iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67184

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

ToxiSafe: Hate Speech Detection System

Vijayshekhar Aratagi[1], Vedavyas N[2], Srinandu K[3], Venu K[4], Hosamani Manikeshwari[5]
Ballari Institute of Technology and Management

Abstract: ***In today's interconnected digital world, the internet serves as a vital platform for communication and information sharing. However, it has also become a source of harmful content, including hate speech expressions intended to demean or discriminate based on identity. This project addresses the growing challenge of hate speech in online spaces by developing an advanced detection system. Leveraging the BERT model, the application accurately identifies hate speech in textual content. Furthermore, it incorporates multimedia analysis using tools like Pydub, MoviePy, and Speech Recognition to transcribe audio and video content for processing, enabling detection across diverse formats. A standout feature of the system is its integration of LIME (Local Interpretable Model-agnostic Explanations), which enhances transparency by highlighting specific words or phrases contributing to flagged content. Built on Flask, the system ensures user-friendliness, delivering results in an accessible format. This project has wide applications, from moderating social media platforms to aiding researchers and educators. It represents a meaningful step toward fostering respectful digital interactions, combining technological innovation with ethical responsibility***

I. INTRODUCTION

In today's fast-paced digital era, the internet has become an integral part of our lives, connecting people worldwide and enabling the exchange of ideas and information. However, despite its many benefits, the internet has also become a hub for harmful content, including hate speech. Hate speech—words or expressions intended to demean, insult, or discriminate against individuals based on their identity—has the potential to harm individuals, create divisions, and foster negativity in society. As digital spaces expand, addressing the challenge of hate speech has become a priority. This project introduces a powerful and user-friendly solution for detecting and understanding hate speech across multiple forms of media. At its core is BERT (Bidirectional Encoder Representations from Transformers), an advanced natural language processing model that excels in interpreting the context and meaning of text with high accuracy. The system processes inputs, whether plain text, audio, or video, and classifies them as "Hate Speech" or "Not Hate Speech." Beyond detection, the system integrates LIME (Local Interpretable Model-agnostic Explanations) to provide detailed explanations by highlighting specific words or phrases responsible for the classification, fostering transparency and trust. Moreover, the application extends its functionality to handle multimedia content, including audio and video. With tools like Pydub, MoviePy, and Speech Recognition, it transcribes spoken words into text for analysis using the BERT model. Built on Flask for ease of use, the system ensures results are presented in an accessible format, making it ideal for moderating hate speech in today's diverse digital landscape

II. LITERATURE REVIEW

A. Multimodal Hate Speech Detection Using Deep Learning and Transfer Learning

This paper investigates automated techniques for classifying hate speech across multiple media formats on social media, including text, voice, images, and video. The study focuses on leveraging deep learning and transfer learning models, particularly analyzing data collected from Twitter. By utilizing multilingual BERT, CNN, and MLP architectures, the research demonstrates effective methods for recognizing hateful, offensive, and neutral speech in multilingual and multimodal contexts, improving accuracy in detecting hate speech across diverse formats.

B. Optical Character Recognition (OCR) for Automated Text Extraction and Digital Transformation

This study explores the role of Optical Character Recognition (OCR) systems in automating text extraction from images. It highlights how OCR technology converts scanned images into editable digital formats with high accuracy, reducing human errors and streamlining data entry processes. The paper also discusses the application of OCR in various industries, including insurance claims processing, legal document digitization, and government record management. Furthermore, advancements in OCR algorithms supporting multilingual and complex document recognition are examined, enhancing their usability in diverse scenarios.

C. A Comprehensive Review of Text-Based Hate Speech Detection Techniques

The authors present an extensive literature review on text-based hate speech detection, analyzing various machine learning and deep learning methods such as TF-IDF, lexicon-based approaches, CNNs, RNNs, and Transformer-based models like BERT. The study reviews datasets, feature extraction techniques, and evaluation metrics commonly used in hate speech detection research. It also highlights the challenges of defining and detecting hate speech due to language nuances, cultural sensitivities, and contextual variations, providing insights into best practices and open research problems in the field.

D. Computational Approaches for Hate Speech Detection in Text: A Survey of Machine Learning Algorithms

This survey explores the computational and methodological approaches used for hate speech detection in text, covering algorithms such as Support Vector Machines (SVM), Naïve Bayes, CNNs, and RNNs. The study evaluates their effectiveness in detecting various hate speech subtypes, including race, gender, and ethnicity. It also discusses key datasets, including Twitter and Reddit, and emphasizes the importance of preprocessing techniques like tokenization, stemming, and lemmatization in improving model accuracy. Additionally, the paper examines the societal and regulatory aspects of hate speech detection, analyzing how legal frameworks like EU regulations influence detection standards and algorithmic responsibility.

