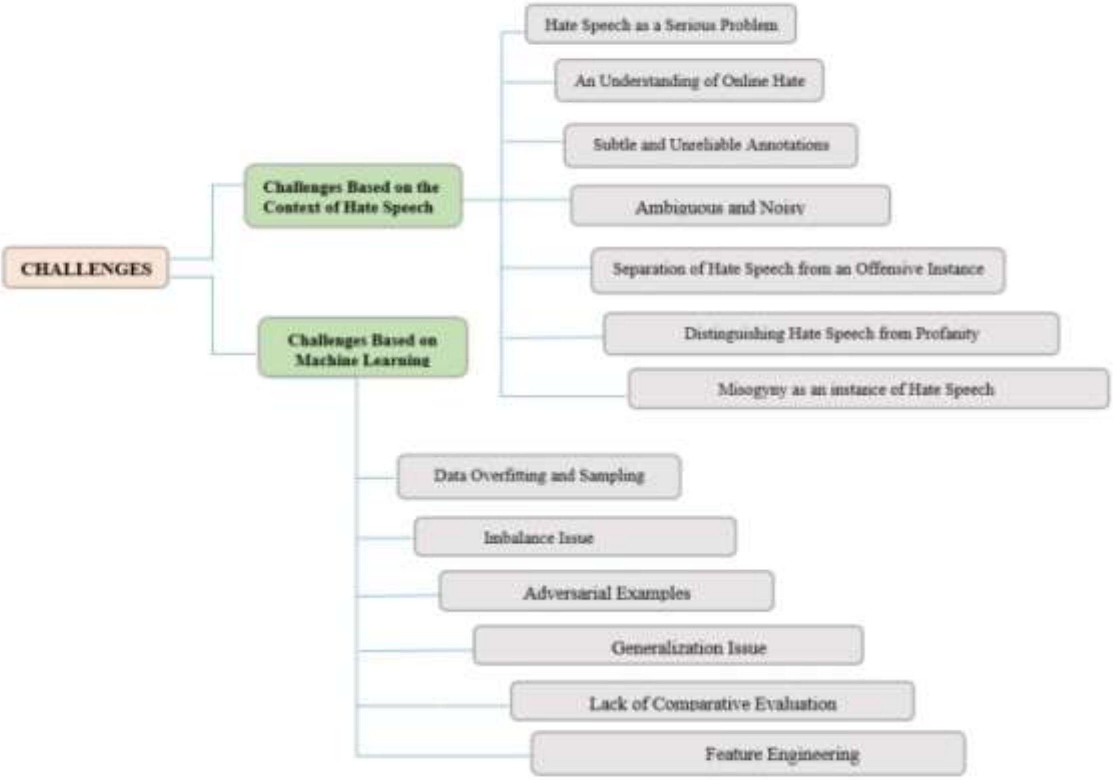


FIGURE 3 - General twitter data collection phases.


CHALLENGES
Challenges Based on the Context of Hate Speech
Challenges Based on Machine Learning
Hate Speech as a Serious Problem
An Understanding of Online Hate
Subtle and Unreliable Annotations
Ambiguous and Noisy
Separation of Hate Speech from an Offensive Instance
Distinguishing Hate Speech from Profanity
Misogyny as an instance of Hate Speech
Data Overfitting and Sampling
Imbalance Issue
Adversarial Examples
Generalization Issue
Lack of Comparative Evaluation
Feature Engineering

`

FIGURE 4 - The taxonomy of the selected studies is based on the existing challenges.

CONCLUSION

Identifying and categorizing offensive language online shows great potential through the use of machine learning algorithms like Naive Bayes for detecting hate speech. By analyzing different characteristics and training the model on a vast dataset of labeled data, Naive Bayes is able to accurately categorize text as either hate speech or non-hate speech. Nevertheless, it is crucial to emphasize that the efficacy of detecting hate speech with Naive Bayes, or any machine learning method, greatly depends on the caliber and variety of the training dataset. Hence, it is essential to meticulously select and verify the training data set to accurately depict the various forms of hate speech found across different environments and societies. Moreover, one must take into account the ethical consequences of employing machine learning for hate speech recognition, like the risk of biased algorithms and the effect on freedom of speech. Hence, it is essential to create and implement these tools responsibly and ethically, considering the wider social, cultural, and political environment.

