
HATEFUL MEME'S, VIDEO'S, AUDIO'S DETECTION

**Prof. Harshali Ragite^{*1}, Ms. Gauri Farde^{*2}, Ms. Gayatri Lute^{*3}, Ms. Prachi Bobade^{*4},
Ms. Heena Patle^{*5}**

^{*1,2,3,4,5}Department Of Computer Science And Engineering, Wainganga College Of Engineering & Management, Nagpur, Maharashtra, India.

ABSTRACT

The Proliferation of online social media platforms, video sharing websites, and live streaming services has led to a surge in the spread of hateful content, including hate speech, harassment, and discriminatory behavior. To combat this issue, this project aims to develop an artificial intelligence (AI) powered system for detecting hateful content in audio and video recordings. The proposed system utilizes a deep learning-based approach, combining convolutional neural networks and recurrent neural networks to analyze audio and video files and identify patterns and characteristics indicative of hateful content. The system is trained on a large dataset of labeled audio and video files and evaluated on a separate test dataset.

Keywords: Hateful Content Detection, Hate Speech Detection, Audio And Video Analysis , Artificial Intelligence Online Safety, Social Media Regulation, Hate Crime Prevention.

I. INTRODUCTION

The rapid growth of online social media platforms, Video sharing websites, and live streaming services has revolutionized the way people communicate, interact, and share information. However, this increased online activity has also led to a surge in the spread of hateful content, including hate speech, harassment, and discriminatory behavior. Hateful content can take many forms, including audio and video recordings that promote violence, discrimination, or hatred against individuals or groups based on their race, ethnicity, gender, religion, or other personal characteristics.

The spread of such content can have severe consequences, including the perpetuation of hate crimes, the marginalization of vulnerable communities, and the erosion of social cohesion.

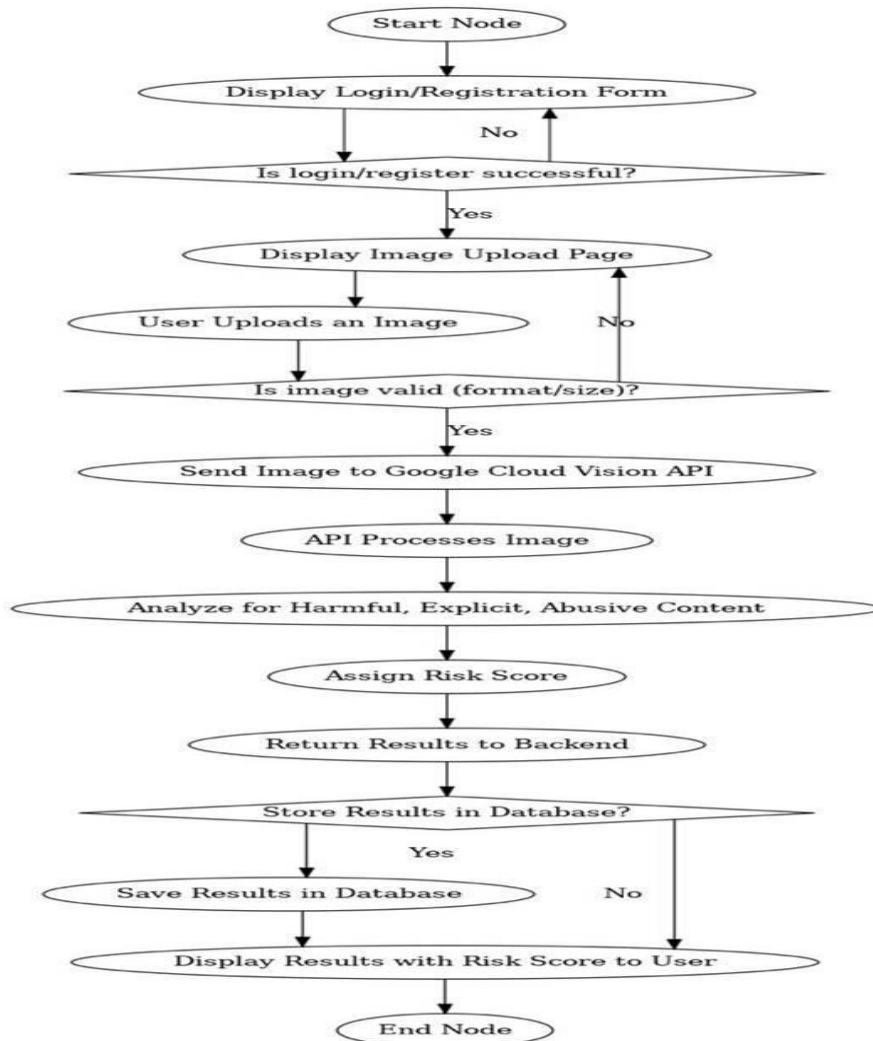
Despite the efforts of social media platforms, online communities, and law enforcement agencies to combat hateful content, the problem persists. One of the main challenges is the sheer volume of online content, which makes it difficult to detect and remove hateful material in a timely and effective manner.