E. Machine Learning-Based Hate Speech Detection in Video Content.

This research focuses on the challenges of detecting hate speech in video content, particularly in handling non-textual data such as audio and visual elements. The authors propose a machine learning-based pipeline that extracts audio from videos, converts speech to text, and classifies transcribed content as hateful or non-hateful. They employ sentiment analysis to assess tone and utilize algorithms like Naïve Bayes and Random Forest for classification. The study also highlights dataset challenges, emphasizing the need for higher-quality audio processing and diverse datasets to improve the effectiveness of multimedia-based hate speech detection.

III. PROBLEM DEFINITION

With the exponential growth of social media and digital communication platforms, hate speech has become a major concern, leading to online toxicity, harassment, and societal divisions. Manual moderation of such content is not only time-consuming and labor-intensive but also prone to bias and inefficiencies. The increasing volume of user-generated content surpasses human moderation capabilities, necessitating the development of automated methods for precise hate speech detection. This project aims to enhance online safety and content moderation by implementing a machine learning-based system that effectively identifies and mitigates hate speech. By leveraging advanced Natural Language Processing (NLP) techniques, this system ensures scalable, accurate, and unbiased detection, contributing to a safer digital space.

IV. METHODOLOGY

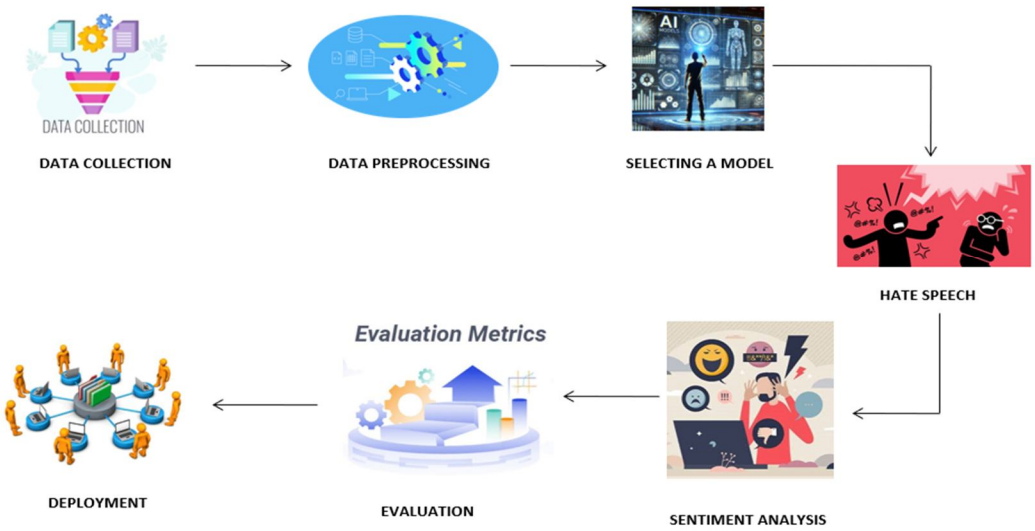


Fig1: system architecture flow.

A. Data Collection

This phase involves gathering data from various sources such as social media platforms, forums, news websites, or public comment sections. The collected data may include text, audio, or video that potentially contains hate speech. The diversity and quality of the data are crucial for building an effective detection system.

B. Data Preprocessing

Before feeding the data into the model, preprocessing is necessary to clean and prepare it for analysis. Tasks include removing irrelevant content (e.g., special characters, URLs), normalizing text (e.g., lowercasing, stemming), and handling missing or noisy data. For multimedia content, tools like Speech Recognition are used to transcribe audio and video into text for further analysis.

C. Selecting a Model

At this stage, an appropriate machine learning or deep learning model, such as BERT (Bidirectional Encoder Representations from Transformers), is chosen. This model is trained on the preprocessed data to classify content as "Hate Speech" or "Not Hate Speech." The selection depends on the complexity of the data and the desired accuracy.

D. Hate Speech Detection

The trained model identifies harmful content in the data. It classifies whether the input contains hate speech based on the context and meaning of the words or phrases.

E. Sentiment Analysis

Beyond classification, the system performs sentiment analysis to understand the tone and intent of the content. This provides additional insights into the emotional undertone of the detected hate speech.

