CENSORSHIP FOR ABUSIVE CONTENT ON WEB

Thidiksh M.S[1], Prof. Mahalakshmi M[2], Prof. Arvind G[3]

[1] *PG student, Master of Computer Application, Maharaja Institute of Technology Mysore, Karnataka, India*

[2] *Assistant Professor, Master of Computer Application, Maharaja Institute of Technology Mysore, Karnataka, India*

[3] *Assistant Professor, Master of Computer Application, Maharaja Institute of Technology Mysore, Karnataka, India*

ABSTRACT

In the modern digital age, the internet has become an integral part of everyday life, providing an open platform for free expression and exchange of information. However, with the tremendous growth of online content, there has been an alarming increase in the dissemination of abusive and harmful material, including hate speech, cyberbullying, harassment, and explicit violence. Such content poses significant threats to individuals' safety, mental well-being, and social cohesion. The implementation of effective censorship for abusive content on the web requires a multi-faceted approach that considers technological advancements, user feedback, ethical principles, and legal norms. By striking a balance between safeguarding online safety and preserving free expression, a safer and more inclusive digital environment can be fostered for present and future generations.

Keyword: - *Keyword Matching, Keyword Storage, Age Classification.*

1. INTRODUCTION

The rapid growth and ubiquity of the internet have revolutionized the way people communicate, share information, and interact globally. The digital age has brought about unprecedented opportunities for creativity, collaboration, and knowledge dissemination. However, this transformation has also brought with it a dark side - the proliferation of abusive and harmful content on the web. The unchecked dissemination of hate speech, cyberbullying, harassment, and violent material has become a pressing concern, posing significant threats to individuals, communities, and the social fabric at large.

To deal with this issue, we are building a model in which we can classify the content based on its description and title and by using an age detection model we will detect the age of the user and by Using ML models and datasets we will match the content and the users age and take appropriate action. The system also censors the abusive comments and will block it from being posted.

2. LITERATURE SURVEY

Machine Learning-Based Detection of Cyber-bullying on Social-Media

[1] "Machine Learning- Based Detection of Cyberbullying on Social-Media" by A. Gupta and R. Singh focuses on machine learning-based approaches to detect abusive content on social media. It examines the features and classifiers used for cyberbullying detection and evaluates the accuracy and efficiency of different methods. The study highlights the potential of artificial intelligence in content moderation.

The Role of Internet Intermediaries in Reducing Harmful Content Online

[2] "The Role of Internet Intermediaries in Reducing Harmful Content Online" by B. Smith and C. Davis explores the roles and responsibilities of internet intermediaries, such as social media platforms and internet service providers, in reducing harmful content. It discusses the legal and policy frameworks that shape their content moderation practices and analyzes the challenges faced by intermediaries in tackling abusive content while respecting user rights.

Automated Hate Speech Detection and the Problem of Offensive Language

[3] "Automated Hate Speech Detection and the Problem of Offensive Language" by Schmidt, Anna, and Wiegand, Michael revolves around investigating the use of natural language processing techniques for hate speech detection. The paper evaluates the effectiveness of machine learning models in identifying offensive language and discusses the complexities in defining and detecting abusive content.

The literature survey presented above provides a glimpse of the diverse research conducted on the topic of censorship for abusive content on the web. These studies shed light on the complexities of the issue, the impact on individuals and society, and the various technological, ethical, and regulatory challenges involved in content moderation. Together, they contribute to an informed understanding of the problem and pave the way for further exploration and improvement in addressing abusive content online.

3. PROPOSED SYSTEM

The user can either sign-in with their Google Account or create a new account through Firebase Authentication. After successful sign-in or registration the user allows access to their webcam for image capture. The webcam image along with other user details, is stored in firebase Realtime Database. The user is then redirected to Content Search Window where they can enter keywords for content search. The entered content is stored for content filtering. When the user presses the search button, the Content Search Window retrieves the users age prediction and the entered keyword from Firebase Realtime Database. The main Joblib file processes the users age and keyword with the content age group dataset stored in Firebase Realtime database. Firebase Realtime Database is updated with the permission status (Allowed or Not Allowed) for displaying the content or showing the message "You are not eligible for this content". Finally, the Content Display Window displays either the eligible content or the ineligibility message based on the permission status.