REFERENCES

[1]. V. B. Ohol, S. Patil, I. Gamne, S. Patil, S. Bandawane, "Social Shout – Hate Speech Detection Using Machine Learning Algorithm", 2023 International Research Journal of Modernization in Engineering Technology and Science (IRJMETS), 2023 pp.-2582-5208, doi:7.868, Volume:05/Issue:05/May-2023.

[2]. P. William, R. Gade, R. e. Chaudhari, A. B. Pawar and M. A. Jawale, "Machine Learning based Automatic Hate Speech Recognition System," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 315-318, doi: 10.1109/ICSCDS53736.2022.9760959.

[3]. B. Pawar, P. Gawali, M. Gite, M. A. Jawale and P. William, "Challenges for Hate Speech Recognition System: Approach based on Solution," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 699-704, doi: 10.1109/ ICSCDS53736.2022.9760739.

[4]. V. Mercan, A. Jamil, A. A. Hameed, I. A. Magsi, S. Bazai and S. A. Shah, "Hate Speech and Offensive Language Detection from Social Media," 2021 International Conference on Computing, Electronic and Electrical Engineering(ICE Cube), 2021, pp. 1-5, doi: 10.1109/ICECube53880.2021.9628255.

[5]. Bhatia, P., Jain, R., & Kar, S. (2020). Automatic detection of hate speech: A survey. Journal of Ambient Intelligence and Humanized Computing, 11(9), 3837-3855.

[6]. Thakur, V., & Jain, A. (2020). A review on hate speech detection using machine learning techniques. Journal of Ambient Intelligence and Humanized Computing, 11(11), 5021-5034.

[7]. Kumar, A., & Zhang, L. (2020). Detecting hate speech on Twitter using a convolutional neural network. In Proceedings of the IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP) (pp. 95-101).

[8]. D. Elisabeth, I. Budi and M. O. Ibrohim, "Hate Code Detection in Indonesian Tweets using Machine Learning Approach: A Dataset and Preliminary Study," 2020 8th International Conference on Information and Communication Technology (ICoICT), 2020, pp. 1-6, doi: 10.1109/ICoICT49345.2020.9166251

[9]. A.Arango, J. Pérez, and B. Poblete, ‘‘Hate speech detection is not as easy as you may think: A closer look at modelvalidation (extended version),’’ Inf. Syst., vol. 105, Mar. 2022, Art. no. 101584

[10]. F. Alkomah and X. Ma, ‘‘A literature review of textual hate speech detection methods and datasets,’’ Information,vol. 13, no. 6, p. 122, 2022, doi:10.3390/info13060273.

[11]. R. T. Mutanga, N. Naicker, and O. O. Olugbara, ‘‘Detecting hatespeech on Twitter network using ensemble machine learning,’’ Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 3, pp. 331–339, 2022, doi: 10.14569/IJACSA.2022.0130341.

[12]. T. X. Moy, M. Raheem, and R. Logeswaran, ‘‘Hate speech detection in English and non-English languages: A review of techniques and challenges,’’ Webology, vol. 18, no. 5, pp. 929–938, Oct. 2021, doi: 10.14704/WEB/V18SI05/WEB18272

[13]. M. K. A. Aljero and N. Dimililer, ‘‘Genetic programming approach to detect hate speech in social media,’’ IEEE Access, vol. 9, pp. 115115–115125, 2021, doi: 10.1109/ACCESS.2021.3104535.

[14]. F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, ‘‘Machine learning techniques for hate speech classification of Twitter data: Stateofthe- art, future challenges and research directions,’’ Comput. Sci. Rev., vol. 38,Nov. 2020, Art. no. 100311.

[15]. Shaikh, S. and S.M. Doudpotta, Aspects Based Opinion Mining for Teacher and Course Evaluation. Sukkur IBA Journal of Computing and Mathematical Sciences, 2019. 3(1): p. 34-43.