F. Evaluation Metrics

The performance of the model is assessed using metrics such as accuracy, precision, recall, and F1-score. These metrics help fine-tune the model to improve its detection capabilities.

G. Deployment

After successful evaluation, the system is deployed as an application or API. It can then be integrated into various platforms, enabling real-time hate speech detection and moderation to create safer digital spaces.

V. RESULTS AND EVALUATION

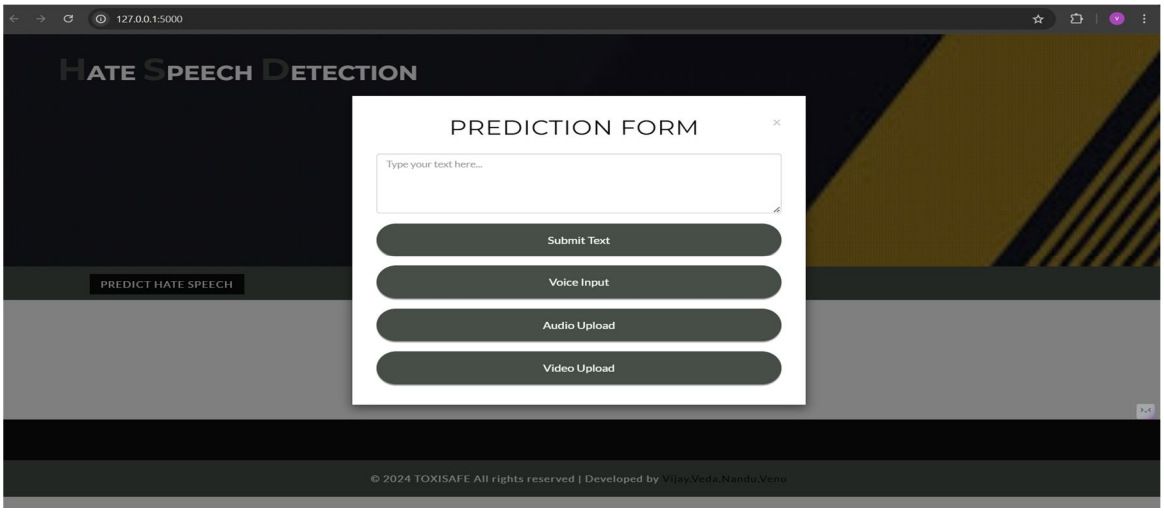


Figure *2* Home Page

This figure showcases the main interface of the hate speech detection tool. It acts as the central access point for users, enabling them to provide input in various formats, such as text, audio, or video. Through this form, users can conveniently interact with the system, choosing the most suitable input type for their needs. The interface reflects the system's flexibility in handling multiple formats, making it accessible for different use cases.

