HATE SPEECH DETECTION IN ONLINE SHORT-FORM VIDEOS THROUGH AUDIO CLASSIFICATION

[1]B.Nirupama, [2]Nara Taruni, [3]Bingi Sathwika,[4] Pyata Tarun,

Assistant Professor in department Of IT Teegala Krishna Reddy Engineering College

nirupama.varagani@gmail.com

UG Scholars In Department of IT Teegala Krishna Reddy Engineering College

[2]narataruni@gmail.com ,[3] ssathvika208@gmail.com ,[4] tarun.p.1793@gmail.com

Abstract

In this study, we pioneer the development of an audio-based hate speech classifier from online, short-form TikTok videos using traditional machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machines. We scraped over 4746 videos using the TikTok API tool and extracted audio-based features such as MFCCs, Spectral Centroid,Rolloff, Bandwidth, Zero-Crossing Rate, and Chroma values as primary feature sets. Results show that using the extracted predictors for hate speech detection can obtain upto 78.5% accuracy on an optimized Random Forest model, crossing the 50% benchmark for models in this task. In addition, comparing the Information Gain scores and globally learned model weights identified that Spectral Rolloff and MFCCs are top predictors in discriminating hate speech for the Filipino language. Index Terms—hate speech, tiktok, audio classification, machine learning, speech processing.

I INTRODUCTION

The rise of hate speech on the internet has become more apparent in the recent years. With the use of different major social media platforms, its dissemination has become faster than ever. Across the globe, the number of hateful contents on the internet has continuously increased endangering not only individuals but society on a wider spectrum. However, the definition of hate speech remains a controversial topic amongst many political discourses whereas its main discussion is whether hate speech falls within the scope and protective rights of freedom of expression. More democratic regions including Australia, New Zealand, France, and Denmark legislation see hate

Speech as a form of criminal offense. On the other hand, in United States this could be considered as unconstitutional because of the first amendment which is a law that protects the

basis of expression of moral and political convictions. As defined from Encyclopedia Britannica, hate speech is a form of expression that attacks an individual or group based on its

ethnicity, race, religion, social status, gender, sexual orientation, physical or mental disability, and others. The impact of hate speech can negatively affect an individual's mental health and behavioral upbringing, in worst case scenarios it could contribute to incitement of hate crimes and violence. Thus, the need for extensive research on automatic hate speech detection has undeniably become an important open problem in the fields of Natural Language Processing (NLP) as more and more victims or groups are being endangered from potential exposure to malicious contents. Identifying hateful contents in many digital platforms is a challenging task as contents vary in input types such texts, images, videos, and audio data.Many researchers have investigated different approaches in hate speech detection such as using Twitter tweets and Facebook memes1. However, in this study, the focus is on speech or via a human voice input type to detect malicious content. Speech is defined as thought expressed through sound or spoken words and is considered as a good basis to recognize emotions. Different social media platforms enable different kind of contents to be published, but when it comes to video sharing, Tiktok is one of the most popular sites globally. Thus, in this paper, we pioneer the development of a hate speech detection model using extracted Filipino speech data from shortform TikTok videos. To build the models, we explore the use traditional audio-based features for training model with machine learning algorithms such as SVM, Logistic Regression, and Random Forest. To differentiate, the type of hate speech detection task that this study tackles only cater to the use of speech data without the additional process to transcribe the speech in contrast to previous studies. We hope that this study may serve as an initiative to encourage other researchers in all fields of study to also work on hate speech detection on the TikTok platform.

II LITERATURE SURVEY

Hate speech on social media: Global comparisons,"

Hate speech and hate crimes are trending. In the past five years, there has been an upsurge in extreme nationalist and nativist political ideology in mainstream politics globally. In the United States, the President regularly mobilizes a political constituency by vilifying Mexican immigrants as "criminals" and "rapists" who "infest" America, and by promoting a "zero tolerance" policy at the border that punitively separates children from their parents, including persons exercising their right to apply for asylum.2 Data suggest a connection between this rise in rhetoric to increases in hate crimes in the United States.3 Similar trends are evident abroad as well.In the United Kingdom, the 2016 Brexit referendum elicited conspicuous expressions of anti-Muslim and anti-immigrant sentiment and coincided with the sharpest increase in religiously and racially motivated hate crimes ever recorded in British history**.**

