Review On Hate Speech Recognition Using Text Analysis with The Goal of Reducing Inequality on Digital Platforms

Author: Swetha S (2348568)

Aim:

This literature review aims to highlight the interdisciplinary character of the topic and its potential for societal effect by defining hate speech, synthesizing the important discoveries and trends in hate speech identification using text analysis. This study contributes to a greater understanding of the difficulties and potential in combating online hate speech while preserving the concepts of ethical Artificial intelligence and free speech by pointing out gaps in existing research and exploring future approaches.

Abstract:

Due to the rise of hate speech online and its negative impact on people and communities, hate speech recognition has become an important focus of research in the field of text analysis. This literature review offers a comprehensive review of the state of the art. text analysis-based hate speech recognition studies and technological improvements. Surveying the various text analysis methods used for hate speech identification takes up a sizeable section of the article. Natural language processing models, conventional machine learning methods, and more recent developments in deep learning architectures are all included. In addition to illuminating potential directions for future study, this literature review aims to provide insights into the existing issues and developments in hate speech identification using text analysis. The comprehensive analysis adds to the larger conversation on monitoring hate speech in the age of digital media by emphasizing the value of multidisciplinary cooperation and technical innovation in promoting welcoming and civil online communities.

Keywords:

Hate speech; Text Analysis; Natural Language Processing; machine learning models; digital media; Artificial intelligence

Introduction:

The prevalent accessibility and immediate nature of internet communication platforms have created previously unthinkable chances for global interaction and discussion in the current digital age. However, this quick advancement in communication and technology has also given rise to new difficulties, the most significant of which is the spread of hate speech. Hate speech poses serious risks to societal cohesiveness, individual well-being, and democratic principles because of its offensive, discriminatory, and biased nature. Thus, it has become urgently important for scholars, lawmakers, and technological innovators to identify and eliminate hate speech in all of its expressions.

In order to combat the threat of hate speech, there has been an increasing consciousness in recent years that computational tools must be used to their full potential. Natural language processing (NLP)'s text analysis subfield has become an essential instrument in this process. Text analysis techniques provide the ability to effectively identify and minimize instances of hate speech across various internet platforms by automating the process of hate speech identification and categorization. As a result, developers and practitioners have focused more on creating and improving hate speech identification models using cutting-edge machine learning methods.

Literature review:

Based on Liriam Sponholz ‘s chapter titled "Hate Speech" [1] by in a book called "Challenges and perspectives of hate speech research" edited by C. Strippel, et al. we try to receive a formal definition of hate speech. The piece examines the conceptual difficulties associated with research on hate speech. It talks about the issue with the idea of "hate speech" and the necessity of having a precise grasp of its meaning and use. The article emphasizes the use of ideas in assessing and dealing with societal concerns. The definition of hate speech includes elements like assaults with an identity-based focus and symbolic nature. In hate speech research, idea stretching, shrinkage, and inflating lead to conceptual problems that make it difficult to compare empirical data. hatred speech is not a blanket word for all concerns relating to conflict in online communication, and it cannot be substituted by phrases like "online hate." To successfully combat extreme kinds of symbolic prejudice, it is important to understand how hate speech is communicated in public. The

idea of hate speech should be further developed by scholars in order to overcome conceptual obstacles and address this socially relevant issue

[2](Muhammad Okky Ibrohim et al. July2023) in "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," explains the developments and difficulties in identifying hate speech and abusive language in Indonesian social media. The identification of Indonesian hate speech and abusive language (HSAL) has been studied using a variety of techniques. Techniques for preprocessing, modelling, and feature extraction are all included in the procedures. Preprocessing Technique: To convert the alphabet into lowercase style and exclude extra characters like punctuation, URLs, and Twitter handles from text datasets, researchers frequently utilize case folding. Additionally, stop words are eliminated, and text normalization and stemming techniques may be used. Features and Models Extraction: To analyses and categorize HSAL, researchers have employed both supervised and unsupervised techniques. FP-Growth, association rule models, and Latent Dirichlet Allocation (LDA) are examples of unsupervised techniques. Classic machine learning models such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest Decision Tree (RFDT), k-Nearest Neighbor (kNN), Latent Semantic Analysis (LSA), Maximum Entropy, and Artificial Neural Network (ANN) are used in supervised methods.