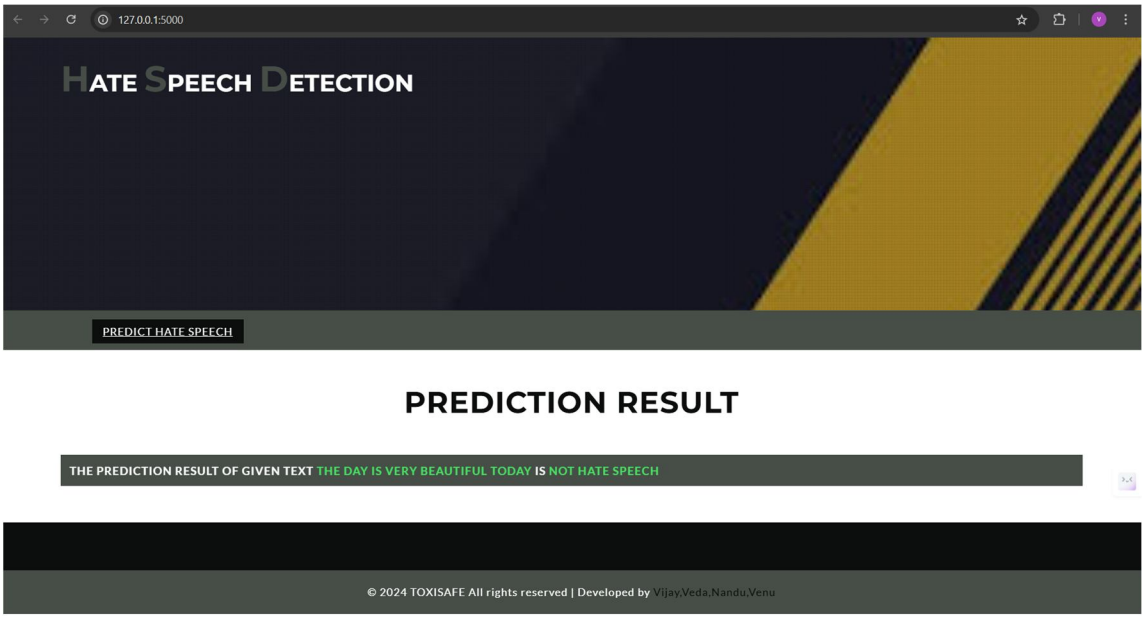


Figure 3 Textual hate speech recognition

This figure highlights how the system processes text inputs to detect hate speech. Users can enter text for analysis, and the system will classify it as either "Hate Speech" or "Not Hate Speech." This step demonstrates the core functionality of the tool, showing how it utilizes advanced text processing techniques to analyze and interpret language effectively. It reflects the model's precision in identifying patterns that signify harmful content.

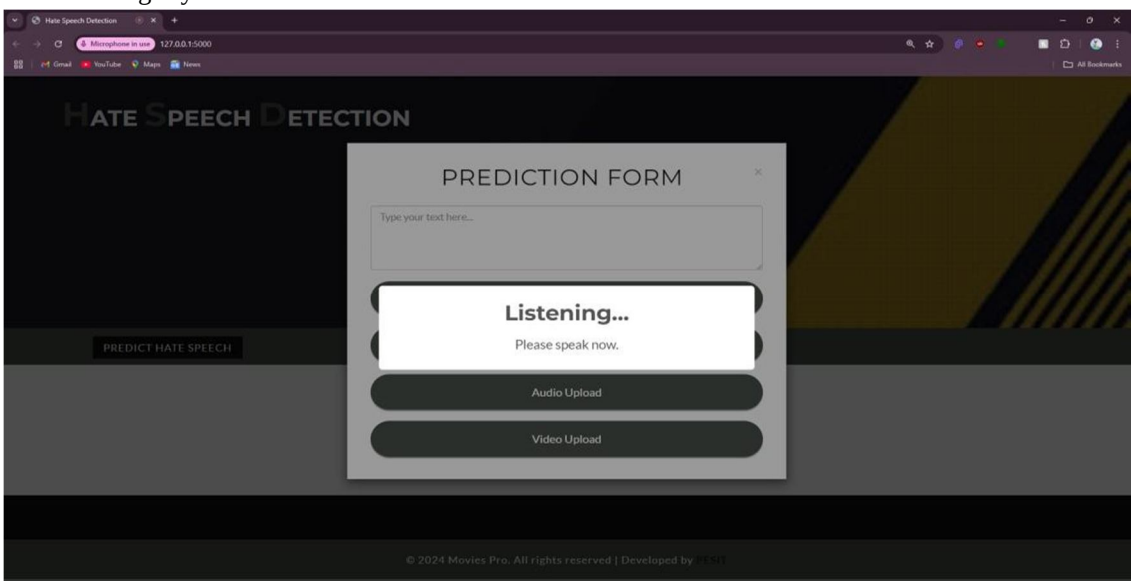


Figure 4 Voice Input for Hate speech detection

This image represents the feature that allows users to upload audio files for analysis. The system processes these files by transcribing the spoken words into text, which is then analyzed for hate speech. This capability broadens the scope of the tool, enabling it to detect harmful language conveyed through verbal communication. It's particularly useful in contexts where speech-based

Hate speech need to be moderated.

