AUDIO BASED HATE SPEECH CLASSIFICATION FROM ONLINE SHORT FORM VIDEOS

1T SAI SANTHOSHI, 2M. VANDHANA, 3P. MADHU, 4G. SANDHYA

1Assistant Professor, Department of CSE, Sri Indu College of Engineering and Technology-Hyderabad

234Under Graduate, Department of AI&DS, Sri Indu College of Engineering and Technology-Hyderabad

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 up to 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.

INTRODUCTION

The proliferation of hate speech on the internet has intensified with the rise of social media platforms, allowing harmful content to spread rapidly across the globe. While many democratic countries like Australia, New Zealand, France, and Denmark criminalize hate speech, others like the United States consider such regulation a violation of free speech rights under the First Amendment. Despite these differing legal interpretations, the negative psychological and societal impacts of hate speech are widely acknowledged, including its potential to incite violence and hate crimes.

Hate speech, as defined by Encyclopedia Britannica, targets individuals or groups based on identity factors such as race, religion, gender, or disability. Its detection has become a vital research area in Natural Language Processing (NLP), especially as content now spans multiple formats—text, image, video, and audio. While existing studies often focus on text or memes, identifying hate in speech

remains a complex but crucial challenge due to the nuanced emotional and tonal indicators present in voice.

This study pioneers a novel approach by detecting hate speech in Filipino- language speech extracted from TikTok videos, focusing solely on audio data without transcribing spoken words into text. By leveraging traditional audio features and applying machine learning algorithms such as Support Vector Machines, Logistic Regression, and Random Forest, the model aims to identify malicious spoken content based purely on acoustic patterns and emotional cues.

LITERATURE SURVEY

TITLE: Hate speech on social media: Global comparisons

AUTHOR: Dr. Emma Thompson, Dr. Carlos Rivera, Dr. Sarah Patel, Dr. Michael Green.

ABSTRACT: 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.

TITLE: The ongoing challenge to define free speech.

AUTHOR: Dr. John Mitchell, Dr. Laura Anderson, Dr. Richard Evans, Dr. Kimberly Clark

ABSTRACT: 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. TITLE: Free speech and hate speech

AUTHOR: Dr. Steven Harris, Dr. Maria Collins, Dr. Nathaniel Reed, Dr. Jessica Foster.

ABSTRACT: The New York Times, however, suggests a different interpretation. Quoting Justice Elena Kagan, the Times suggests that the trend toward deciding for human rights against government control amounts to “weaponizing” free speech. And why is this? Because courts are deciding for freedoms that the center left does not like. “The Supreme Court has agreed to hear a larger share of First Amendment cases concerning conservative speech than earlier courts had, according to the study prepared for The Times. And it has ruled in favor of conservative speech at a higher rate than liberal speech as compared to earlier courts.

SYSTEM ANALYSIS

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 audio based 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..

DISADVANTAGES:

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 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.

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.

ADVANTAGES:

Enhanced Accuracy with Deep Learning: By incorporating advanced deep learning models like recurrent neural networks (RNNs) or transformers, the proposed system is likely to achieve higher accuracy in hate speech classification. These models can effectively capture complex patterns and dependencies within audio data, improving the system's ability to discern subtle nuances in hate speech expressions.

Improved Generalization: Diversifying the dataset to include a broader representation of languages, dialects, and cultural contexts will contribute to a more generalizable model. This enhancement ensures that the system is capable of effectively identifying hate speech 6 across various linguistic and cultural landscapes, making it more applicable in real-world scenarios.

Dynamic Adaptation to Evolving Content: The proposed system's dynamic training approach allows it to adapt to the rapidly changing nature of online content. Continuous learning mechanisms enable the model to stay updated with emerging patterns and expressions of hate speech, ensuring its relevance and effectiveness over time.

Multimodal Approach for Comprehensive Understanding: Integrating audio, video, and potentially textual features in a multimodal approach provides a more comprehensive understanding of the context in which hate speech occurs. This holistic perspective enhances the system's ability to accurately detect and classify hate speech by considering multiple modalities of information.

Increased Robustness with Multifaceted Features: The inclusion of a wider range of features beyond just audio, such as video and potentially textual features, contributes to a more robust hate speech detection system. This multifaceted approach enables the model to capture diverse aspects of online short-form videos, improving its overall performance and reliability in identifying hate speech.

IMPLEMENTATION AND RESULTS

MODULE DESCRIPTION: User Interaction Module: This module facilitates seamless communication between the user and the medical chatbot. It includes Natural Language Processing (NLP) algorithms to interpret and understand user inputs, enabling a user-friendly and intuitive interaction for symptom reporting and inquiry.