II. METHODOLOGY

The methodology of the hateful memes detection project involves several stages.

1. First, a large dataset of memes is collected from various online sources, including social media platforms and online forums. The collected memes are then annotated as hateful or non-hateful by human evaluators.
2. Next, the images and text in the memes are pre-processing using techniques such as resizing, tokenization, and steaming. Features are then extracted from the preprocessed images and text using convolutional neural networks and word embedding.
3. A machine learning model, such as a CNN or recurrent neural network, is then trained on the extracted features to learn patterns and relationships between the images and text.
4. The trained model is then evaluated using metrics such as accuracy, precision, recall, and F1-score, and its performance is compared to baseline models. Finally, the trained model is deployed in a suitable environment, such as a web application or API, and its performance is continuously monitored and updated to ensure its accuracy and effectiveness.

III. MODELING AND ANALYSIS



IV. RESULTS

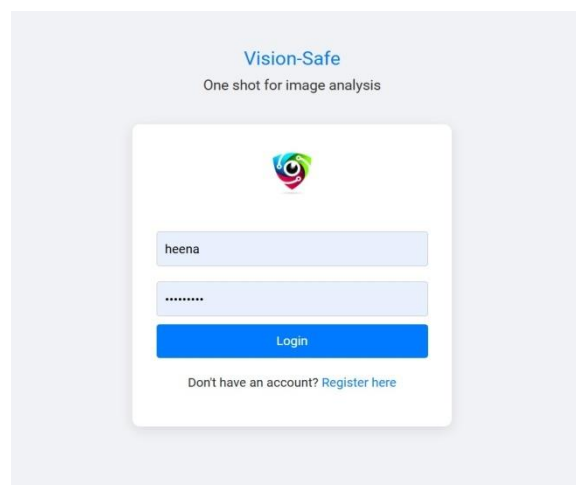
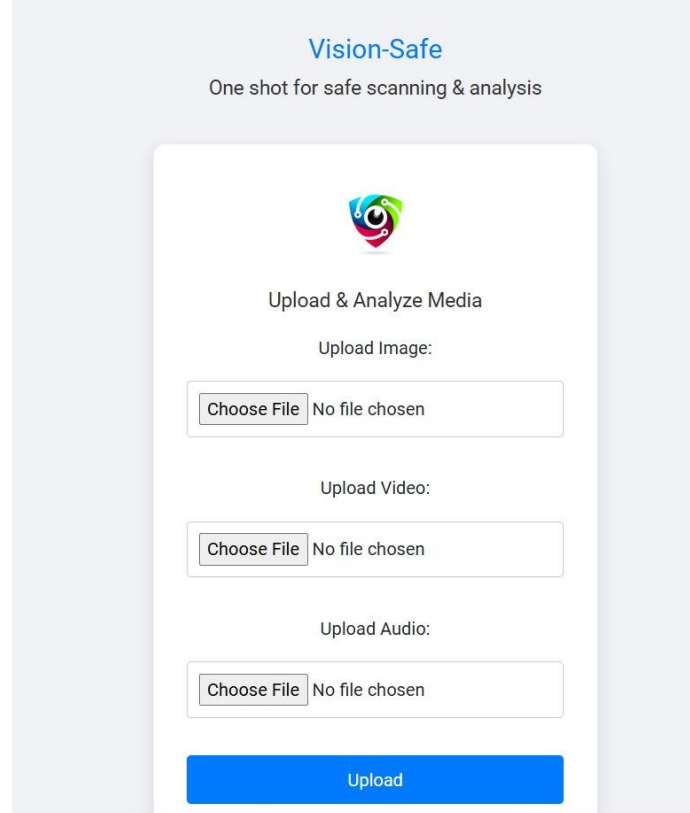


Fig.1

The interface is titled 'Vision-Safe' with the tagline 'One shot for safe scanning & analysis'. It features a central white box with a colorful eye-like logo. Below the logo, the text 'Upload & Analyze Media' is displayed. There are three sections for file uploads: 'Upload Image:', 'Upload Video:', and 'Upload Audio:'. Each section has a 'Choose File' button and a text area that currently says 'No file chosen'. At the bottom of the white box is a large blue 'Upload' button.