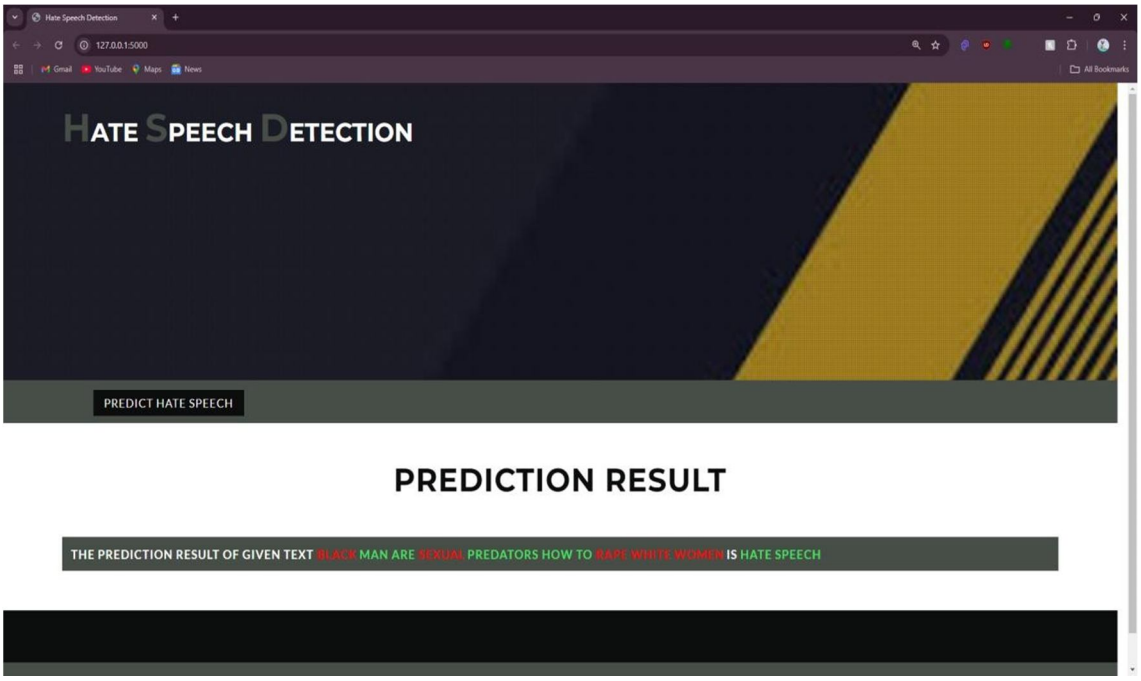


Figure 5 Prediction result for audio input.

This figure showcases the outcome of analyzing an audio file. It displays the transcribed text along with the detection result, indicating whether the content contains hate speech. This step emphasizes the integration of transcription and detection functionalities, ensuring that even spoken content can be reviewed for harmful language. It also provides transparency in the process by presenting both the transcription and the classification result.

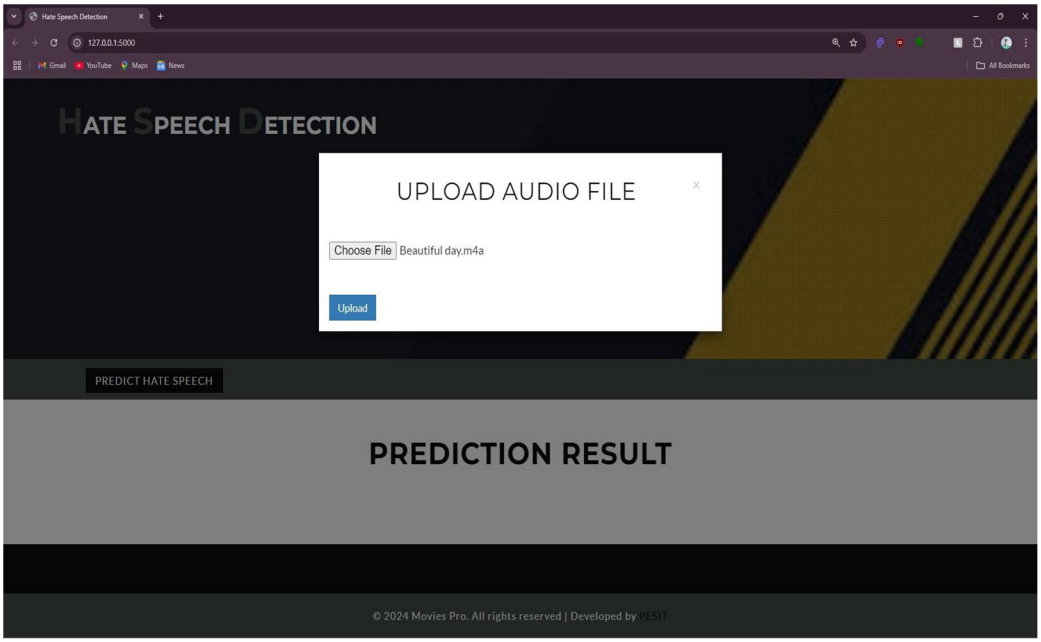


Figure 6 A Window pertaining to Audio files

This image focuses on the interface for handling audio files. It allows users to upload or manage their audio inputs, ensuring that the process is simple and user-friendly. This feature streamlines the workflow for users, making it easier to submit audio data for analysis without technical complexity. It demonstrates the system's ability to accommodate diverse user needs and efficiency.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