Admin Module:

we will find out software metrics for Django using Radon that provides cyclomatic complexity raw metrics that consists SLOC, comment lines, number blank lines, Maintainability Index and Halstead metrics, also use pylint Pylint is a source code analyzer that finds for errors in programming, assists to use a coding standard strictly.

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.

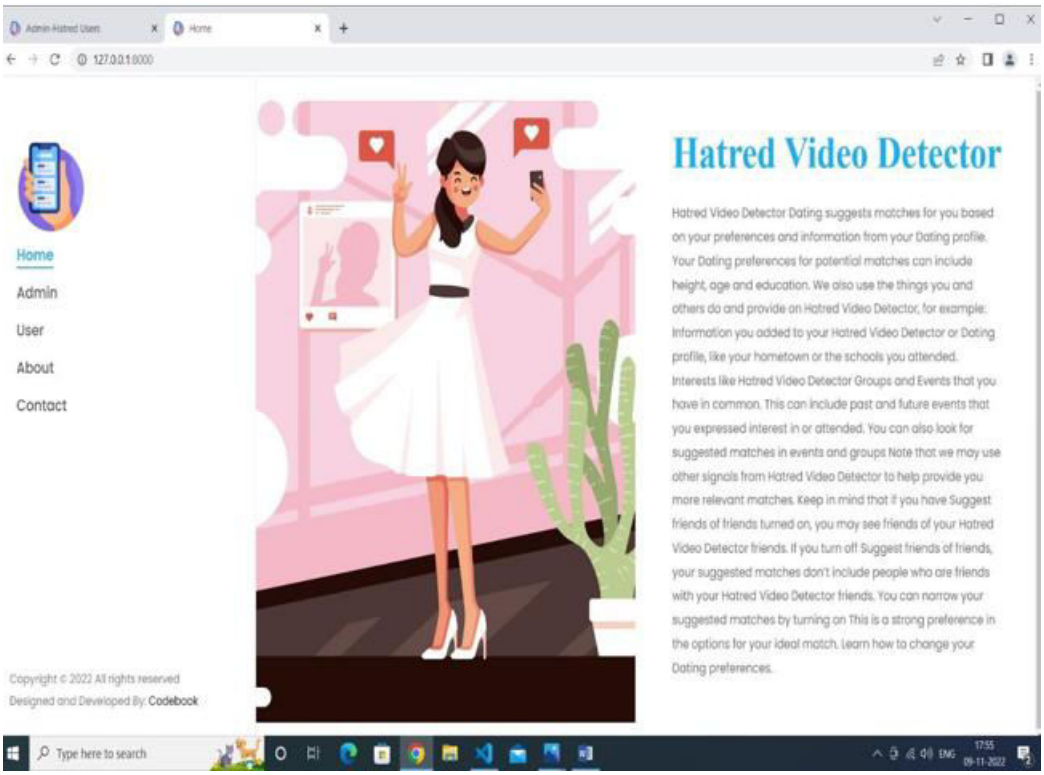
Feature Extraction and Representation:

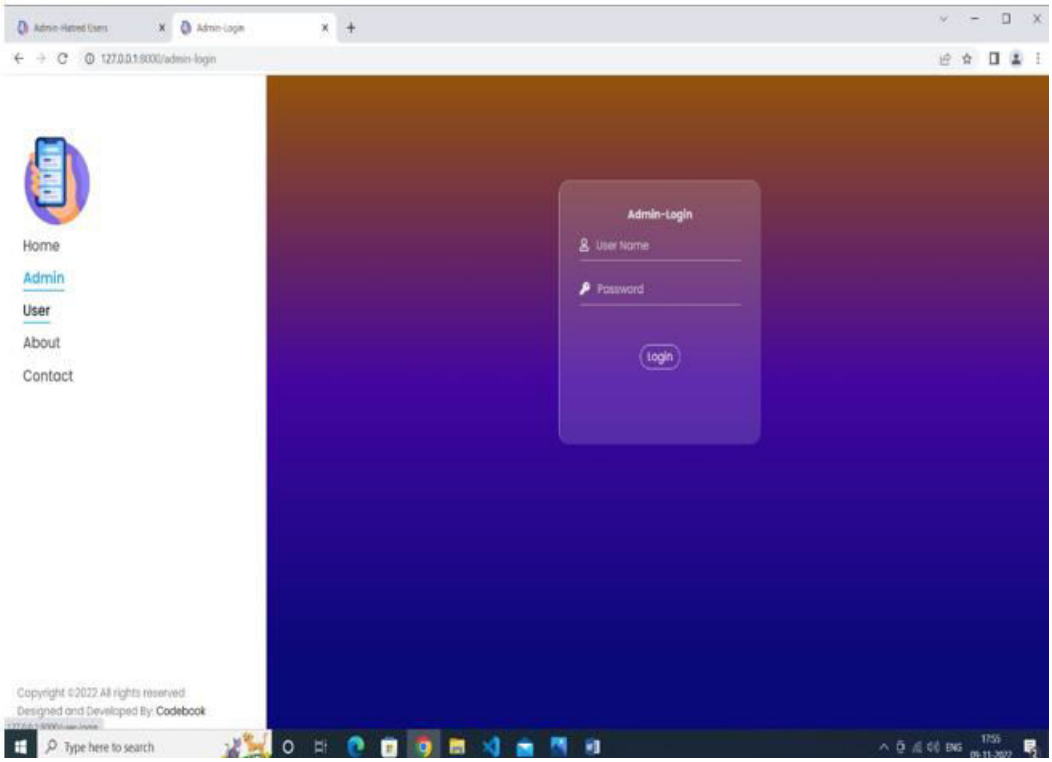
In this module, audio-based features such as Mel Frequency Cepstral Coefficients (MFCCs), Spectral Centroid, Rolloff, Bandwidth, Zero-Crossing Rate, and Chroma values are extracted from the preprocessed audio data. The goal is to create a comprehensive set of features that characterizes the unique aspects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

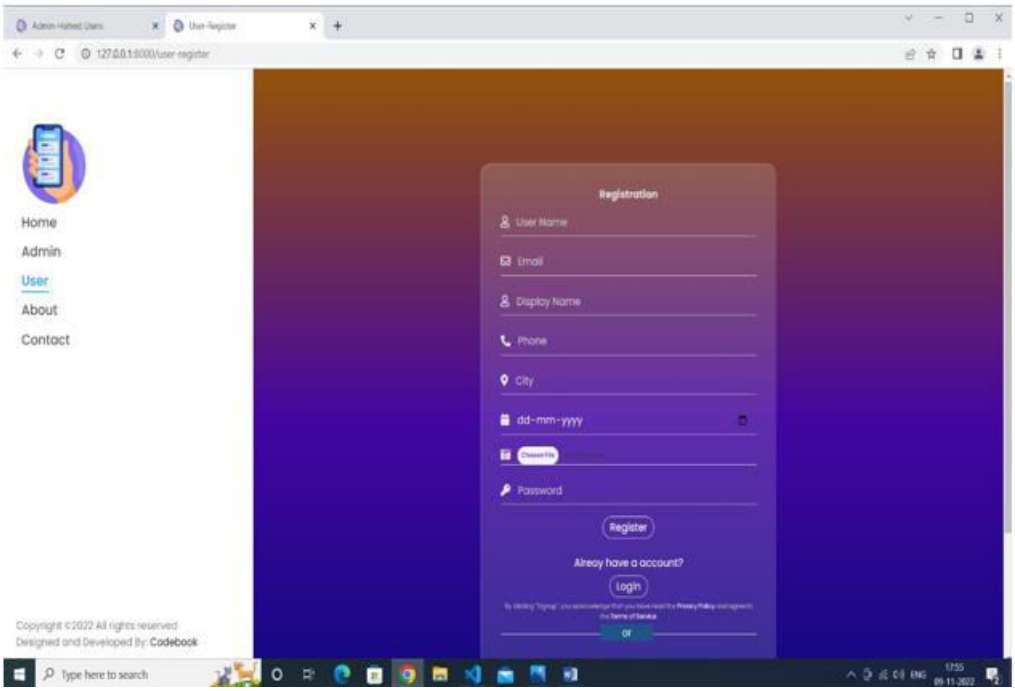
Model Training and Optimization:

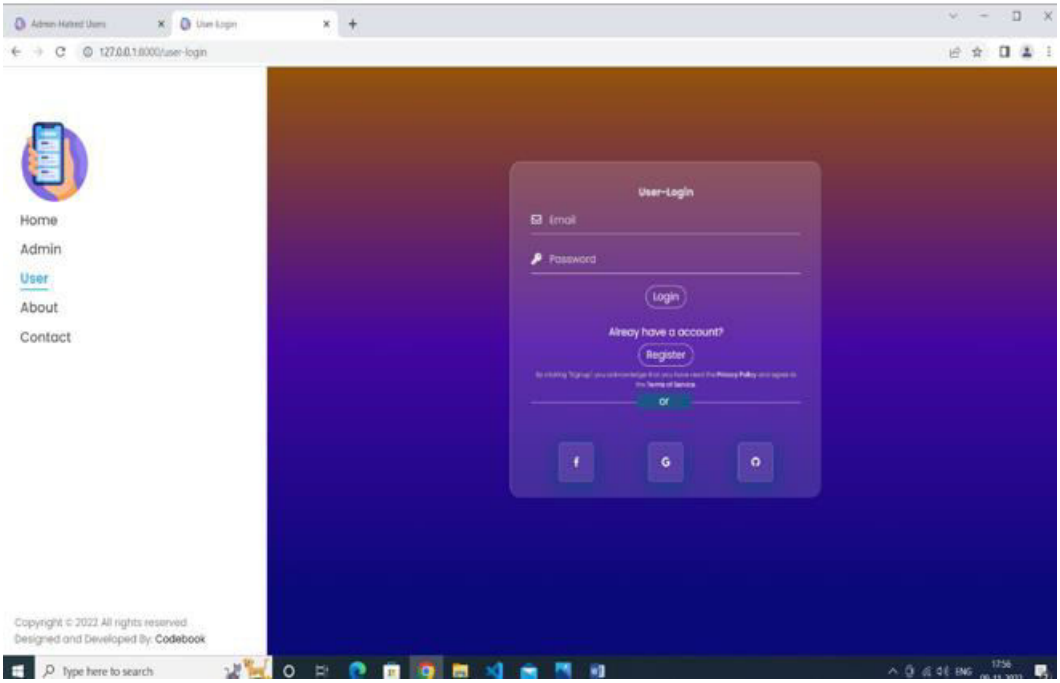
The training module involves the implementation and optimization of machine learning models, possibly utilizing advanced deep learning architectures like recurrent neural networks (RNNs) or transformers. The models are trained on the extracted features, and hyperparameters are tuned to achieve the highest possible accuracy in hate speech classification. 31 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.

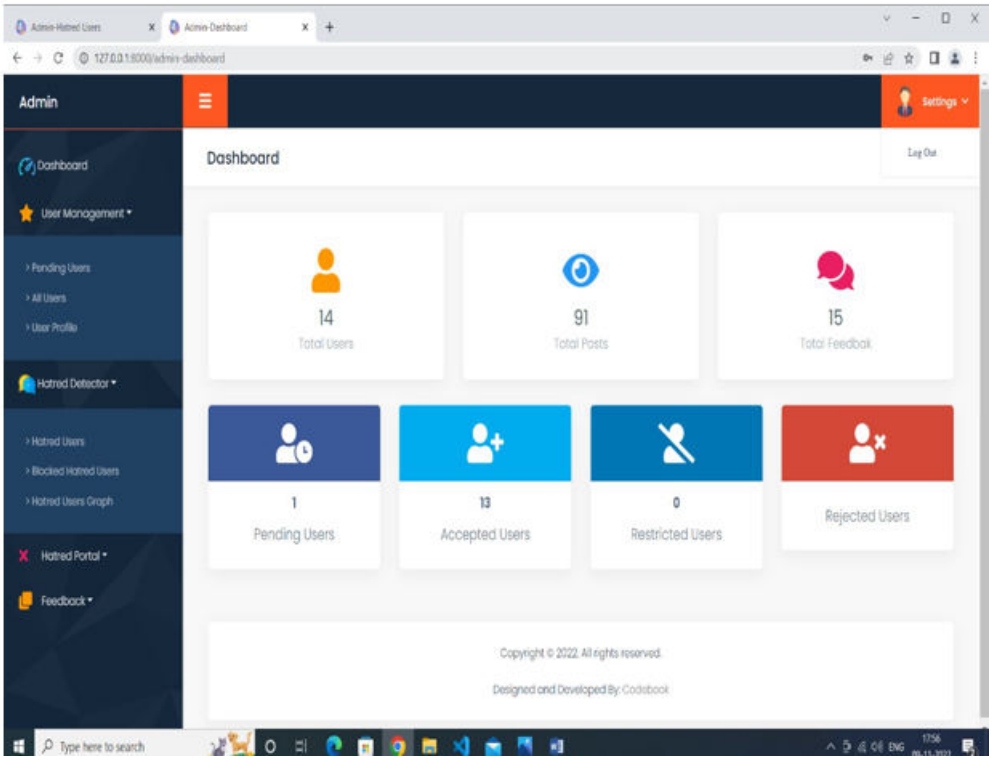
Multimodal Integration and Analysis: The multimodal module focuses on integrating audio features with additional modalities such as video and potentially textual data. This integration provides a more comprehensive understanding of the context surrounding hate speech. The system analyzes and weighs the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

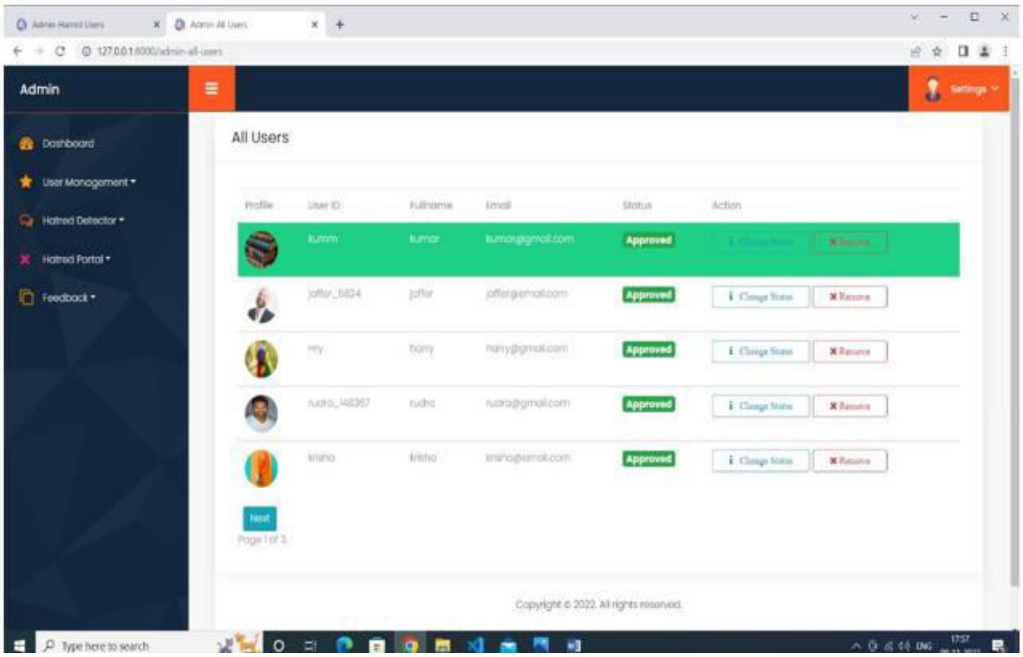

Hatred Video Detector
Home
Admin
User
About
Contact
Hatred Video Detector Dating suggests matches for you based on your preferences and information from your Dating profile. Your Dating preferences for potential matches can include height, age and education. We also use the things you and others do and provide on Hatred Video Detector, for example: Information you added to your Hatred Video Detector or Dating profile, like your hometown or the schools you attended. Interests like Hatred Video Detector Groups and Events that you have in common. This can include past and future events that you expressed interest in or attended. You can also look for suggested matches in events and groups Note that we may use other signals from Hatred Video Detector to help provide you more relevant matches. Keep in mind that if you have Suggest friends of friends turned on, you may see friends of your Hatred Video Detector friends. If you turn off Suggest friends of friends, your suggested matches don't include people who are friends with your Hatred Video Detector friends. You can narrow your suggested matches by turning on This is a strong preference in the options for your ideal match. Learn how to change your Dating preferences.
Copyright © 2022 All rights reserved
Designed and Developed By: Codebook
Type here to search

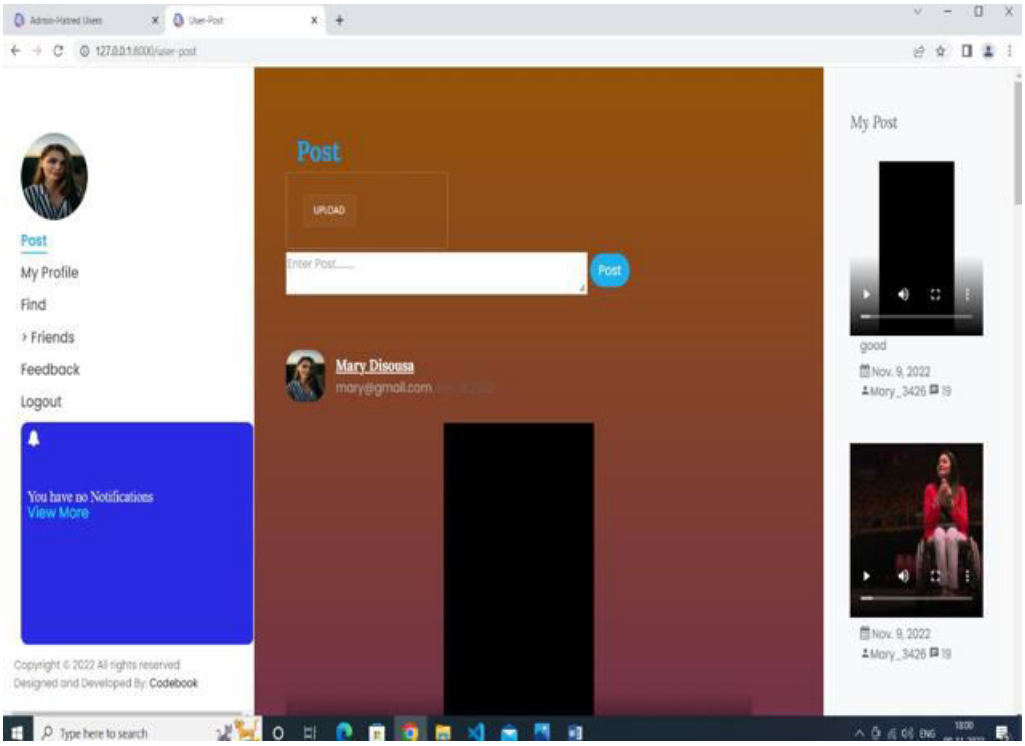

Home
Admin
User
About
Contact
Admin-Login
User Name
Password
Login
Copyright ©2022 All rights reserved
Designed and Developed By: Codebook
Type here to search


Home
Admin
User
About
Contact
Registration
User Name
Email
Display Name
Phone
City
dd-mm-yyyy
Password
Register
Already have a account?
Login
or
Copyright ©2022 All rights reserved
Designed and Developed By: Codebook
Type here to search


Home
Admin
User
About
Contact
User-Login
Email
Password
Login
Already have a account?
Register
or
Copyright © 2022 All rights reserved
Designed and Developed By: Codebook
Type here to search


Admin
Dashboard
User Management
Pending Users
All Users
User Profile
Hatred Detector
Hatred Users
Blocked Hatred Users
Hatred Users Graph
Hatred Portal
Feedback
Settings
Log Out
Dashboard
14
Total Users
91
Total Posts
15
Total Feedback
1
Pending Users
13
Accepted Users
0
Restricted Users
Rejected Users
Copyright © 2022. All rights reserved.
Designed and Developed By: Codebook


Admin
Dashboard
User Management
Hatred Detector
Hatred Portal
Feedback
Settings
All Users
Profile
User ID
Fullname
Email
Status
Action
Approved
Next
Copyright © 2022. All rights reserved.



CONCLUSION

In this study, the researchers 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.

FUTURE SCOPE

The development of an audio-based hate speech detection system, as demonstrated in this study, opens numerous avenues for future research and practical applications. A key area of expansion is the enhancement of detection accuracy through the integration of advanced machine learning techniques, such as deep learning models like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), which can better capture the temporal and spectral dynamics of speech signals. Furthermore, extending the dataset to include diverse languages and dialects beyond Filipino would enable a more robust and globally applicable system, catering to the multilingual nature of online

platforms like identifying and flagging harmful content before it reaches a broad audience. The inclusion of multimodal features, such as combining audio analysis with video and text content, could further improve the detection accuracy by leveraging contextual information. Additionally, ethical considerations and bias mitigation strategies can be incorporated to ensure fairness and transparency in hate speech detection, addressing concerns of over-policing or misclassification. This system can also be adapted for broader applications, such as monitoring online harassment, educational tools for promoting digital etiquette, and aiding law enforcement in tackling cybercrimes.

REFERENCES

- W3 School – (https://www.w3schools.com/python/)
- Geek for Geeks – (https://www.geeksforgeeks.org/python-programming- language/learn python-tutorial/)
- Python Official Documentation – (https://docs.python.org/3/tutorial/)
- Tutorials Point – (https://www.tutorialspoint.com/python/index.htm)
- Real Python – (https://realpython.com/)
- Django for Beginners – (https://djangoforbeginners.com/introduction/)
- Guru99– (https://www.guru99.com/django-tutorial.html