The ongoing challenge to define free speech

Freedom of speech, Supreme Court Justice Benjamin Cardozo declared more than 80 years ago, “is the matrix, the indispensable condition of nearly every other form of freedom.” Countless other justices, commentators, philosophers, and more have waxed eloquent for decades over the critically important role that freedom of speech plays in promoting and maintaining democracy. Yet 227 years ater the irst 10 amendments to the U.S. Constitution were ratiied in 1791 as the Bill of Rights, debate continues about the meaning of freedom of speech and its First Amendment companion, freedom of the press. his issue of Human Rights explores contemporary issues, controversies, and court rulings about freedom of speech and press. his is not meant to be a comprehensive survey of First Amendment developments, but rather a smorgasbord of interesting issues.

Content regulation and the first amendment

As a general matter, government may not regulate speech “because of its message, its ideas, its subject matter, or its content.”

1 “It is rare that a regulation restricting speech because of its content will ever be permissible.”

2 The constitutionality of content-based regulation is determined by a compelling interest test derived from equal protection analysis: the government “must show that its regulation is necessary to serve a compelling state interest and is narrowly drawn to achieve that end.”

3 Narrow tailoring in the case of fully protected speech requires that the government “choose[ ] the least restrictive means to further the articulated interest.”

4 Application of this test ordinarily results in invalidation of the regulation

Decision level combination of multiple modalities for recognition and analysis of emotional expression

Emotion is expressed and perceived through multiple modalities. In this work,we model face, voice and head movement cues for emotion recognition and we fuse classifiers using a Bayesian framework. The facial classifier is the best performing followed by the voice and head classifiers and the multiple modalities seem to carry complementary information, especially for happiness. Decision fusion significantly increases the average total unweighted accuracy, from 55% to about 62%. Overall, we achieve average accuracy on the order of 65-75% for emotional states and 30-40% for neutral state using a large multi-speaker, multimodal database. Performance analysis for the case of anger and neutrality suggests a positive correlation between the number of classifiers that performed well and the perceptual salience of the expressed emotion.

III EXISTING SYSTEM

The existing system of the project involves the development of an audio-based hate speech classifier specifically designed for online short-form videos on the TikTok platform. The researchers employed traditional machine learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machines, to create a robust hate speech detection model.Using the TikTok API, they collected a dataset comprising over 4746 videos, from which audiobased features such as MFCCs (Mel-Frequency Cepstral Coefficients), Spectral

Centroid, Rolloff,Bandwidth, Zero-Crossing Rate, and Chroma values were extracted as primary feature sets.The study achieved promising results, demonstrating that hate speech detection using these extracted predictors could achieve up to 78.5% accuracy, surpassing the 50% benchmark for models in this specific task. Furthermore, the comparison of Information Gain scores and globally learned model weights revealed that Spectral Rolloff and MFCCs emerged as the top predictors in discriminating hate speech, particularly for the Filipino language. This research lays the foundation for effective hate speech detection in the dynamic context of short-form videos on social media platforms.

LIMITATIONS

***Data Bias and Generalization*:** One limitation is the potential bias in the dataset collected through the TikTok API. If the dataset is not diverse or representative enough, the model may struggle to generalize well to various types of hate speech or linguistic nuances, limiting its real world applicability.

Language and Cultural Specificity: The study focuses on hate speech detection for the Filipino language, and this specialization may limit the model's effectiveness when applied to other languages and cultures. Hate speech expressions and linguistic nuances can vary significantly across different regions and communities.

Algorithmic Complexity: While traditional machine learning algorithms like Logistic Regression,Random Forest, and Support Vector Machines were used, the study may not account for the complexities and subtleties of hate speech, especially in the evolving landscape of online communication. More advanced deep learning models might be required for improved performance**.**

***Dynamic Nature of Online Content*:** Online platforms, including TikTok, are dynamic environments with rapidly changing content and trends. The model's training data might not capture the evolving nature of hate speech, leading to potential performance degradation over time as new expressions and patterns emerge.

