Mini Project Report On

TOXITRACK AI(A comment toxicity analyzer and filter)

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science & Engineering

By

Basil Sabu (U2103059)

Basil Eldo (U2103058)

Avin Pulikkan (U2103056)

Christo Shaju (U2103069)

Under the guidance of

Ms.Sherine Sebastian

Department of Computer Science & Engineering Rajagiri School of Engineering & Technology (Autonomous) (Affiliated to APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039

May 2024

CERTIFICATE

This is to certify that the mini project report entitled ”TOXITRACK AI” is a bonafide record of the work done by Basil Sabu (U2103059), Basil Eldo (U2103058), Christo Shaju (U2103069), Avin Pulikkan(U2103056), submitted to the APJ Ab dul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.

Ms.Sherine Sebastian Mr Harikrishnan M Project Guide Project Coordinator
Assistant Professor Assistant Professor Dept. of CSE Dept. of CSE RSET
RSET

Dr.Preetha K.G.
Head of the Department
Designation
Dept. of CSE
RSET

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude towards Dr P. S. Sreejith, Principal of RSET, and Dr. Preetha K.G., Head of the Department of Computer Science and Engineering for providing me with the opportunity to undertake my mini project, ”TOXITRACK AI”.

I am highly indebted to my project coordinator Mr.Harikrishnan M, Assistant Pro fessor, Department of Computer Science and Engineering and Ms.Sherine Sebastian ,Assistant Professor,Department of Computer Science for their valuable support.

It is indeed my pleasure and a moment of satisfaction for me to express my sincere grat itude to my project guide Ms.Sherine Sebastian for her patience and all the priceless advice and wisdom she has shared with me.

Last but not the least, I would like to express my sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Basil Sabu

Basil Sabu

Avin Pulikkan

Christo Shaju

Abstract

TOXITRACK AI is an advanced comment analysis tool designed to enhance online communication by detecting and managing toxicity in user interactions. This project focuses on identifying and quantifying various toxicity parameters, including identity hate, insult, obscene language, severe toxicity, and threats. Using a

sophisticated machine learning model, TOXITRACK AI analyzes each comment entered by users and provides a detailed percentage breakdown of these toxicity parameters.

In addition to the analysis feature, TOXITRACK AI incorporates a chat space where users can create rooms to discuss diverse topics. This public chat space allows people to communicate safely by continuously monitoring comments for toxic content. The comments in these chat rooms are analyzed in real-time by the model, and if any toxicity parameter exceeds a threshold the overall toxicity of the chat is flagged, and the offending comment is blurred to maintain a healthy conversation environment.

The graphical representation of toxicity levels provides users with a clear understand ing of the nature and extent of toxicity in their comments, promoting self-awareness and encouraging more respectful interactions. TOXITRACK AI aims to foster a safer and more positive online community by leveraging artificial intelligence to support effective content moderation and enhance user experiences, creating a public chat space where people can communicate freely and safely.

Contents

List of Figures

List of Tables

List of Abbreviations

Chapter 1

Introduction

1.1 Background

The internet and social media have revolutionized global communication, but they also face challenges with toxic comments that disrupt conversations and harm users.TOXITRACK AI addresses this problem by using advanced machine learning to detect and analyze toxic comments in real-time. The project provides graphical representations of various toxicity parameters, helping users understand and manage online behavior. A unique feature of TOXITRACK AI is its public chat space, where users can create rooms for discussions. The system monitors these chats, and if any comment's toxicity exceeds toxicity level it blurs the comment to maintain a respectful environment.

1.2 Problem Definition

The aim of TOXITRACK AI is to detect and manage toxic comments in online interac tions by providing real-time analysis of various toxicity parameters and ensuring a safe and respectful communication environment in public chat spaces.

1.3 Scope and Motivation

TOXITRACK AI will develop a tool to detect and quantify toxicity in comments, cov ering identity hate, insults, obscene language, severe toxicity, and threats. It will feature a user-friendly interface with graphical toxicity representations and a public chat space where users can create rooms and chat. The system will blur comments exceeding a set toxicity threshold, ensuring a respectful environment. The project aims to be adaptable and scalable for integration with various online platforms.

The motivation for TOXITRACK AI is to address the widespread issue of toxic

online behavior, which disrupts communities and harms individuals. Traditional moderation methods struggle with the volume and complexity of content. By using AI for real-time analysis and moderation, TOXITRACK AI aims to create safer and more respectful on line spaces, enhancing user experiences and promoting healthier digital communities.

1.4 Objectives

1. Develop an advanced machine learning model to detect and quantify various tox icity parameters in online comments.
2. Create a user-friendly interface that visually represents toxicity levels through graphical charts.
3. Implement a public chat space where users can create rooms and engage in dis cussions on diverse topics.
4. Integrate real-time monitoring and moderation to identify and blur comments ex ceeding a predefined toxicity threshold.
5. Ensure the system is adaptable and scalable for integration with various online platforms.
6. Promote a safer and more respectful online community by enhancing user aware ness and managing toxic behavior effectively.

1.5 Challenges

The challenges involved in TOXITRACK AI include accurately detecting and quantifying nuanced toxic behavior in real-time, managing the complexity and volume of online com ments, and ensuring the model can adapt to diverse linguistic and contextual variations across different platforms.

1.6 Assumptions

The machine learning model can be trained on a diverse dataset representative of various types of toxic comments.Users will engage in the chat space responsibly, and the system

will effectively blur toxic comments without significantly disrupting conversations. The platform will have sufficient computational resources to perform real-time analysis and moderation.

1.7 Societal / Industrial Relevance

TOXITRACK AI has significant societal and industrial relevance. In society, it can be applied to social media platforms, online forums, and community chat rooms to foster safer and more respectful interactions. By reducing the prevalence of toxic comments, TOXITRACK AI helps create a healthier digital environment, which can improve mental well-being and encourage positive engagement. In the industry, it can be integrated into customer service platforms, e-commerce websites, and any online communication tools to enhance user experience, ensure compliance with community guidelines, and maintain brand reputation by preventing harmful interactions.

1.8 Organization of the Report

The report for TOXITRACK ai contains all-encompassing details of the project divided acrossseveral chapters. Chapter 2 delves into the Software Requirements Specifications, System Features andRequirements of the system. Chapter 3 details the system Archi tecture, Database User interface design, Datasetsused, Module Division and more. The report makes use of tables, images and diagrams to give a visual representation ofthe project.

Chapter 2

Software Requirements Specification

2.1 Introduction

Purpose

The purpose of the TOXITRACK AI project is to develop a software application that analyzes user comments for various toxicity parameters, such as identity hate, insults, obscene language, severe toxicity, and threats. The goal is to provide real-time feedback and graphical representations of these toxicity levels to help foster a healthier online com munication environment. Additionally, the project includes a chat space where users can create rooms and engage in discussions, with automated monitoring and moderation to ensure conversations remain respectful and non-toxic.

Product Scope

TOXITRACK AI is a comprehensive tool aimed at identifying and visualizing the toxi city levels in user comments based on parameters such as identity hate, insult, obscene language, severe toxicity, and threats. The primary purpose of this software

is to foster a healthier online communication environment by providing real-time toxicity analysis and moderation features. The key benefits and objectives of TOXITRACK AI include:

Real-time Toxicity Analysis: Immediate evaluation of comments for various toxicity parameters to provide instant feedback and moderation.

Graphical Representation: Visualization of toxicity levels through bar graphs, allowing users to easily interpret the severity of comments.

Chat Room Monitoring: Users can create chat rooms where comments are continuously monitored. If any comment exceeds a 50User Engagement: Encourages respectful and non-toxic interactions among users by providing clear feedback on the nature of their comments.

The software aligns with corporate goals of promoting safe and inclusive online spaces. It supports business strategies focused on enhancing user experience and maintaining community standards

2.2 Overall Description

Product Prespective

TOXITRACK AI is a new, self-contained software application designed to enhance online communication by monitoring and analyzing the toxicity levels of user comments. It is not a follow-on member of an existing product family nor a replacement for any current systems. The software functions as an independent tool with capabilities to interface with various chat platforms, providing real-time toxicity analysis and moderation.

Product Functions

Real-time Toxicity Analysis: Evaluate user comments for various toxicity parameters such as identity hate, insults, obscene language, severe toxicity, and threats.Graphical Representation: Display the percentage of toxicity parameters in bar graphs for easy in terpretation. Chat Room Monitoring: Allow users to create chat rooms, where comments are analyzed in real-time. If a comment exceeds 50User Notifications: Inform users about the toxicity level of their comments to promote

respectful interactions.