Vision-Safe

One shot for safe scanning & analysis

Upload & Analyze Media

Upload Image:

Choose File No file chosen

Upload Video:

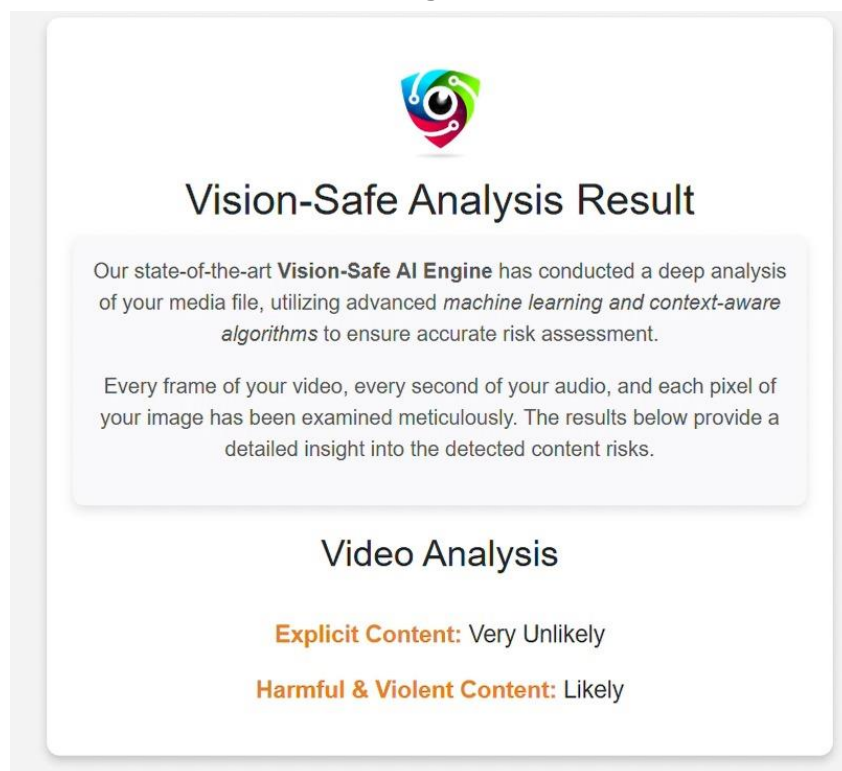
Choose File No file chosen

Upload Audio:

Choose File No file chosen

Upload

Fig.2

The interface shows the 'Vision-Safe Analysis Result'. It starts with the same colorful eye-like logo. Below it, the title 'Vision-Safe Analysis Result' is centered. A text box explains that the 'Vision-Safe AI Engine' has conducted a deep analysis using 'machine learning and context-aware algorithms'. Another text box states that every frame of the video, every second of the audio, and each pixel of the image have been examined. Below this, the section 'Video Analysis' is shown, followed by two results: 'Explicit Content: Very Unlikely' and 'Harmful & Violent Content: Likely'.

Vision-Safe Analysis Result

Our state-of-the-art **Vision-Safe AI Engine** has conducted a deep analysis of your media file, utilizing advanced *machine learning and context-aware algorithms* to ensure accurate risk assessment.

Every frame of your video, every second of your audio, and each pixel of your image has been examined meticulously. The results below provide a detailed insight into the detected content risks.

Video Analysis

Explicit Content: Very Unlikely

Harmful & Violent Content: Likely

Fig.3

V. CONCLUSION

The Hateful Audio and Video Detection Project aimed to develop an artificial intelligence powered system to detect and flag hateful content in audio and video files. The project's primary objective was to create a robust and accurate system that can identify hate speech, harassment, and discriminatory behavior in online content.

Throughout the project, we explored various machine learning and deep learning techniques, including convolutional neural networks, recurrent neural networks, and transfer learning. We also investigated the use of natural language processing techniques to analyze text-based content and identify hate speech.

VI. REFERENCES

- [1] Williams, ML, Burnap P, Javed A, Liu H, Ozalp S. Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. The British Journal of Criminology. 2020; 60(1):93-117.
- [2] Schieb C, Preuss M, editors, Governing hate speech by means of counter speech of face book. 66th ica annual conference, at fukuoka, japan; 2016.
- [3] Konikoff D. Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies. Policy & Internet. 2021.
- [4] Newton C. The secret lives of Facebook moderators in Amercia. The Verge. 2019; 25