From the article titled [3]Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions (Femi Emmanuel Ayo et al. September 2020), Twitter's Design Methodology for Hate Speech Classification can be studied like Metadata Extraction -where Using the Twitter Streaming API, the metadata extraction process comprises gathering Twitter datasets. We manually classified as hate speech or non-hate speech 25,000 English tweets with hate speech phrases from Hatebase.org. The database was compiled with the hate speech category for further analysis. Preprocessing: To begin preprocessing, all letters in the gathered tweets are changed to lowercase, and phrases are tokenized to create feature vectors. The data are now ready for the subsequent stage of feature extraction. Feature extraction entails sifting through the preprocessed hate tweets to extract semantic information. Bags of words made from trigrams and wordngrams are compared for semantic aspects with corpus-based produced opinion lexicons.

S. Jahan and M. Oussalah, in “A systematic review of hate speech automatic detection using natural language processing,” have briefed about the The selection of keywords was the initial step that was taken and how the search parameters were divided into six categories: hate speech, sexism, racism, cyberbullying, abusive, and insulting [4]. This is because the idea of hate speech encompasses a wide range of hate categories, this gives the best possibility of finding a sizable amount of pertinent work.

“SocialHaterBERT [5]: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles,” by (G. Del Valle-Cano et al. April 2023) , the HaterBERT technique is identified. BERT stands for Bidirectional Encoder Representations from Transformers . A foundation model called HaterBERT was created to categorize text on Twitter as either hate speech or non-hate speech. It is based on the well-known transformer-based model BERT. HaterBERT uses a variety of libraries, including HuggingFace, TensorFlow, Keras, PyTorch, and Keras, and modifies the transformer. Tokenizing the phrases, putting special tokens at the starting point and conclusion, giving token IDs, and fixing the leet alphabet are all phases in the preprocessing process. A classifier with the name of BertForSequenceClassification is also included in the model. Overall, HaterBERT has outperformed earlier classifiers in the identification of hate speech.

Finding from the literature Review:

From the literature review , it is found that,

1. Hate Speech in digital media is inevitable but not unsolvable.
2. Many detection techniques have been implemented and many more are constantly being evolved to make the world a better place to live in.
3. The definition of hate speech and what possibly could categorized as hate speech, the existing models of hate speech detection and the upcoming and most widely used models are the takeaways.

Methodology used:

Few of the methodologies that are used to detect speech would be . Classic machine learning models such as Naive Bayes, Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), Random Forest Decision Tree (RFDT), k-Nearest Neighbor (kNN), Latent Semantic Analysis (LSA), Maximum Entropy, and Artificial Neural Network (ANN).

The model provided in this paper has been created in partnership with ONDOD, which is a practical contribution. This study has opened a previously text-only field of study and paved the door for further investigation by confirming that the features of the users contribute significantly to recognizing hatred in the network. This work's classifier may be included into a continuous live monitoring program . The tool's output may be analyzed by the appropriate authorities to find spikes and triggers in hate speech as well as to come up with mitigating plans.

Research Gap:

Hate speech regulation is still a topic of discussion. It is still unclear if using legal action or other strategies (such counter-speech and education) is the appropriate course of action. With hate speech detection comes the concept of free speech and the imposing of free speech

References

[1]Sponholz, Liriam. (2023). Hate Speech. 10.48541/dcr.v12.9.

[2]M. O. Ibrohim and I. Budi, “Hate speech and abusive language detection in Indonesian social media: Progress and challenges,” *Heliyon*, p. e18647, Jul. 2023, doi: 10.1016/j.heliyon.2023.e18647.

[3]F. E. Ayo, O. Folorunso, F. T. Ibharalu, and I. A. Osinuga, “Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions,” *Computer Science Review*, vol. 38, p. 100311, Nov. 2020, doi: 10.1016/j.cosrev.2020.100311.

[4]S. Jahan and M. Oussalah, “A systematic review of hate speech automatic detection using natural language processing,” *Neurocomputing*, vol. 546, p. 126232, Aug. 2023, doi: 10.1016/j.neucom.2023.126232.

[5]G. Del Valle-Cano, L. Quijano-Sánchez, F. Liberatore, and J. Gómez, “SocialHaterBERT: A dichotomous approach for automatically detecting hate speech on Twitter through textual analysis and user profiles,” *Expert Systems With Applications*, vol. 216, p. 119446, Apr. 2023, doi: 10.1016/j.eswa.2022.119446