Operating Enviornment

Hardware: Standard server hardware with sufficient processing power and memory to handle real-time analysis. Operating System: Compatible with major operating systems such as Windows, macOS, and Linux. Software: Built using Flask for the backend, lever aging the transformers library for text classification, and Firebase for authentication and database services.

Design and Implementation constraints

Corporate/Regulatory Policies: Compliance with data privacy laws such as GDPR. Hard ware Limitations: Real-time processing requirements necessitate efficient use of resources. Interface Requirements: Seamless integration with external chat applications via APIs.

Technology Stack: Flask, transformers library, Firebase. Security Considerations: Secure user data handling and storage, robust authentication mechanisms.

Assumptions and Dependencies

Third-party Components: Utilization of the transformers library for toxicity analysis, and Firebase for authentication and database services. Development Environment: As sumes availability of necessary development tools and environments. External Factors: Dependence on third-party APIs for integration with chat platforms, which may affect functionality if they change.

2.3 External Interface Requirements

User Interfaces

The user interface (UI) of TOXITRACK AI will be designed to ensure an intuitive

and user-friendly experience. The logical characteristics of each interface between the software product and the users are as follows: GUI Standards: The design will follow modern web design standards, ensuring a consistent look and feel across the application. Screen Layout Constraints: The UI will be responsive, adapting to various screen sizes from desktops to mobile devices. Layouts will be designed to provide a seamless experience regardless of the device used. Standard Buttons and Functions: Common elements will include buttons for login, signup, comment submission, and navigation between different chat rooms. Each screen will have consistent placement for these buttons to maintain familiarity. Keyboard Shortcuts: Accessibility features will include keyboard shortcuts, such as pressing ”Enter” to send a message, to improve user experience. Error Message Display Standards: Error messages will be clearly displayed near the relevant input fields or at the top of the screen to inform users of issues such as invalid input or failed actions.

Hardware Interfaces

The hardware interfaces describe the interaction between the software and the hard ware components of the system: Supported Device Types: The system will be compatible with standard server hardware and client devices, including desktops, laptops, tablets, and smartphones. Data and Control Interactions: The server will process text inputs from clients and provide real-time updates to them. This requires reliable network connec tivity and adequate processing power to handle simultaneous requests. Communication Protocols: Standard internet communication protocols (HTTP/HTTPS) will be used to facilitate data exchange between client devices and the server.

Software Interfaces

The software interfaces detail the connections between TOXITRACK AI and other software components: Connections with Specific Software Components: Databases: The system will use Firebase Realtime Database to store chat data, toxicity values,

and chat room topics. Operating Systems: The backend server will run on an operating system compatible with Flask (e.g., Linux, Windows). Tools and Libraries: Key libraries and tools include the transformers library for toxicity analysis and Firebase SDK for authenti cation and database operations. Data Items and Messages: The system will handle input data such as user comments and output data including toxicity analysis results. Data interchange will use JSON format for consistency and ease of parsing. Service Require ments: The system needs to support real-time analysis, data storage, user authentication, and session management.

Communication Interfaces

The communications interfaces describe the requirements for the communication func tions of the product: Protocols: The system will use HTTP/HTTPS protocols for secure data transmission between client and server. Message Formatting: JSON format will be used for data exchange to ensure compatibility and ease of use. Communication Stan dards: The system will adhere to standard web communication protocols, such as RESTful API design, to facilitate interaction with external systems.

2.4 System Features

Toxicity Analysis

Description and Priority This feature analyzes user comments for various toxicity parameters such as identity hate, insult, obscene content, severe toxicity, and threats. It is a high-priority feature as it forms the core functionality of the TOXITRACK AI system.

Stimulus/Response Sequences Stimulus: User submits a comment in the chat. Re sponse: The system analyzes the comment for toxicity and returns the percentage levels of each toxicity parameter. If any parameter exceeds 50Functional Requirements REQ-1: The system shall tokenize and encode the input comment. REQ-2: The system shall use the pre-trained BERT model to analyze the comment.

REQ-3: The system shall calculate and return the probability for each toxicity parameter. REQ-4: The system shall mark the overall toxicity of the comment as true if any param eter exceeds 50percentage.

REQ-5: The system shall return a JSON object with the analysis

results. User Authentication

Description and Priority This feature provides secure user authentication using Fire base for login and signup processes. It is of high priority to ensure that only authorized users can access the chatrooms.

Stimulus/Response Sequences Stimulus: User submits login or signup form. Response: The system validates the credentials using Firebase Authentication and creates a user ses sion. Functional Requirements

REQ-1: The system shall provide a login form for existing users.

REQ-2: The system shall provide a signup form for new users.

REQ-3: The system shall authenticate users using Firebase Authentication. REQ-4: The system shall create a session for authenticated users.

REQ-5: The system shall redirect authenticated users to the chat

interface. 8

Chat Interface

Description and Priority This feature allows users to create chat rooms, enter com ments, and view the toxicity analysis of comments in real-time. It is of medium priority as it facilitates the primary user interaction with the system.

Stimulus/Response Sequences Stimulus: User creates a chat room and enters com ments. Response: The system stores the chat data, analyzes comments for toxicity, and updates the display with the results. Functional Requirements

REQ-1: The system shall allow users to create chat rooms.

REQ-2: The system shall display a chat interface for users to enter

comments. REQ-3: The system shall store chat data in Firebase Realtime Database. REQ-4: The system shall analyze comments for toxicity in real-time. REQ-5: The system shall blur comments marked as toxic. REQ-6: The system shall update the chat interface with toxicity analysis results. Data Storage

Description and Priority This feature stores chat data, overall toxicity values, and chat room topics in Firebase Realtime Database. It is of high priority to ensure data persistence and retrieval.

Stimulus/Response Sequences Stimulus: User interacts with the chat interface. Re sponse: The system stores and retrieves chat data and toxicity analysis results from the database. Functional Requirements REQ-1: The system shall store chat data in Firebase Realtime Database. REQ-2: The system shall store overall toxicity values for each chat. REQ-3: The system shall store chat room topics. REQ-4: The system shall retrieve stored data for display in the chat interface. Graphical Representation

Description and Priority: This feature provides a graphical representation of toxicity analysis results. It is of medium priority to help users visualize the toxicity levels of

comments. Stimulus/Response Sequences Stimulus: User submits a comment for analy sis. Response: The system generates a bar graph displaying the toxicity levels for each parameter. Functional Requirements REQ-1: The system shall generate a bar graph for toxicity analysis results. REQ-2: The system shall display the bar graph in the chat interface. REQ-3: The system shall update the graph in real-time as new comments are analyzed.

2.5 Other Nonfunctional Requirements

Perfomance Requirements

Response Time: The system must analyze a comment and return toxicity results within 2 seconds to ensure real-time interaction. Throughput: The system should han dle up to 100 concurrent users without performance degradation.Scalability: The system should be able to scale horizontally to accommodate an increasing number of users and chat rooms.Resource Utilization: CPU and memory usage should be optimized to prevent server overload, especially during peak times.

Safety Requirements

Data Integrity: Ensure the integrity of chat data and toxicity analysis results by im plementing robust data validation and error-handling mechanisms.Error Handling: The system must handle errors gracefully, providing clear messages to users without expos ing technical details.Backup and Recovery: Regular backups of the database should be scheduled, and recovery procedures must be in place to prevent data loss in case of system failure. User Protection: Implement measures to protect users from exposure to harmful or offensive content by automatically blurring comments identified as toxic.

Security Requirements

Authentication: Use Firebase Authentication to ensure secure login and signup

cesses, protecting user credentials. Authorization: Implement role-based access control to restrict access to certain functionalities based on user roles (e.g., admin, user).Data Encryption: Encrypt sensitive data in transit and at rest to prevent

unauthorized ac cess.Privacy: Ensure user data privacy by complying with relevant data protection regu lations (e.g., GDPR, CCPA).Logging and Monitoring: Implement logging and monitoring to detect and respond to security incidents promptly.