The application also employs the machine learning to detect and censor abusive and inappropriate comments on the platform using machine learning models.

3.1 System Architecture

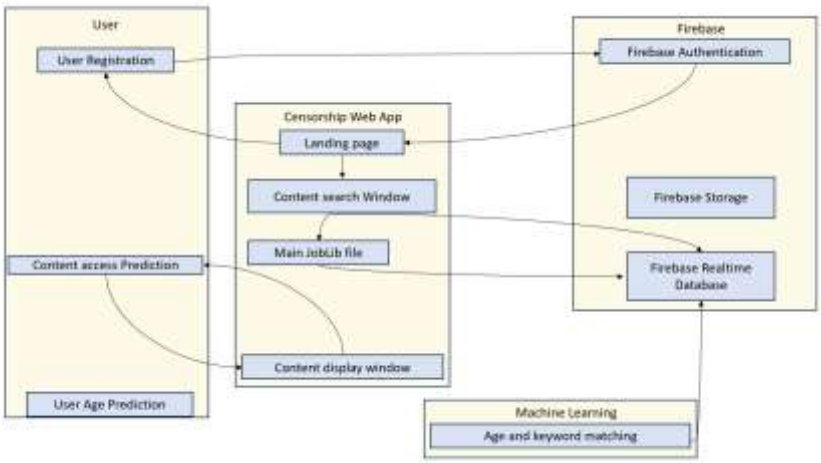

Fig-1: System Architecture

A. Video Classification

The description and title of the video are examined for their content. A CNN architecture like age detection module is used to train the content classification model. The trained content classification model receives the preprocessed content descriptions and titles and predicts the classification of the content (such as positive, negative, or abusive). The categorization aids in processing and filtering content according to its type and appropriateness

B. Age Prediction

A dataset of image files is needed for the projects age detection and content classification. Preprocessing of these images include scaling them to a standard size and leveling the pixels values to [0,1] range. The age detection model developed using Convolutional Neural Network (CNN). The CNN architecture consists of convolutional and pooling layers followed by fully connected layers. The preprocessed images fed into the trained age detection model which predicts the age of the user based on the image. The predicted age is used to determine whether the user is eligible to access certain content.

C. Comment Censorship

Like the age detection and content classification algorithms, the comment censoring model is trained using CNN architecture. The trained comment censorship model receives input from the preprocessed comments and forecasts how offensive or improper a comment is. The model aids in recognizing and creating a safer environment for users by eliminating harmful remarks from the platform.

D. Content Handling

Based on the predicted age and content classification, the platform decides how to handle the content. If the user's age is within a certain range, they are allowed to access specific content. Otherwise, access is restricted. The content files are displayed, processed, and filtered based on the classification, ensuring that inappropriate or abusive content is not shown to users.

4. CONCLUSIONS

On social media, inappropriate and abusive content can have major repercussions for the users who come across it and can foster a hostile and toxic environment. To make their platforms a secure and friendly place for its users, social media is a continuous task that calls for a combination of technological and manual moderation. Social media platforms need to take this issue seriously and create appropriate solutions.

5. REFERENCES

[1] *A. Gupta, R. Singh, "Machine Learning-Based Detection of Cyberbullying on Social Media", International Conference on Big Data Analytics, 2019*

[2] *B. Smith, C. Davis the Role of Internet Intermediaries in Reducing Harmful Content Online Policy & Internet, 2021*

[3] *Schmidt, Anna, and Wiegand, Michael, "Machine Learning-Based Detection of Cyberbullying on Social-Media", Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*