***Limited Audio Feature Set*:** The study primarily relies on a set of audio features such as MFCCs, Spectral Centroid, Rolloff, Bandwidth, Zero-Crossing Rate, and Chroma values. While these features provide valuable information, the exclusion of other relevant features or the absence of a multimodal approach (combining audio with video or text features) might limit the model's ability to comprehensively capture the nuances of hate speech in short-form videos.

IV PROPOSED SYSTEM

The proposed system aims to address the limitations of the existing model by incorporating advanced techniques and considerations to enhance the accuracy and robustness of hate speech classification in online short-form videos. The system will explore the utilization of state-of-the-art deep learning

models, such as recurrent neural networks (RNNs) or transformers, to capture
intricate patterns and dependencies within audio data, enabling more nuanced discrimination of hate speech. Additionally, efforts will be made to diversify the dataset, ensuring a broader representation of languages, dialects, and cultural contexts to improve the model's generalization capabilities. The proposed system will also implement a dynamic training approach that adapts to the evolving nature of online content by incorporating continuous learning mechanisms. To overcome the limitations of a singular focus on audio features, a multimodal approach will be explored, integrating audio, video, and potentially textual features to provide a more comprehensive understanding of the context in which hate speech occurs. This proposed system aims to push the boundaries of hate speech detection in online short-form videos, fostering a more inclusive and effective model for combating harmful content across diverse linguistic and cultural landscapes.

V IMPLEMENTATION

Data Collection and Preprocessing: This module involves the retrieval of online short-form videos from TikTok using the TikTok API. The collected data undergoes preprocessing to clean and format the audio content. Additionally, linguistic and cultural metadata may be extracted to enhance dataset diversity.

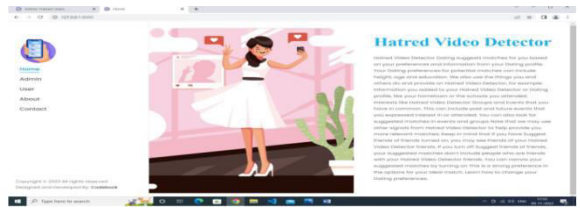
Feature Extraction and Representation: In this module, audio-based features such as Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Centroid, Rolloff, Bandwidth, Zero-Crossing Rate, and Chroma values are extracted from the preprocessed audio data. The goal is to create a comprehensive set of features that characterizes the unique aspects of hate speech in the given context.

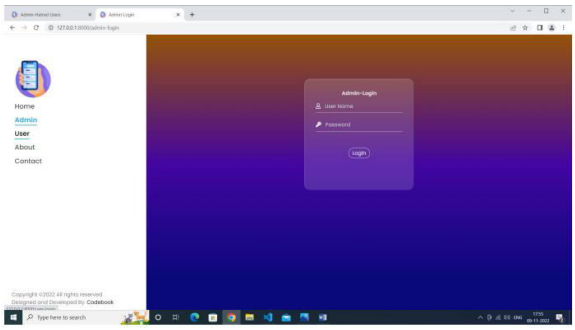
Model Training and Optimization: The training module involves the implementation and optimization of machine learning models, possibly utilizing advanced deep learning architectures like recurrent neural networks (RNNs) or transformers. The models are trained on the extracted features, and hyper parameters are tuned to achieve the highest possible accuracy in hate speech classification.

Dynamic Learning and Continuous Updating: To address the dynamic nature of online content, This module implements continuous learning mechanisms. The model is designed to adapt and Update itself as new data becomes available, ensuring that it remains effective in identifying
Evolving patterns of hate speech over time.

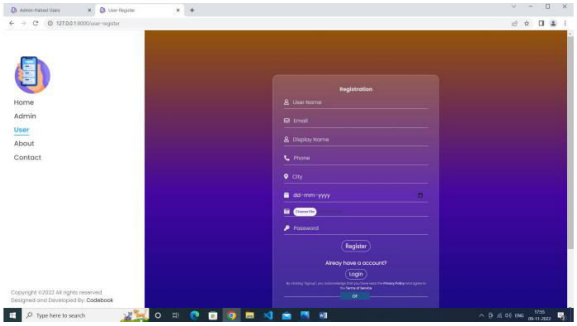
Multimodal Integration and Analysis: The multimodal module focuses on integrating audio features with additional modalities such as video and potentially textual data. This integration provides a more comprehensive understanding of the context surrounding hate speech. The system analyzes and weighs the contributions of each modality to enhance the overall accuracy and reliability of hate speech classification in online short-form videos.