Software Quality Attributes

Adaptability: The system should be adaptable to changes in toxicity detection mod els or additional toxicity parameters. Availability: The system should maintain a min imum uptime of 99.9 percentage,ensuring high availability for users.Correctness: The system must accurately analyze and report toxicity levels without false positives or neg atives.Flexibility: The system should support integration with other services and allow easy modifications to the frontend and backend. Robustness: The system should be ro bust, handling unexpected inputs or situations without crashing.Testability: The system should be designed to allow easy testing of individual components and overall function ality.Usability: The system should be user-friendly, with an intuitive interface and clear instructions, to ensure a positive user experience.

Chapter 3

System Architecture and Design

3.1 System Overview

TOXITRACK AI is designed to analyze user comments for various toxicity parameters and provide a graphical representation of the analysis. It also includes a chat functionality where users can create rooms and discuss different topics. The system ensures that if any comment in a chat room exceeds a defined toxicity threshold, the overall toxicity of the chat is flagged, and the chat is blurred to protect users from harmful content. The architecture is divided into frontend, backend, and database components, integrating several technologies to ensure seamless functionality.

Detailed Architecture Description Frontend: The frontend of TOXITRACK AI is built using HTML, CSS, and JavaScript. It serves the following purposes: User Interface: Provides a user-friendly interface for users to interact with the sys tem, including login, signup, creating chat rooms, and participating in chats. Graphical Representation: Displays the toxicity analysis results using bar graphs, illustrating the percentage of different toxicity parameters. Real-time Updates: Uses JavaScript to enable real-time updates in chat rooms, ensuring that users can see new messages and toxicity warnings as they happen. Backend The backend is developed using Flask, a lightweight Python web framework, to handle the server-side logic. Key functionalities include: API Endpoints: Provides endpoints for user authentication, comment submission, and toxicity analysis. Toxicity Analysis: Integrates with a pre-trained BERT model (unitary/toxic-bert) to analyze comments. The analysis results include percentages for various toxicity parameters such as identity hate, insult, obscene, severe toxic, and threat. Session Management: Manages user sessions to ensure secure access to the chat rooms and other functionalities. Database Firebase is used for both authentication and real-time database functionalities:

Figure 3.1: System Architectural Diagram

Authentication: Firebase Authentication handles user login and signup processes se curely. Realtime Database: Firebase Realtime Database stores chat data, including the chat messages, overall toxicity values, and the topics of chat rooms. It ensures real time synchronization, so users see updates immediately. Detailed Process Outline User Authentication:

Users sign up or log in through the frontend. The frontend communicates with Firebase Authentication to verify user credentials. Upon successful authentication, a session is created for the user, and they are redirected to the main interface. Creating and Joining Chat Rooms:

Authenticated users can create new chat rooms or join existing ones. Each chat room has a unique topic stored in the Firebase Realtime Database. The frontend retrieves and displays the list of chat rooms for user selection. Comment Submission and Analysis:

Users submit comments through the chat interface. The frontend sends the comment to the Flask backend via an API call. The backend uses the BERT model to analyze the comment for toxicity. The analysis results, including percentages for each toxicity parameter, are returned to the frontend. Displaying Toxicity Results:

The frontend displays the toxicity analysis results in a bar graph format. If any parameter exceeds 50The system blurs the chat if the overall toxicity is flagged, ensuring that harmful content is obscured. Storing Chat Data:

All chat messages, along with their toxicity analysis results, are stored in the Firebase Realtime Database. The database maintains the state of each chat room, including the topic and overall toxicity status. Real-time Synchronization:

The Firebase Realtime Database ensures that any new messages or changes in toxicity status are immediately reflected across all users in the chat room. Users receive real time updates, making the chat experience seamless and dynamic. Key Components Flask Backend: Handles API requests, integrates with the BERT model, and manages session logic. BERT Model: Provides the core functionality for toxicity analysis. Firebase Au thentication: Secures user login and signup processes.

Firebase Realtime Database: Stores chat data and synchronizes updates in real time. Frontend Interface: Built with HTML, CSS, and JavaScript to provide an interactive and responsive user experience. This ar chitecture ensures that TOXITRACK AI is scalable, secure, and capable of providing real-time toxicity analysis and feedback to users, fostering a safer and more respectful online communication environment.