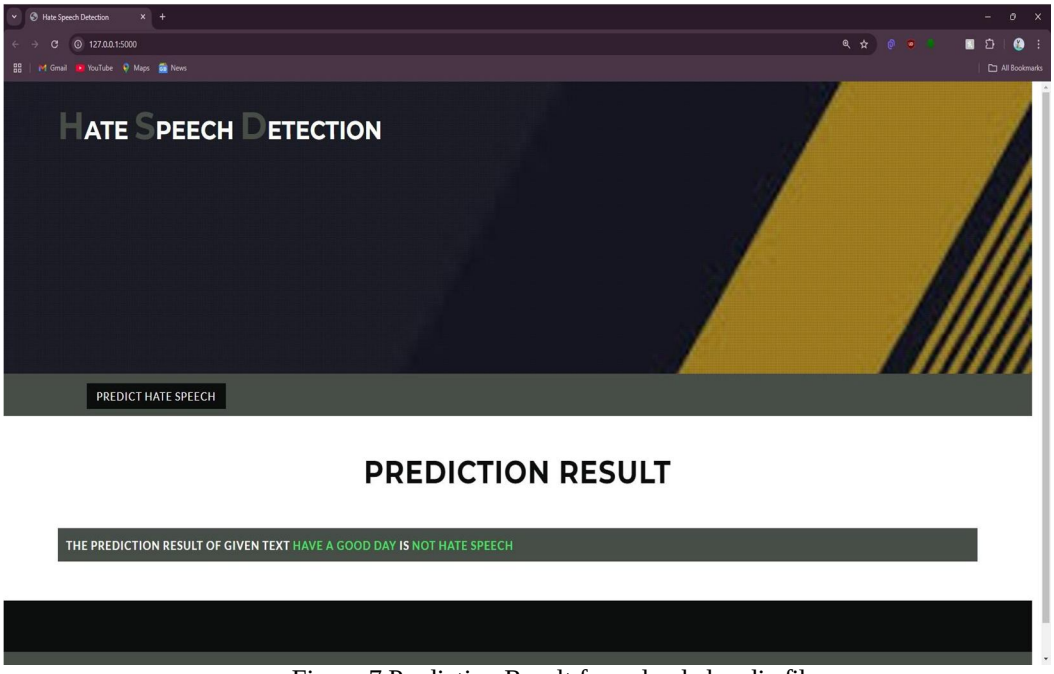


Figure 7 Prediction Result for uploaded audio file

This figure highlights the results after an audio file has been processed. It not only shows the system's classification of the audio (e.g., "Hate Speech" or "Not Hate Speech") but also includes a confidence score, indicating the reliability of the result. This provides users with additional insights into the analysis, reinforcing trust in the system's capabilities.