VI RESULTS

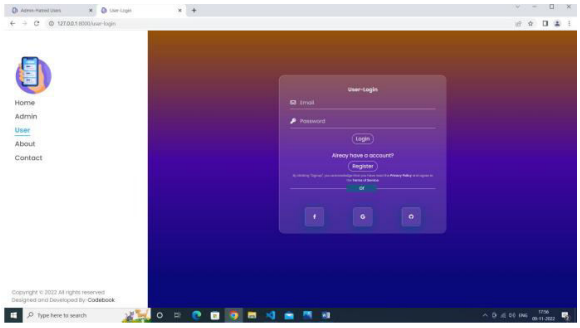




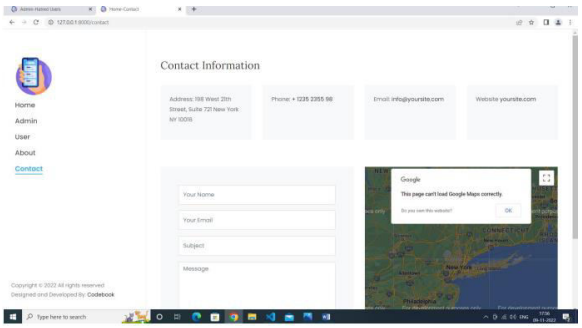
Admin



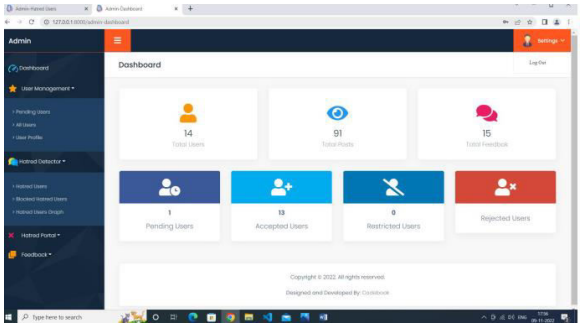
User Registration



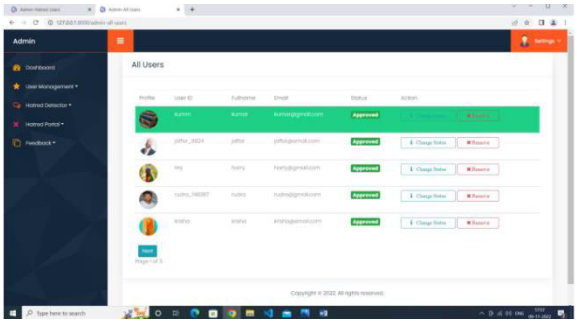
User login



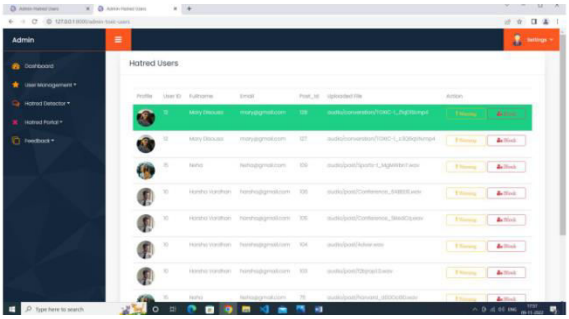
Contact Information



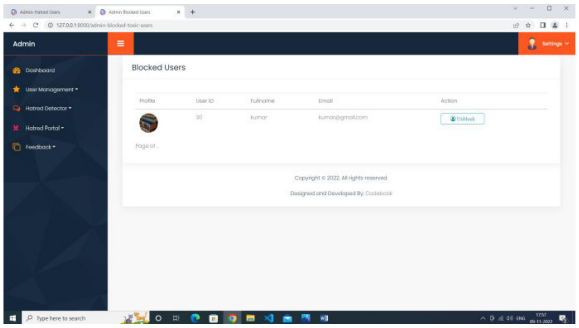
Dashboard



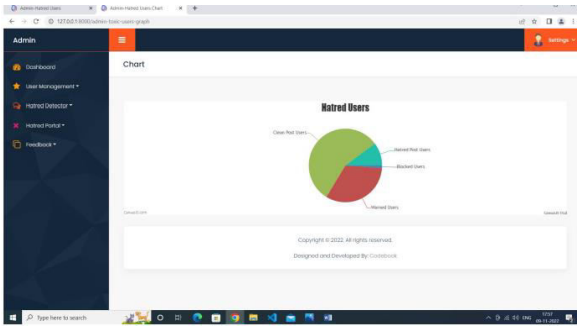
All Users



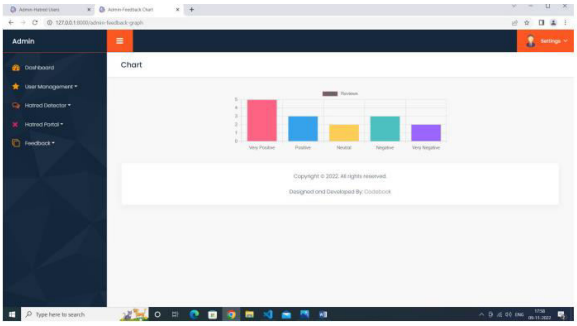
Hatred Users



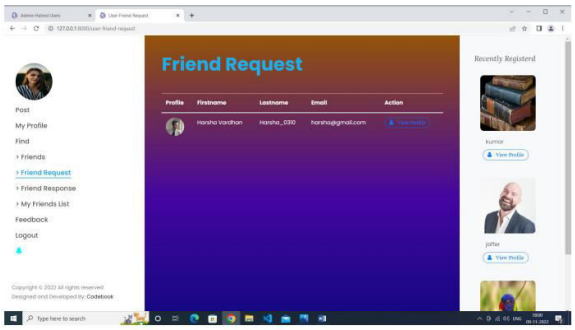
Blocked users



Chart



Feedback Chart



Friend Request

VII CONCLUSION

The paper explored machine learning classification algorithms, Support Vector Machine, Logistic Regression, and Random Forest, to develop a hate speech classifier from online short-form TikTok videos. Evaluation metrics such as accuracy, precision, recall and F1 score were used to leverage the performances of the models. Results obtained from training showed that Random Forest Classifier model provided the best performance with an accuracy of 78%. Performing a two-pronged feature selection technique using Information Gain and extraction of global learned weights showed that the two most contributive audio-based feature suitable for hate speech detection are Spectrall Rolloff and Mel Frequency Ceptral Coefficients. For future works, a finergrained approach on the hate speech detection task as well as application beyond discourse onc politics.

REFERENCES