3.2 Architectural Design

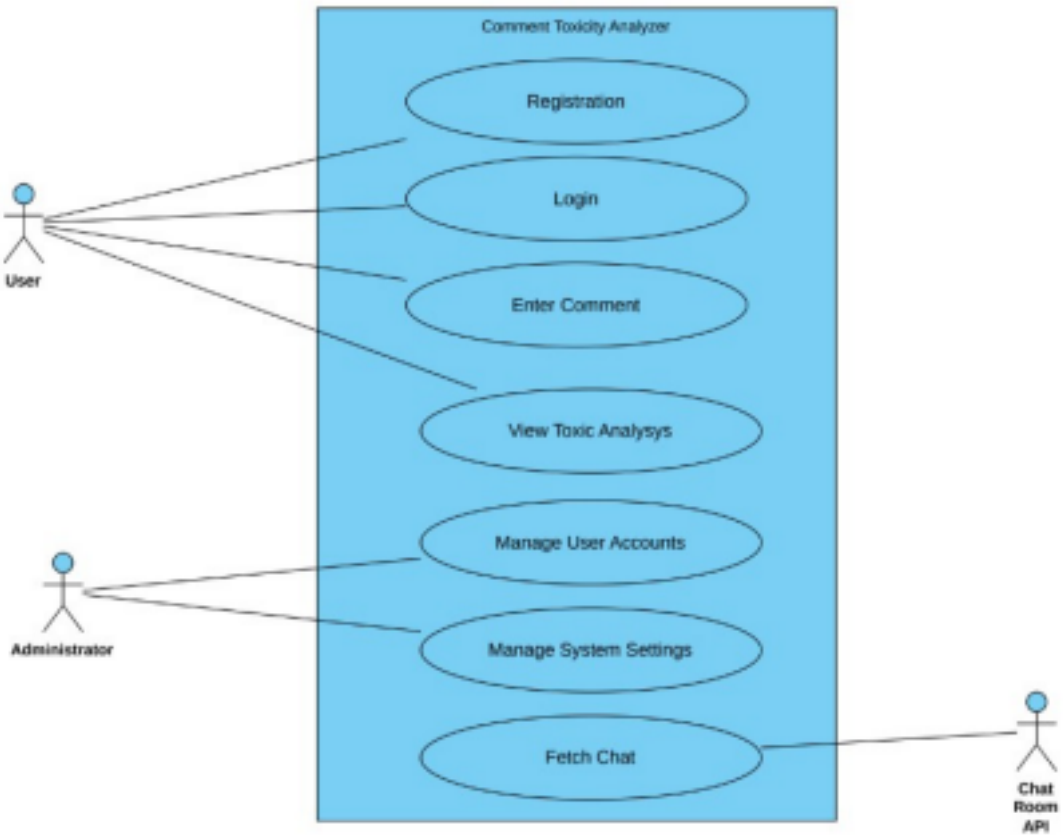

Figure 3.2: Usecase diagram

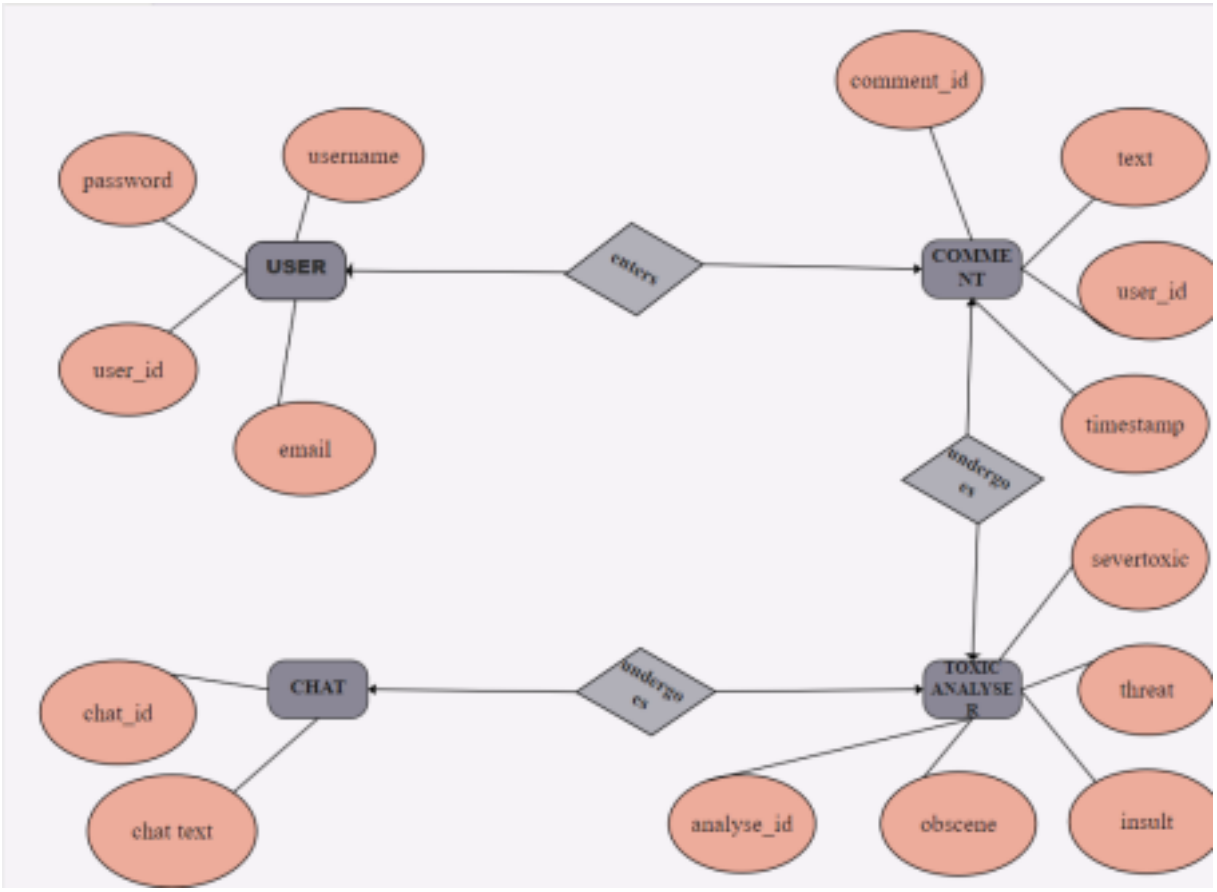

Figure 3.3: ER Diagram

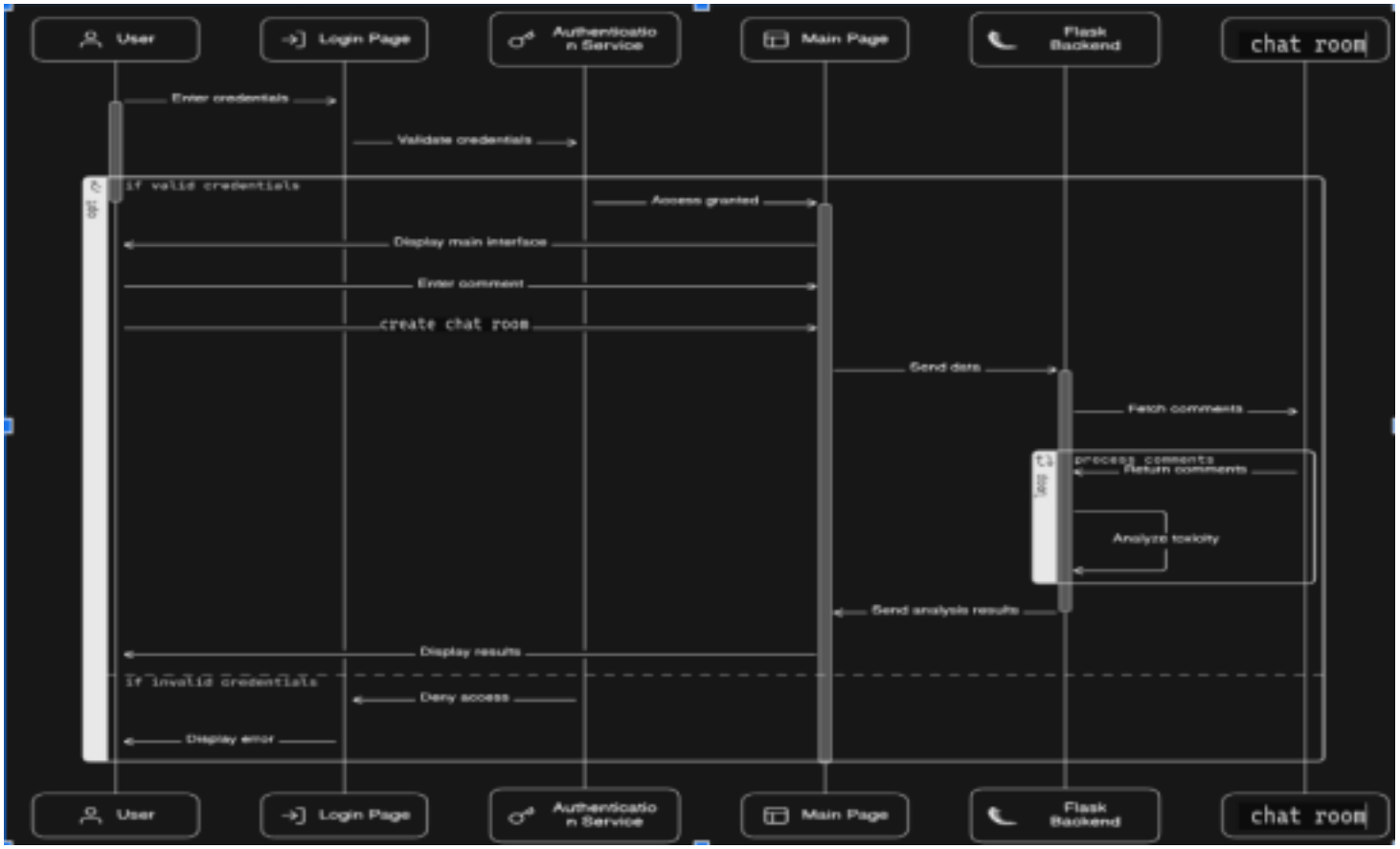

Figure 3.4: Sequence Diagram

3.3 Dataset Identified

Data Source The dataset selected for this project is sourced from Kaggle, a renowned platform for datasets and machine learning resources. The dataset specifically focuses on toxic comments classification and includes a diverse range of comments labeled with toxicity parameters such as identity hate, insults, obscenities, severe toxicity, threats, and general toxicity.

Dataset Properties The dataset exhibits the following properties:

- Size: The dataset consists of a substantial number of comments, providing a rich and diverse collection for training and testing the toxicity detection model.
- Labeling: Each comment in the dataset is labeled with one or more toxicity pa rameters, allowing for multi-label classification tasks.
- Variety: The dataset encompasses a wide range of comment types, reflecting real world scenarios and diverse language usage.
- Balanced Distribution: The distribution of toxic and non-toxic comments is balanced, ensuring unbiased model training and evaluation.
- Quality: The dataset is curated and annotated with care, ensuring accuracy and reliability in the labeling of toxicity parameters.

Dataset Location The dataset can be accessed and downloaded from Kaggle at the following location:

https://www.kaggle.com/datasets

Sample Subsets To provide an overview of the dataset's content, sample subsets can be highlighted. For example:

- Sample 1: Not sure about a heading of 'Fight for Freedom' what will it contain?
- Sample 2: Are you threatening me for disputing neutrality? I know in your country it's quite common to bully your way through a discussion and push outcomes you want. But this is not Russia.
- Sample 3: I went there around the same time he did, and that certainly was not the case at the time. Later on they stopped taking children from such a young age.

3.4 Proposed Methodology

3.4.1 Text Toxicity Analysis

We will use a pre-trained transformer-based model for text classification, specifically the transformers library in Python. This library provides easy access to a variety of

pre trained models such as BERT, RoBERTa, and DistilBERT, which are known for their performance in natural language processing tasks.

Input Preprocessing: Comments entered by users will undergo basic text prepro cessing steps such as tokenization, lowercasing, and removal of stop words to prepare them for analysis.

Toxicity Parameter Classification: We will use the pre-trained model to classify each comment into different toxicity parameters, including identity hate, insult, obscene, severe toxic, and threat. Each parameter will be assigned a binary label (0 for not present, 1 for present) based on the comment's content.

Threshold Determination: To calculate the percentage of each toxicity parameter, we will set a threshold value (e.g., 50

Percentage Calculation: The percentage of each toxicity parameter will be calcu lated based on the number of comments in a chat room that contain the parameter above the threshold, divided by the total number of comments in the room.

Overall Toxicity Determination: If any toxicity parameter exceeds the threshold in a chat room, the overall toxicity of the chat will be marked as true.

3.4.2 Graphical Representation

For the graphical representation of toxicity parameters, we will use a bar graph. The x-axis of the graph will represent different toxicity parameters, while the y-axis will represent their corresponding values (percentage). Each bar's height will represent the percentage of a specific toxicity parameter in the chat room, providing a visual representation of the toxicity distribution.

3.4.3 Chat Room Functionality

The chat room functionality will be implemented using Flask for the backend and HTML, CSS, and JavaScript for the frontend. Firebase Realtime Database will store chat data,

including comments and toxicity parameter classifications. Firebase authentication will be used for user login and signup.

3.5 User Interface Design

Figure 3.5: page 1

Figure 3.6: page 2

Figure 3.7: page 3

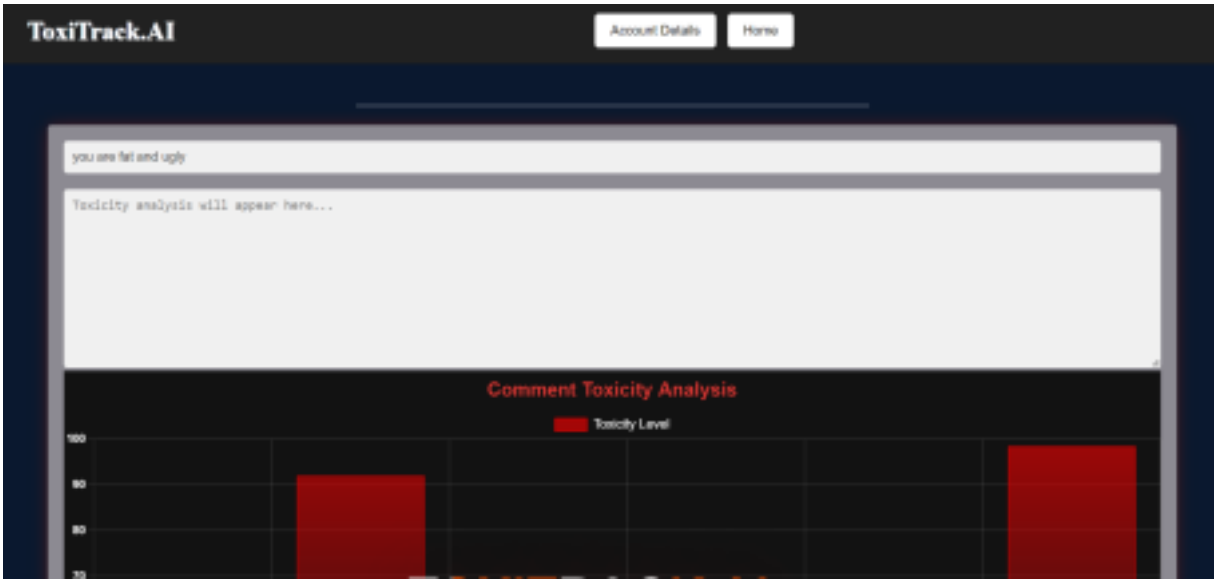

Figure 3.8: page 4

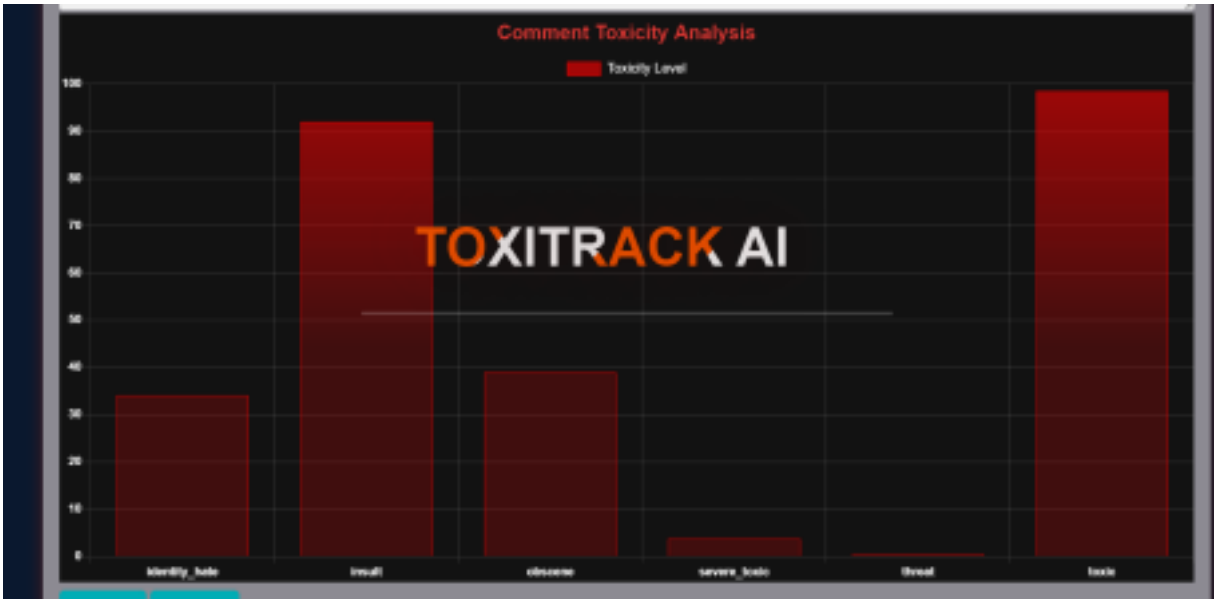

Figure 3.9: page 5

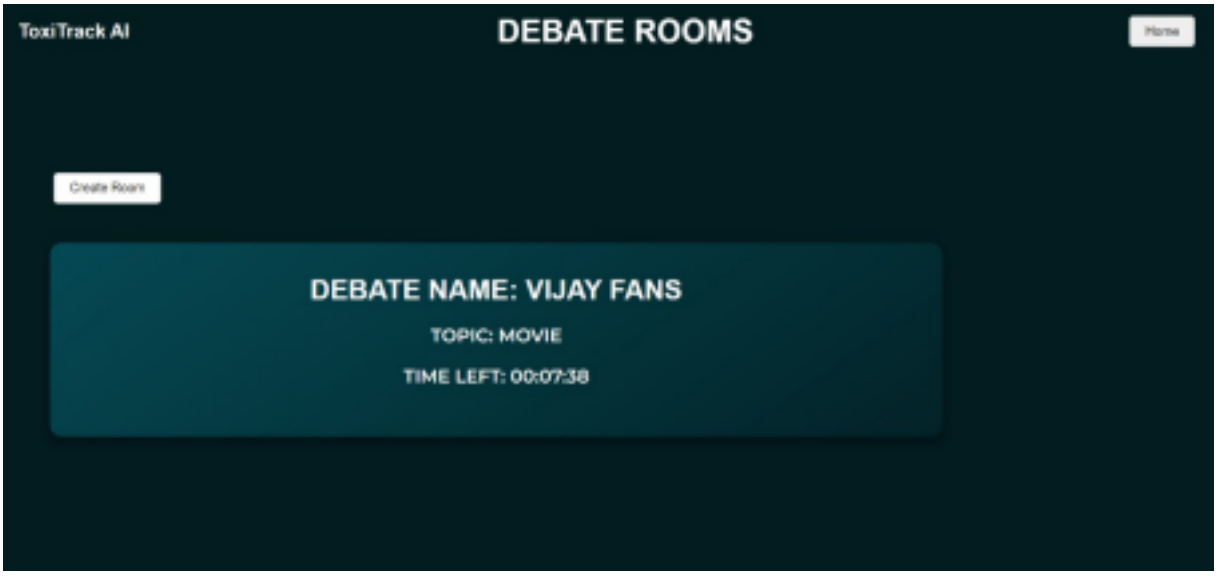

Figure 3.10: page 6

3.6 Database Design

Database Selection For the TOXITRACK AI project, we have chosen to use Firebase Realtime Database as our database solution. Firebase offers real-time data

synchro nization and is well-suited for applications that require real-time updates, such as chat applications. The decision to use Firebase was based on several factors:

Real-time Updates: Firebase Realtime Database provides real-time synchronization, allowing us to update and synchronize chat data and toxicity parameters in real time, providing a seamless user experience.

Scalability: Firebase is highly scalable, which is crucial for our project as we anticipate a large number of users and messages. Firebase can handle the scalability requirements of our application without compromising performance.

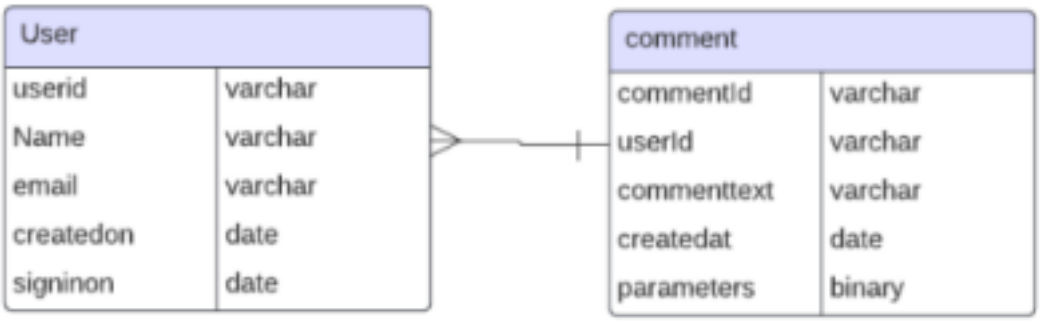

Figure 3.11: Database Design Diagram

3.7 Description of Implementation Strategies

3.7.1 Text Analysis for Toxicity Parameters

- Use the transformers library to preprocess comments and extract features for toxicity parameters like identity hate, insult, obscene, severe toxic, and threat.
- Fine-tune a pre-trained transformer model (e.g., BERT, RoBERTa) for text classi fication to predict these toxicity parameters.
- Threshold the predicted probabilities to determine if a comment is toxic in a specific category.

3.7.2 Graphical Representation

- Use a JavaScript library like Chart.js or D3.js to create a graphical

representation of the percentage of each toxicity parameter.

- Update the graph dynamically as new comments are analyzed.

3.7.3 Chatroom Implementation

- Use Flask for the backend to handle chatroom functionalities.
- Use Firebase Realtime Database to store chat messages and toxicity parameters for each message.
- Use Firebase Authentication for user login and signup.

3.7.4 Real-time Toxicity Detection

- Implement a WebSocket connection between the frontend and backend to enable real-time communication.
- When a new message is sent in the chatroom, send it to the backend for toxicity analysis.
- If any toxicity parameter exceeds 50%, mark the overall toxicity of the chat as true and blur the message.

3.7.5 Deployment

- Deploy the backend Flask application on a cloud platform like Heroku or AWS.
- Host the frontend on a static hosting service or integrate it with the backend de ployment.

3.8 Module Division

1. Comment Analysis Module

   Description: This module analyzes comments for toxicity parameters like identity hate, insult, obscene, severe toxic, and threat. It calculates the percentage of each

   parameter in the comment.

Assigned to: Basil Eldo

2. Graphical Representation Module

Description: This module is responsible for displaying the percentage of toxicity parameters in a graphical representation, making it easier for users to understand the toxicity levels.

Assigned to: Christo Shaju

3. Chat Room Module

Description: This module handles the creation of chat rooms, user interactions, and comment submission. It also integrates the comment analysis module to analyze comments in real-time.

Assigned to: Basil Sabu

4. Toxicity Threshold Module

Description: This module checks the percentage of toxicity parameters in com ments. If any parameter exceeds 50%, it marks the overall toxicity of the chat as true and triggers the blurring of the chat.

Assigned to: Avin pullikan

5. User Interface Module

Description: This module handles the frontend of the application, including the design and layout of the graphical representation of toxicity parameters and the chat room interface.

Assigned to: Christo Shaju

6. Backend Integration Module

Description: This module integrates the frontend with the backend, ensuring smooth communication between the different modules and proper functioning of the application.

Assigned to: Basil Sabu

7. Database Management Module

Description: This module manages the storage of chat data, toxicity parameters,

and other relevant information in the database. It ensures data integrity and secu rity.

Assigned to: Basil Eldo

3.9 Work Schedule - Gantt Chart

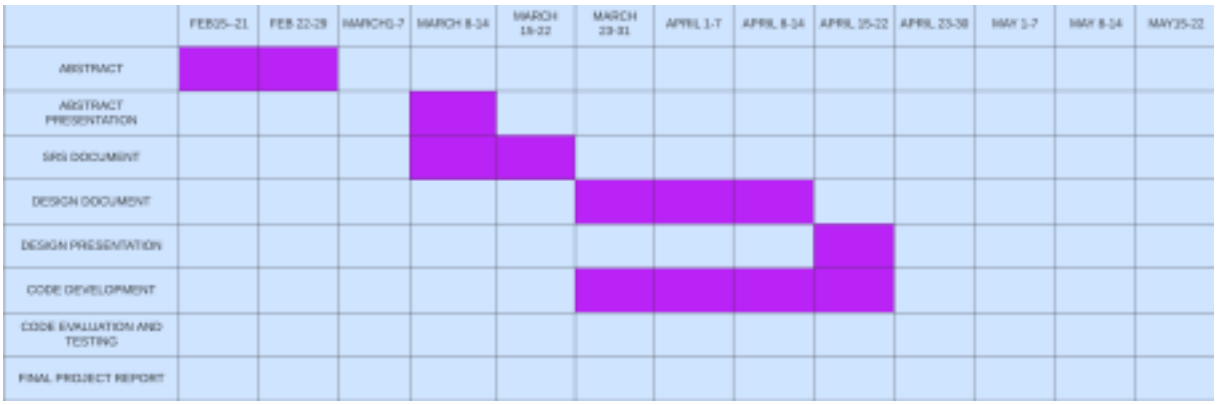

Figure 3.12: Gantt Chart

Chapter 4

Results and Discussions

4.1 Overview

Finally we created a software which analyzes a comment enter by the user and rep resent it in graphical format and as an application of the model we created a chatroom section.Users can create multiple chat rooms and chat and debate on various topics the software ensure healthy and toxic free enviornment because any such interaction would be filtered by blurring it.