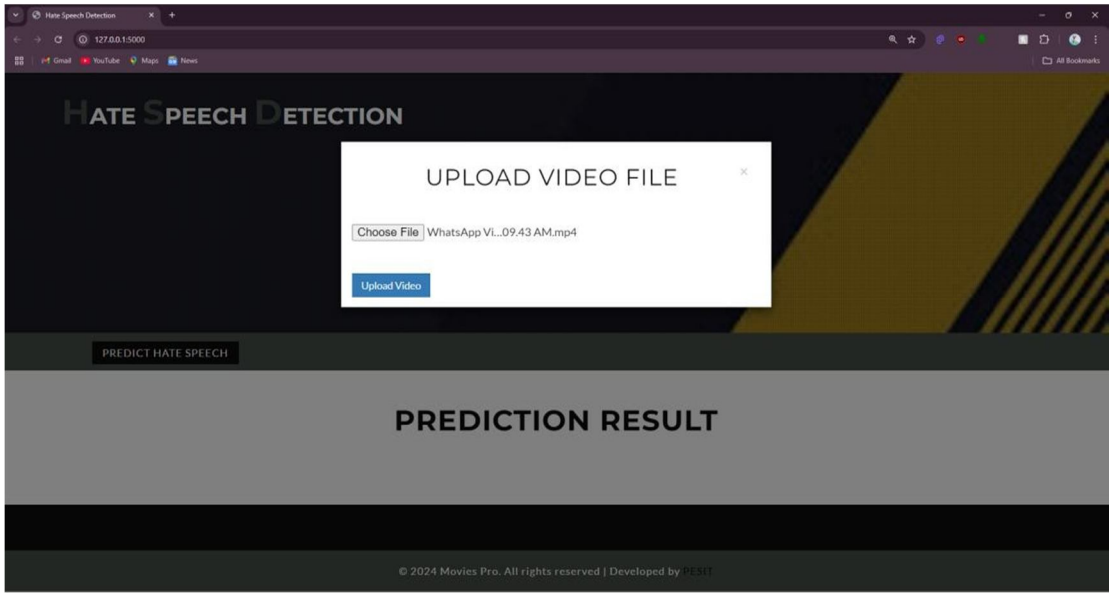


Figure 8 Video input for hate speech detection

This image illustrates the tool's ability to handle video files. It extracts audio from the video, transcribes it into text, and analyzes the content for hate speech. This feature is particularly valuable in today's digital age, where video content is a dominant form of communication. By extending its detection capabilities to videos, the system becomes a comprehensive solution for moderating multimedia content.