[1] Z. Laub, “Hate speech on social media: Global comparisons,” Council on Foreign Relations, vol.7, 2019.

[2] S. Wermiel, “The ongoing challenge to define free speech,” Human Rights, vol. 43, no. 4, p. 82,2018.

[3] J. W. Howard, “Free speech and hate speech,” Annual Review of Political Science, vol. 22, pp.93–109, 2019.

[4] G. R. Stone, “Content regulation and the first amendment,” Wm. & Mary L. Rev., vol. 25, p. 189,1983.

[5] W. M. Curtis, “Hate speech,” November 29 2016.

[6] E. Barendt, "What is the harm of hate speech?," Ethical Theory and Moral Practice, vol. 22, no. 3,pp. 539–553, 2019.

[7] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th international conference on World Wide Web companion, pp.759–760, 2017.

[8] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," arXiv preprint arXiv:2005.04790, 2020.

[9] J. Herrman, "How tiktok is rewriting the world," The New York Times, vol. 10, 2019.

[10] J. Kim, "Bimodal emotion recognition using speech and physiological changes," Robust speech recognition and understanding, vol. 265, p. 280, 2007.

[11] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2462–2465, IEEE, 2010.

[12] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," Proceedings of the ACM on Human-Computer Interaction, vol. 2, no. CSCW, pp. 1–26,2018.

[13] M. Barakat, C. Ritz, and D. A. Stirling, "Detecting offensive user video blogs: An adaptive keyword spotting approach," in 2012 International Conference on Audio, Language and Image Processing, pp. 419–425, IEEE, 2012.