4.2 Testing

Figure 4.1: Result 1

Figure 4.2: Result 2

Figure 4.3: Result 3

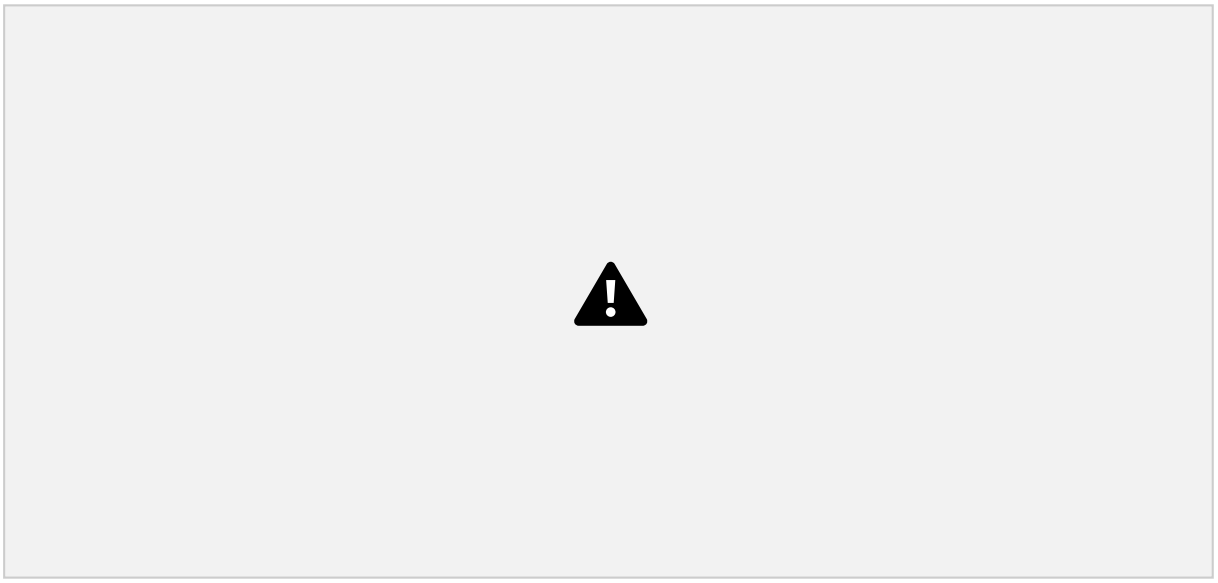

Figure 4.4: Result 4

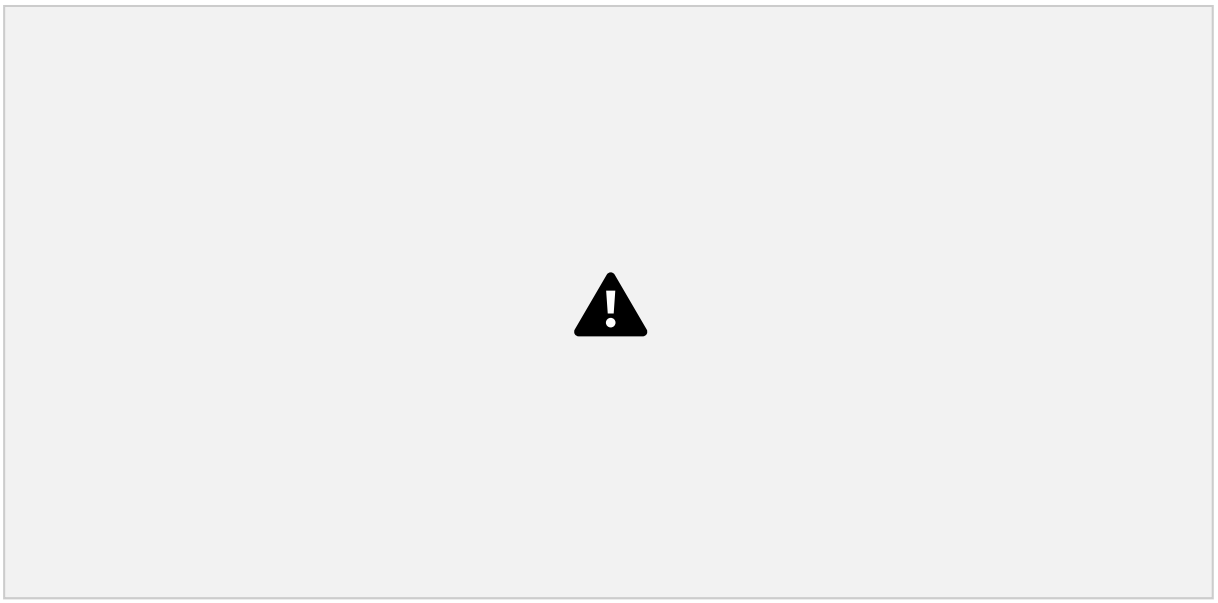

Figure 4.5: Result 5

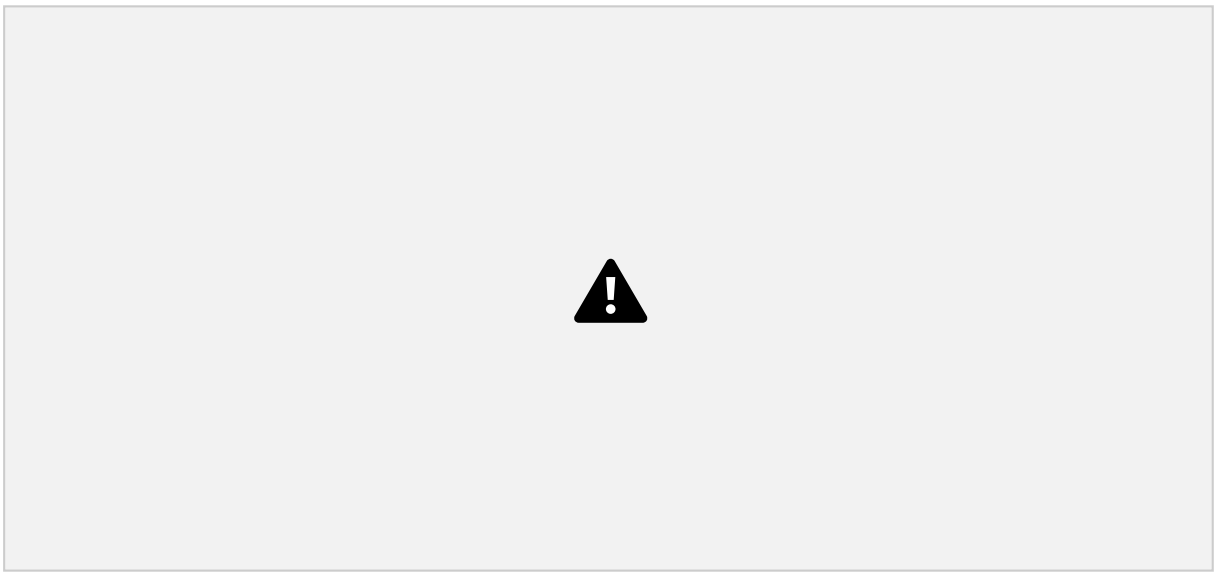

Figure 4.6: Result 6

4.3 Quantitative Results

In this project, we evaluated our model using a confusion matrix, and derived key

perfor mance metrics to assess its effectiveness. The confusion matrix was as follows: This matrix indicates:

• 25 True Negatives (non-toxic comments correctly identified as non-toxic) • 5 False Positives (non-toxic comments incorrectly identified as toxic) • 3 False Negatives (toxic comments incorrectly identified as non-toxic) • 27 True Positives (toxic comments correctly identified as toxic)