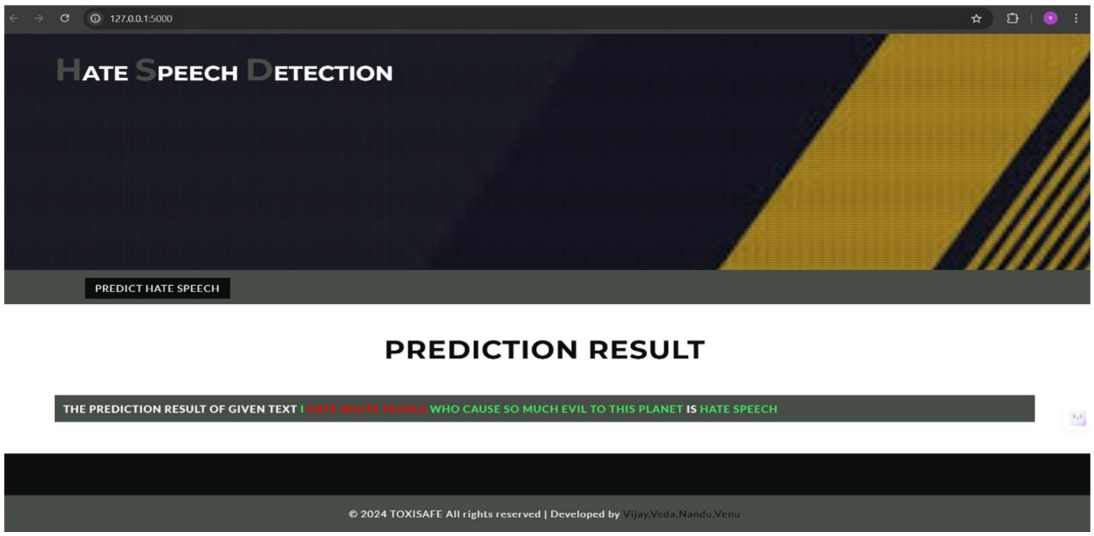


Figure 9 Prediction result for Video input

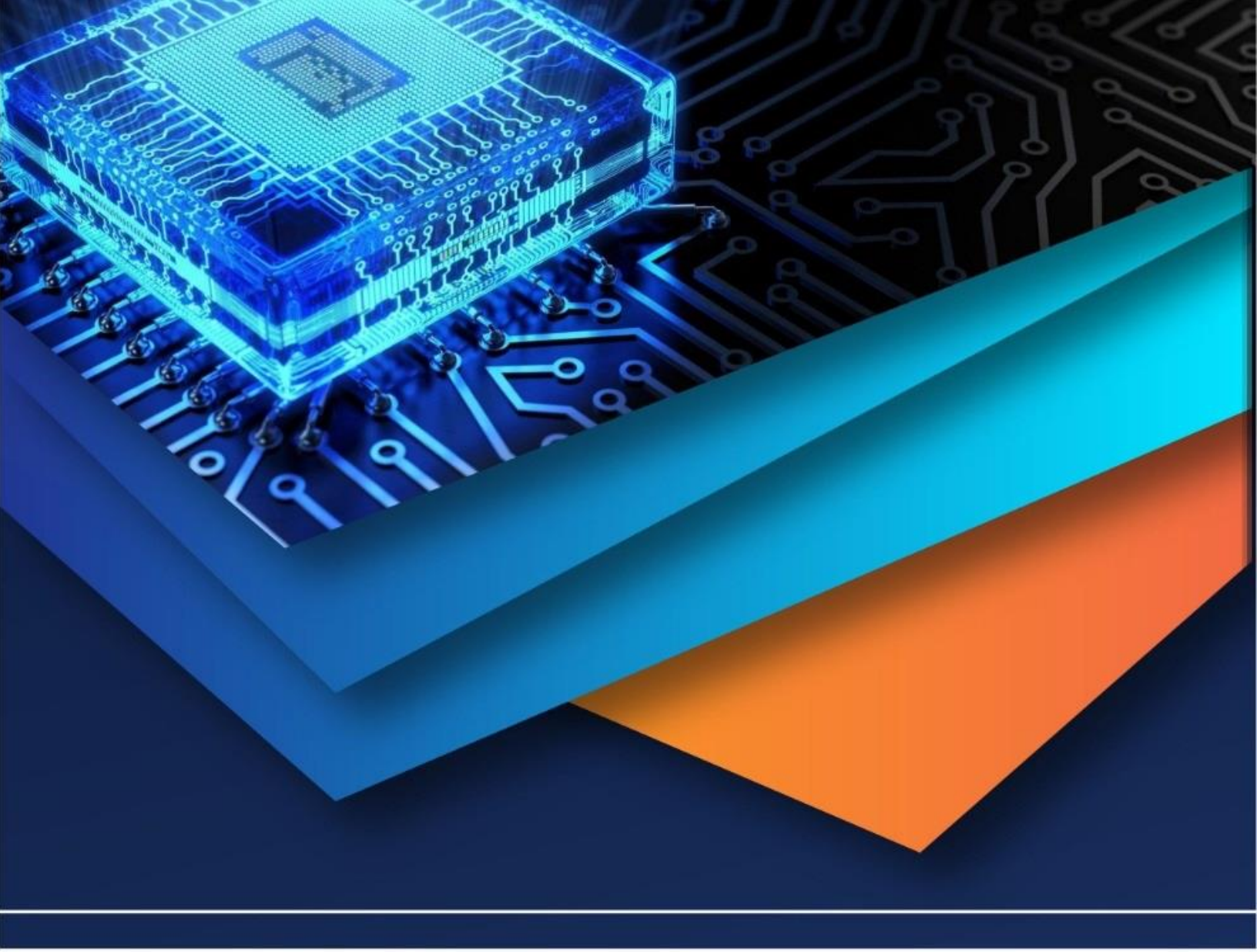
This figure highlights the results after video file has been processed. It not only shows the system's classification of the audio (e.g., "Hate Speech" or "Not Hate Speech") but also includes a confidence score, indicating the reliability of the result. This provides users with additional insights into the analysis, reinforcing trust in the system's capabilities.

VI. CONCLUSION

This project demonstrates the potential of machine learning in addressing the critical issue of hate speech in online content. Our model effectively identifies harmful language, contributing to the creation of safer and more respectful digital spaces. By leveraging advanced natural language processing techniques and diverse data, the project underscores the capability of automated solutions to support content moderation and foster healthier online interactions. The model employs sophisticated algorithms trained on diverse and representative datasets, ensuring its ability to detect nuanced forms of hate speech across different contexts, cultures, and languages. This approach not only enhances its accuracy but also highlights its adaptability to evolving patterns of online discourse. By proactively identifying problematic content, the system empowers platforms to intervene before harm escalates, safeguarding users from toxic interactions.

REFERENCES

[1] Irfan, A., & Kumar, N. (2024). Multi-Modal Hate Speech Recognition Through Machine Learning. In Proceedings of the International Conference on Advanced Computing and Applications.
[2] Mansur, Z., Omar, N., & Tiun, S. (2023). Twitter Hate Speech Detection: A Systematic Review of Methods, Taxonomy Analysis, Challenges, and Opportunities. Journal of Social Media Analytics, 15(3), 45-67.
[3] Mehta, H., & Passi, K. (2022). Social Media Hate Speech Detection Using Explainable Artificial Intelligence (XAI). In Proceedings of the International Conference on Artificial Intelligence Applications.
[4] Wu, C. S., & Bhandary, U. (2020). Detection of Hate Speech in Videos Using Machine Learning. In Proceedings of the International Conference on Multimedia Systems. [5] Mingjun Wei, Qiwei Wu, Hongyu Ji, Jingkun Wang, Tao Lyu, Jinyun Liu , and Li Zhao (2023). A Skin Disease Classification Model Based on DenseNet and ConvNeXt Fusion. Journal of Multidisciplinary Digital Publishing Institute journal.
[5] Alkomah, F., & Ma, X. (2022). A Literature Review of Textual Hate Speech Detection Methods and Datasets. Computational Linguistics Journal, 38(4), 123-145.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY