JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

HATE SPEECH CLASSIFICATION USING MACHINE LEARNING ALGORITHMS

AISHA SAJU [1], BINI WILSON[1], MUMTHAZ SAMAD[1], SALMA.N[1], SABIRA A S[2]

[1]UG Scholar, Department of Computer Science and Engineering

[2]Asst. Prof, Department of Computer Science and Engineering

UKF College of Engineering and Technology, Parippally, Kerala, India

Abstract : Using the Random Forest Classifier algorithm, we propose a complete method that combines text analysis with image classification to counteract toxic speech in online comments and texts. The classification of speech as toxic encompasses terms like "obscene," "toxic," "severe toxic," "threat," "insult," and "identity hate." This poses a serious obstacle to the promotion of constructive online dialogue. Putting these categories together under the heading of "Toxic" speech material will help us build a strong algorithm that can recognize and naturally remove harmful content from social media networks that are accessible online.Apart from textual analysis, we acknowledge the significance of picture content in identifying harmful conduct. Our methodology aims to offer a more complete method for recognizing harmful speech across many modalities by integrating picture classification techniques. Our objective is to create a high-accuracy classifier that can identify harmful content in both text and photos by utilizing machine learning algorithms.Our research intends to help create safer and more inclusive online communities by tackling toxic behavior in both textual and visual modes. We aim to create a more positive online community atmosphere and healthier online interactions by reducing the dissemination of harmful speech.

I. INTRODUCTION

In the contemporary digital era, the prevalence of toxic speech in online communication has grown to be a serious problem, endangering the inclusion, safety, and quality of conversation inside virtual groups. In order to counteract toxic speech in online comments and messages, this research proposes a complete strategy that integrates picture classification with text analysis and makes use of the Random Forest Classifier algorithm.The term "toxic speech" refers to a number of categories, some of which include "Obscene," "Toxic," "Severe Toxic," "Threat," "Insult," and "Identity Hate," each of which presents a different set of difficulties for encouraging positive online interactions. The undertaking seeks to construct a strong model that can detect harmful content from internet-based social media networks by combining these categories under the heading of "Toxic" speech material.The research incorporates picture classification algorithms into its framework in recognition of the significance of image content in detecting toxic behavior. The objective is to achieve a more complete approach to recognizing toxic speech across different modalities.The aim is to create a high-accuracy classifier that can identify harmful content in text and photos by using machine learning methods,to make online places safer and more inclusive by tackling toxic behavior in both textual and visual modes,to create a more positive online community environment and encourage healthier online interactions by reducing the dissemination of toxic speech.This initiative is important because it has the ability to effectively counteract toxic speech, which will improve the general quality of online debate and encourage inclusivity in virtual communities. Through the creation of a powerful system to categorize comments into groups like "Toxic," "Severe Toxic," "Obscene," "Insult," "Threat," or "None," the project enables platform managers and users to effectively identify and address toxic behavior.

II. METHODOLOGY

In general,the hatespeech detection and classification using machine learning algorithms concept originated from the journal [1].This deals with the problem of hatespeech in twitter, Hate speech seems to be an aggressive form of communication that propagates hateful ideas by utilizing false information. A number of protected characteristics, such as gender, religion, color, and disability, are the subject of hate speech. Hate speech can occasionally lead to unwanted crimes because it demoralizes an individual or group of individuals. A Deep Convolutional Neural Network (CNN)-based framework for hate speech identification is presented. This research constraint to exclusively use text-based features for hate speech recognition may be a problem because it may miss subtle aspects of poisonous behavior found in visuals. The proposed research, on the other hand, has the advantage of integrating text and image analysis with the Random Forest Classifier method. The proposed paradigm provides a more thorough method of identifying harmful speech across different modalities by utilizing both textual and visual clues. This helps to create

safer and more welcoming online environments by making it possible to detect and classify hazardous content on online platforms with greater accuracy and effectiveness.

[2] This study investigates the challenges associated with accurately categorizing hate speech due to the subjective assessments and biases of the annotators. Using a collection of tweets, the study highlights how difficult it is to distinguish hate speech from other offensive words.This research emphasizes the difficulty of precisely characterizing toxic communication, especially hate speech, due to differing subjective perspectives and preconceptions. The challenges outlined in this research emphasize the need for trustworthy mechanisms that can recognize and deal with inappropriate activity in virtual communities. In response to these problems, the proposed research aims to develop a comprehensive plan to stop toxic speech by integrating text and visual analysis with machine learning techniques.This uses the Random Forest Classifier algorithm to classify a range of categories, including hate speech, under the overall category of "Toxic" speech content in an effort to address the challenges associated with toxic speech detection. Moreover, hazardous conduct is recognized as multi-modal by employing picture classification techniques, providing a more thorough way to identify unsafe information across several modalities. Furthermore, the project's objectives, which include implementing preprocessing techniques, researching feature extraction tactics, and evaluating system performance, are consistent with the findings and recommendations made in Waseem's study.The study's conclusions, which highlight the necessity of adapting to risky behavior patterns that change over time, are mirrored in the focus on continuous monitoring and process improvement.In view of the subjective character of annotator influence, Waseem's work's conclusions emphasize the need for trustworthy algorithms that accurately categorize harmful speech. The study presents a thorough plan for reducing toxic speech in online communities in order to get over these challenges. Its objective is to promote online connections that are better and inclusive.

[3] Because of the difficulties in identifying objectionable language, the finding of Davidson et al.'s article probably highlights how difficult it is to identify hate speech with accuracy. They might propose that in order to increase the efficacy of automated systems in detecting hate speech on social media platforms, they should take into account a variety of contextual elements and linguistic subtleties. They may also suggest avenues for further investigation or possible enhancements to current detection techniques.This study emphasizes the difficulties in differentiating hate speech from offensive language and the intricacies involved in automated hate speech identification.The suggested research offers an advantage in that it takes a comprehensive approach, utilizing the random forest classifier algorithm to integrate text and image analysis. This allows for more precise identification and categorization of harmful speech across a range of modalities.This method efficiently curbs the dissemination of hazardous content, which helps to create safer online environments.

[4] The survey on countering hate speech online by Gaydhani et al. provides a comprehensive analysis of mitigation and detection strategies. It includes a variety of tactics such social network analysis, natural language processing (NLP) methods, and machine learning algorithms. The study contributes to a better understanding of this important problem in the digital sphere by offering insightful information about the state-of-the-art techniques now employed to combat online hate speech. This survey offers an overview of mitigation and detection strategies for countering hate speech on the internet. Exclusively focus on language analysis and disregard for picture content when identifying harmful conduct could be a disadvantage. However, the proposed research has an advantage because it integrates the Random Forest Classifier method with image and text analysis. Through the integration of image classification methods with text analysis, the suggested framework provides a more all-encompassing method for detecting toxic speech in many modalities. By doing this, hate speech detection becomes more accurate and effective, which helps to create online environments that are safer and more welcoming.

[5] The survey on natural language processing (NLP)-based hate speech detection by Park et al. offers a thorough examination of methods in this field. It covers important topics such feature extraction, dataset generation, classification algorithms, and assessment measures, providing a thorough resource for both practitioners and researchers. The survey furthers the understanding and development of efficient hate speech detection systems utilizing NLP technologies by synthesizing and analyzing current approaches.This survey narrow focus on textual analysis methods without taking into account the use of image content for hate speech detection could be one of its possible weaknesses.The proposed research has the advantage of a comprehensive method that uses the Random Forest Classifier algorithm to merge image and text analysis. The suggested framework provides a more comprehensive method of recognizing poisonous speech across several modalities by taking into account both textual and visual indicators, improving the precision and effectiveness of hate speech detection in online communities.

III. CONCLUSION

This research offers a thorough strategy for preventing toxic conduct in online communication by creating a system that is based on machine learning. Through precise classification of online comments into pre-established toxicity categories, the approach seeks to foster more inclusive and safe online communities. By using sophisticated methodologies including multimodal analysis, fine-grained classification, and dynamic adaptation, the system can successfully evolve and adjust to new harmful speech patterns. Enhancements to user interaction and ethical issues highlight the significance of responsible and cooperative content moderation initiatives. As the project develops, addressing the complex difficulties posed by toxic conduct online will require further work in scalability, performance optimization, and interdisciplinary study. Ultimately, this initiative aims to contribute to

the development of a more pleasant and courteous online environment for all users by utilizing technology to promote healthier online interactions.

REFERENCE

[**1**] Pradeep Kumar Roy ,Asis Kumar Tripathy,Tapan kumar Das And Xiao-Zhi Gao," A Framework for Hate Speech DetectionUsing Deep Convolutional Neural Network.

[2] Waseem, Zeerak. "Are You a Racist or Am I Seeing Things Annotator Influence on Hate Speech Detection on Twitter." Proceedings of the First Workshop on NLP and Computational Social Science. 2016.

[3] Davidson, Thomas et al. "Automated Hate Speech Detection and the Problem of Offensive Language." Proceedings of the International AAAI Conference on Web and Social Media. 2017.

[4] Gaydhani, Tejas et al. "Combating Online Hate Speech: A Survey on Detection and Mitigation Techniques." ACM Computing Surveys. 2020.

[5] Park, Donghoon et al. "A Survey on Hate Speech Detection using Natural Language Processing." Expert Systems with Applications. 2021.