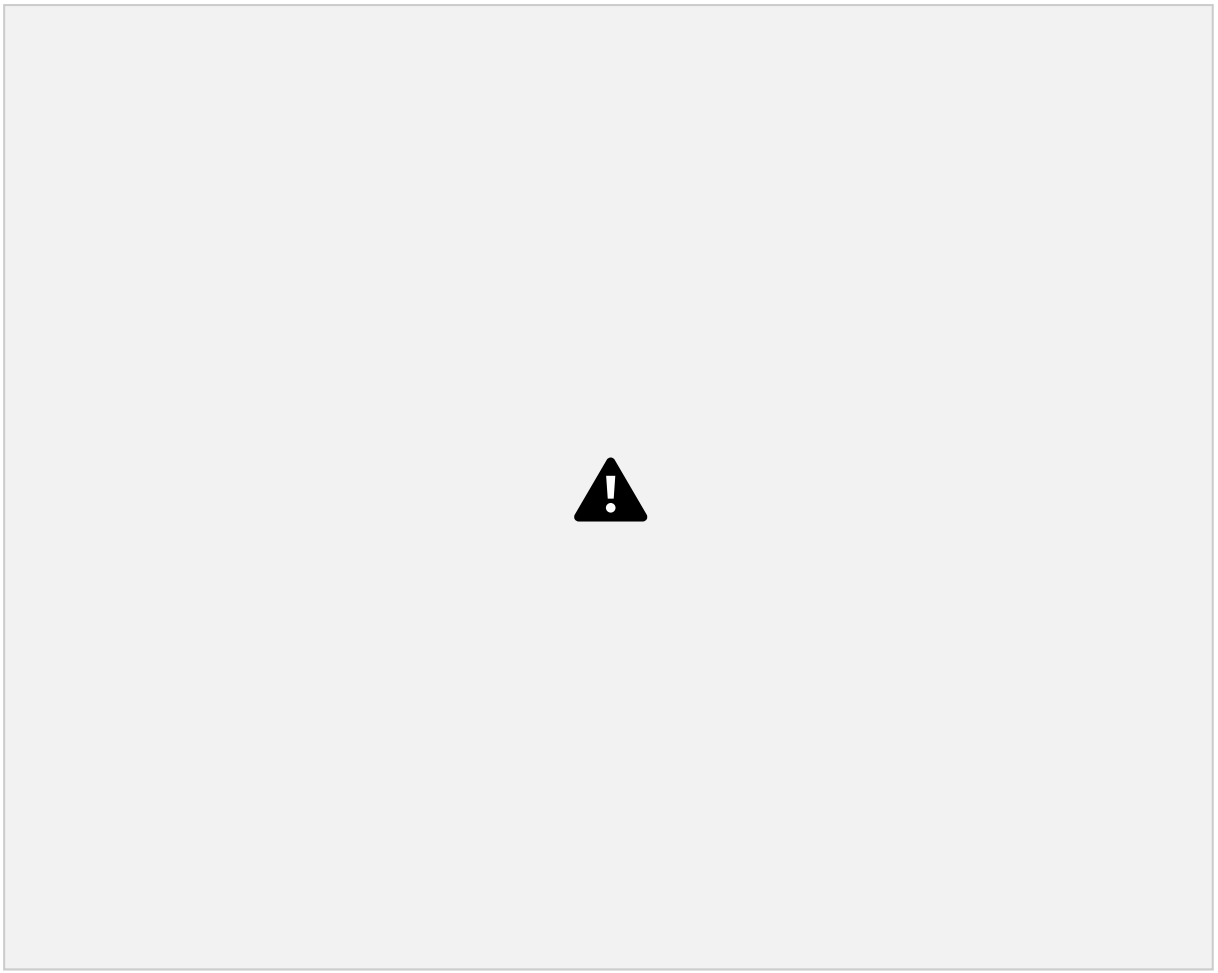

Figure 4.7: Confusion matrix

Figure 4.8: Analysis Accuracy

4.4 Discussion

TOXITRACK AI has been implemented as a web app where users can log in, enter

comments, and analyze the entered comments. Different toxicity parameters are displayed graphically. A chatroom section has been successfully implemented, creating private and public rooms where different communities can discuss their topics. Classes can conduct open scrutiny meetings, and other communities in sports, films, etc., can debate over topics. The app ensures a healthy environment by filtering toxic comments.

Chapter 5

Conclusion

5.1 Conclusion

TOXITRACK AI represents a significant leap forward in fostering a healthier and more engaging online environment. By harnessing the power of comment toxicity analysis and real-time filtering, our software ensures that every interaction within our platform is constructive and free from toxicity.

With the ability to analyze comments and display toxicity parameters graphically, users gain valuable insights into the quality of their interactions. The implementation of a chatroom section further enriches the user experience, allowing diverse communities to engage in meaningful discussions while maintaining a safe and respectful atmosphere.

Whether it’s classrooms conducting open scrutiny meetings, communities discussing sports, films, or any topic under the sun, TOXITRACK AI provides a platform where ideas can flourish without the hindrance of toxic behavior. By blurring toxic comments, we not only protect our users but also encourage responsible online communication.

In conclusion, TOXITRACK AI stands as a beacon for building digital communities that thrive on positivity, collaboration, and mutual respect, setting a new standard for healthy online interactions.

5.2 Future Scope

TOXITRACK AI indeed sounds like a transformative tool for promoting healthy online interactions. Here are some potential future enhancements you could consider for the project:

Enhanced Machine Learning Models: Continuously improving the toxicity detection models by incorporating advanced machine learning techniques, such as deep learning ar chitectures or transformer models like BERT, could enhance the

accuracy and granularity

of toxicity analysis.

Multilingual Support: Expanding language support beyond English to include other languages commonly used online would broaden the reach of TOXITRACK AI and make it more accessible to diverse communities globally.

Customizable Filtering and Moderation: Offering users the ability to customize their toxicity filtering thresholds and moderation settings can empower them to tailor their experience based on their preferences and comfort levels.

User Reputation and Trust Scores: Implementing a system that tracks user behavior and assigns reputation or trust scores based on their interactions can help in identifying and addressing potential sources of toxicity more effectively.

Bibliography

[1] Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P.: Convolutional neural networks for toxic comment classification. In: 10th Hellenic Conference on Artificial Intelligence (2018)

[2] hu, T., Jue, K., Wang, M.: “Comment abuse classification with deep learning. https://web.stanford.edu/class/cs224n/reports/2762092.pdf

[3] Ronan Collobert, Jason Weston, L´eon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res. 12 (Nov. 2011), 2493–253

Appendix A: Presentation 36

TOXITRACKER AI

GUIDED BY:
Ms. SHERINE SEBASTIAN

TEAM MEMBERS:
Basil Sabu
Christo Shaju
Basil Eldo
Avin Pulikkan

CONTENTS

- Problem Definition
- Objectives
- System Design
- Datasets
- Work Division – Gantt Chart
- Software/Hardware Requirements
- Conclusion
- References

INTRODUCTION

- A web-based tool for analyzing comment toxicity and show its graphical analysis
- It is an AI model assessing sentences for **toxicity, severe toxicity, threats, insults, and identity hate.**
- Create a chat space for users to engage in various topics
- Allows different communities to perform various debates in a healthy enviornment

PROBLEM DEFINITION

- Online platforms often face the problem of toxic comments, including profanity, insults, threats, and other harmful language. This toxic behavior creates a hostile environment and discourages people from participating in discussions. There is a need for a web application that can analyze comments for toxicity and provide a safe, moderated chat space for users to engage in respectful and constructive conversations

OBJECTIVES

● Develop a robust toxicity detection Model ● Deploy a system capable of analyzing comments in real-time

● User-Friendly chat and debate space

SYSTEM DESIGN

SYSTEM OVERVIEW

- Architecture Overview
- Process Outline
- Technology Stack
- Security Measures

Architectural Design

4/15/2024

MODULES

Model Module

- Data processing
- Training and Evaluation

MODULE (CONTINUATION)

Authentication Module

- Facilitates user registration
- Secure login for registered users

Chat Space Module

- Interaction with the chat room API
- Functionalities for retrieving chats

MODULE (CONTINUATION)

Website Interface Module

- Provides a user-friendly web interface for comment entry

- provides interface to create room and involve in chats

Flask Modules

- Defines routes for comment analysis, and database interactions
- Middleware functions handle request and error handling

Sequence Diagram

Use Case Diagram

4/15/2024

DATASETS

Source: https://www.kaggle.com/c/jigsaw-toxic-comment classification-challenge/data

WORK DIVISION

Gantt Chart

SOFTWARE/HARDWARE REQUIREMENT

SOFTWARE:

- Python will be used as programming language for development
- BERT model is used
- Flask Python web framework for backend

development

CONTINUATION

- Front-end technologies like HTML, CSS, JavaScript for creating the user interface
- Firebase is used as database for user authentication

HARDWARE

- Operating System: Windows, macOS, Linux
- Processor: Intel Core i5 or equivalent
- ram 8gb
- Internet Connection: Required for accessing ChatroomAPI, Firebase database, hosting the website, and deploying Flask backend
- Graphics: Integrated or discrete graphics card capable of accelerating deep learning computations (NVIDIA GPU recommended for faster training)

RESULTS

RESULT

RESULT

CONCLUSION

Product Description:

- A web app that successfully analyses the comment entered by the user and a chat space for various communities have been implemented

Product Features:

- Analysing comment
- Chat rooms

FUTURE ENHANCEMENTS

- Multilanguage Support
- expanding the features social media platform
- introducing warning system

REFERENCES

- Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G., Plagianakos, V.P.: Convolutional neural networks for toxic comment classification. In: 10th Hellenic Conference on Artificial Intelligence (2018)
- Chu, T., Jue, K., Wang, M.: “Comment abuse classification with deep

learning. https://web.stanford.edu/class/cs224n/reports/2762092.pdf

- Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. J. Mach. Learn. Res. 12 (Nov. 2011), 2493--2537

Thank You!

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING RAJAGIRI SCHOOL OF ENGINEERING & TECHNOLOGY (AUTONOMOUS) RAJAGIRI VALLEY, KAKKANAD, KOCHI, 682039

(Affiliated to APJ Abdul Kalam Technological University)

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision

To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission

To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding pro fessionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)

Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering prob lems and design system components or processes that meet the specified needs with ap propriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex en gineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsi bilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engi neering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsi bilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understand ing of engineering and management principles and apply these to one's own work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and abil ity to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills

The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills

The ability to acquire programming efficiency by designing algorithms and applying stan dard practices in software project development to deliver quality software

products meet ing the demands of the industry.

PSO3: Professional Skills
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes
After the completion of the course the student will be able to:

CO1:
Identify technically and economically feasible problems (Cognitive Knowledge Level: Ap ply)

CO2:
Identify and survey the relevant literature for getting exposed to related solutions and get familiarized with software development processes (Cognitive Knowledge Level: Apply)

CO3:
Perform requirement analysis, identify design methodologies and develop adaptable & reusable solutions of minimal complexity by using modern tools & advanced program ming techniques (Cognitive Knowledge Level: Apply)

CO4:
Prepare technical report and deliver presentation (Cognitive Knowledge Level: Apply)

CO5:
Apply engineering and management principles to achieve the goal of the project (Cognitive Knowledge Level: Apply)

Appendix C: CO-PO-PSO Mapping 6