IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134174

Toxic Comment Detection and Classifier

Adarsh Vinod[1], Adithyan K V[2], Manoranjan M[3], Ramsha Riyaz[4], Mr. Arul N[5]

Student, Computer Science and Engineering, AJIET, Mangalore, India[1-4]

Assistant Professor, Computer Science and Engineering, AJIET, Mangalore, India[5]

Abstract: With the help of a machine learning (ML) model for toxic remark identification, this project presents a locally hosted social media platform that looks like Facebook or Instagram. An active online community is fostered by users' ability to create accounts, publish information, and participate in discussions. By utilizing cutting-edge machine learning algorithms, the platform can identify and eliminate harmful remarks on its own, creating a polite and secure place for users to engage. Proactive moderating is made possible via an email notification system that also instantly informs users of any offensive comments on their posts. With this study, we show how effective machine learning (ML) solutions can be at improving online safety and encouraging positive social media communication.

Keywords: Social Media, offending comment, Toxic Comment Detection, Positive social media communication

I. INTRODUCTION

Social media platforms have revolutionized the way people connect, communicate, and share experiences in the digital age. Not withstanding the advantages of interconnectivity, these platforms frequently encounter obstacles associated with the dissemination of detrimental conduct and content. Hate speech, harassment, and cyberbullying can have a big effect on community dynamics and user experience.

This project intends to create a locally hosted social media network that is modeled after well-known platforms like Facebook and Instagram in response to these difficulties. The main novelty of the platform is the incorporation of a machine learning (ML) model for toxic comment detection, which aims to promote a more secure and encouraging online community.

This platform's main goal is to enable users to participate in meaningful relationships without worrying about coming across harmful stuff. Through the use of cutting-edge machine learning algorithms, the platform analyzes user comments automatically in real-time, recognizing and flagging those that engage in harmful activity. By taking a proactive stance towards content moderation, we hope to foster an accepting and inclusive online community where people are appreciated and respected.

In addition, the platform has an email notification system to provide customers more control over their online experience. The system instantly sends the user an email if it finds a poisonous comment on their article, giving them the opportunity to take the necessary action, such removing the offending comment or blocking the offending user. This research aims to show how effective machine learning (ML)-driven solutions may be in tackling the problems associated with online toxicity by putting these cutting-edge features into practice. The platform prioritizes user safety and well-being in order to foster a respectful and civilized culture in social media spaces, which will enhance and improve everyone’s online experience.

II. PROBLEM STATEMENT

The emergence of social media platforms has completely changed how individuals interact and communicate on the internet. Notwithstanding the advantages of digital connectivity, these platforms frequently encounter noteworthy obstacles associated with the spread of detrimental conduct and information. Incidents of hate speech, harassment, and cyberbullying can negatively affect user experience, causing psychological harm and jeopardizing the platform's credibility. This project aims to address the urgent need for efficient user protection and content control in social media contexts in light of these difficulties. The project's specific goal is to create a social media network that is hosted locally and has sophisticated machine learning (ML) features for the identification and removal of harmful comments.

In addition, the project intends to provide consumers more control over their online experience by putting in place an email notification system that notifies users of any negative comments made on their content.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134174

Through the provision of timely awareness and proactive moderation activities, the platform aims to alleviate the negative effects of toxic conduct and promote a respectful and civilized culture among its user base.

In conclusion, the issue this project attempts to solve is the necessity of promoting a healthy and good online community atmosphere while also battling online toxicity. The project seeks to develop novel machine learning (ML) solutions for user safety and content moderation, with the goal of making social media safer and more pleasant for all users.

III. OBJECTIVES

1. Provide a user-friendly social media platform that is similar to well-known sites like Facebook or Instagram, enabling account creation, content uploading, and user interaction.

2. Put into practice an ML model that uses natural language processing (NLP) to examine user-generated content and pinpoint detrimental conduct in order to identify poisonous remarks in real-time.

3. As soon as comments are submitted, make sure the machine learning model can quickly scan and assess them, highlighting those that show harmful behavior for additional examination or action.

4. Include an email notification system that allows users to be notified as soon as harmful remarks are found on their postings. This gives them the ability to moderate the comments right away.

5. Conduct thorough testing to evaluate the efficacy and performance of the ML model and comment moderation system. Take into account KPIs like accuracy, efficiency, and user happiness.

IV. REQUIREMENT SPECIFICATION

Hardware Requirements

1. Processor: Intel(R) Core i3 & above Versions.
2. System Type: 64-bit operating system, x64-basedprocessor
3. Installed Ram: 8 GB
4. Network Infrastructure: High-speed and reliable network connections to ensure seamless communication between server components and responsiveness for end-users

Software Requirements

1. Platform and Hosting: Hosting for scalability, reliability, and security. Compatibility across various platforms (web, mobile, etc.).
2. Programming Languages and Frameworks: Python for AI/ML algorithms and backend development. TensorFlow or PyTorch for machine learning models.
3. Operating System: Choose a stable and secure operating system that aligns with government IT policies. Common choices include Linux distributions (e.g., CentOS, Ubuntu) or Windows Server.
4. Front-End Technologies: Implement front-end technologies (HTML, CSS, JavaScript) to create an intuitive and user- friendly interface. Consider using a front end framework.

V. SYSTEM DESIGN

A. METHODOLOGY

The Figure shows the steps that are followed in the methodology of this project.

1. **User Registration**: Users register or sign up to gain access to the system, providing necessary information such as username, email, and password.

2. **Input Data (Comment Submission):**After registration, users can submit comments on posts within the social media platform. These comments serve as input data for further processing.

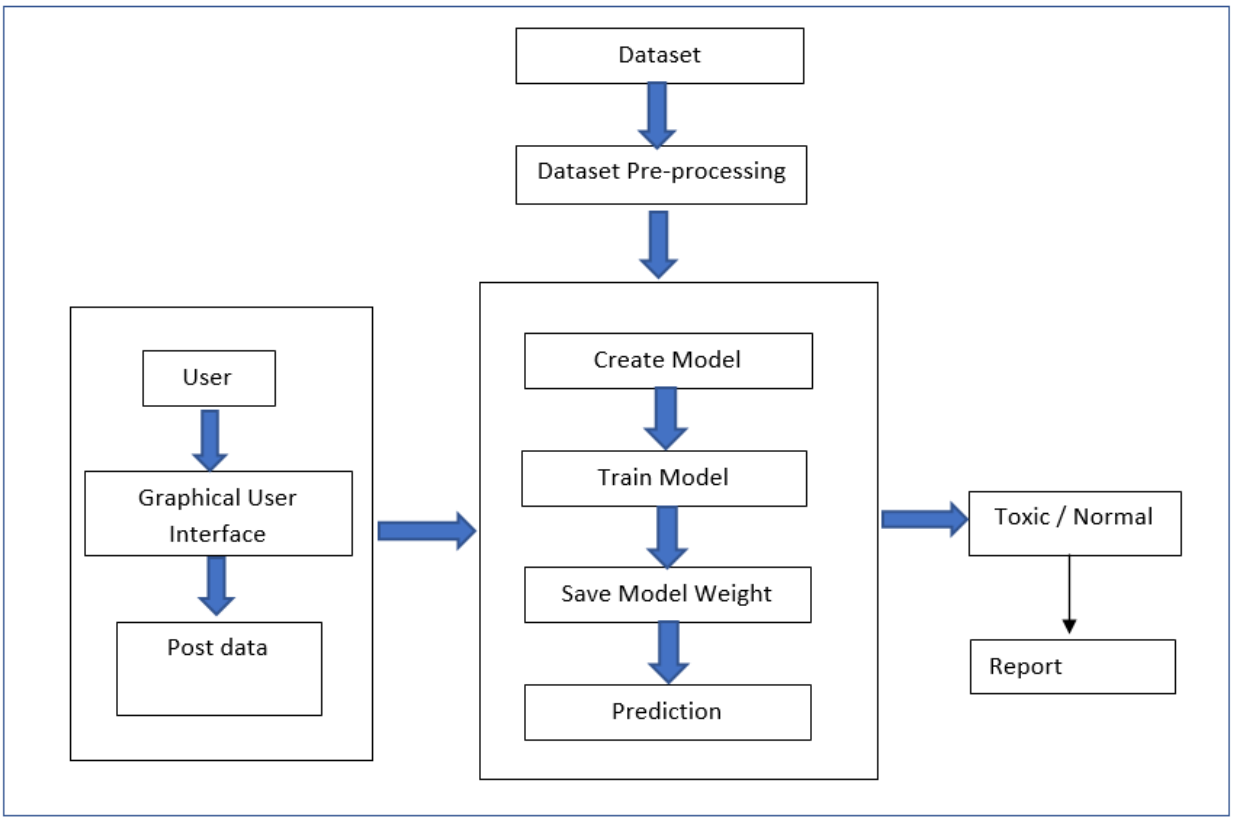
3. **Preprocessing**: The submitted comments undergo preprocessing, which includes tasks like tokenization, lowercasing, and punctuation removal to standardize the text data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134174



4. **Feature Extraction**: Relevant features are extracted from the preprocessed comments. These features could include word frequency, sentiment scores, and linguistic patterns.

5. **Model Building**: A model, such as an AI sentiment tracker or machine learning classifier, is built based on the extracted features to predict the toxicity level or sentiment of the comments.

6. **Training**: The model is trained using labeled data, where comments are labeled as toxic or non-toxic based on predefined criteria or human moderation.

7. **Evaluation**: The trained model's performance is evaluated using validation data, measuring metrics such as accuracy,
precision, recall, and F1-score to assess its effectiveness in detecting toxic comments.

8. **Prediction**: Once the model is trained and validated, it can predict the toxicity level or sentiment of new, unseen comments submitted by users in real-time.

9. **Moderation and Notification**: Toxic comments detected by the model are flagged for moderation, and appropriate actions are taken, such as hiding the comment or notifying the user.

10. **Report Generation**: Finally, a report summarizing the model's performance, including metrics and classification results, is generated periodically to monitor and evaluate the effectiveness of the moderation system.

B. ALGORITHM USED

The algorithms used in this project:

- MultinominalNB

MultinominalNB :

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP).

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134174

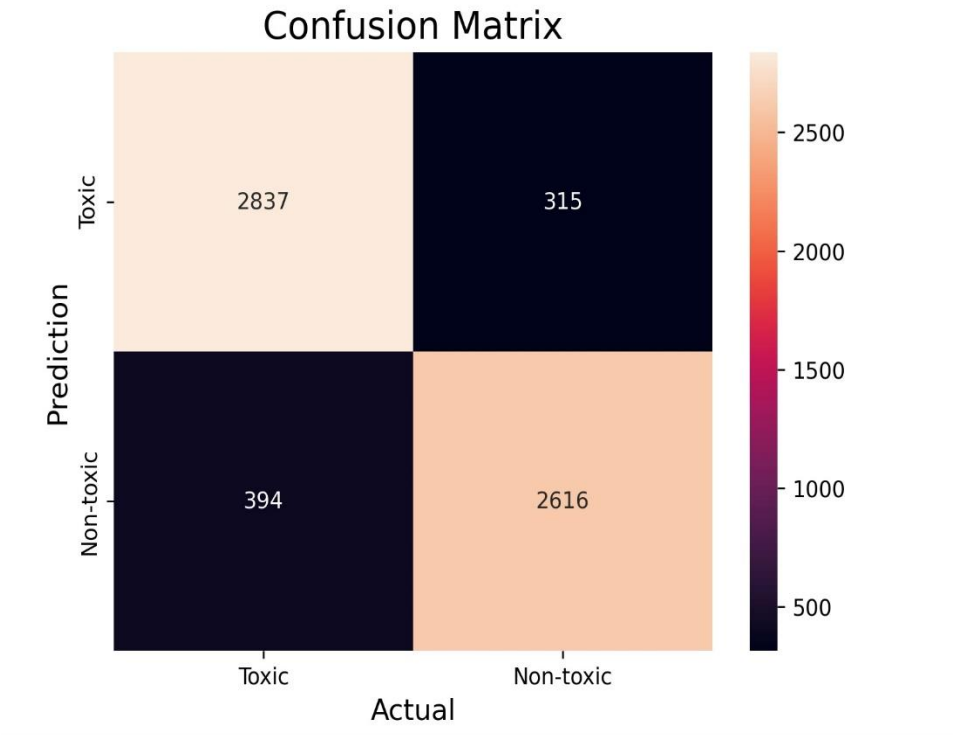
The program guesses the tag of a text, such as an comments or a social m, using the Bayes theorem. It calculates each tag's likelihood for a given sample and outputs the tag with the greatest chance.

Naive Bayes is a probabilistic algorithm family based on Bayes' Theorem. It's "naive" because it presupposes feature independence, which means that the presence of one feature does not affect the presence of another (which may not be true in practice).

Multinomial Naive Bayes is a probabilistic classifier to calculate the probability distribution of text data, which makes it well-suited for data with features that represent discrete frequencies or counts of events in various natural language processing (NLP) tasks.

VI. OUTPUT

Below figure shows the heatmap of this project



Heatmap

The figure shows the confusion matrix ,evaluating the performance of classification model that distinguish between "Toxic" and "Non-toxic". And above we used about 35000 data sets which are 75% are trained and 25% are used for testing the texts. Using these predicted texts we drown heatmap and same is used for output interface.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134174

In Figure 1 we can see the login page of the project which a admin can login through it.

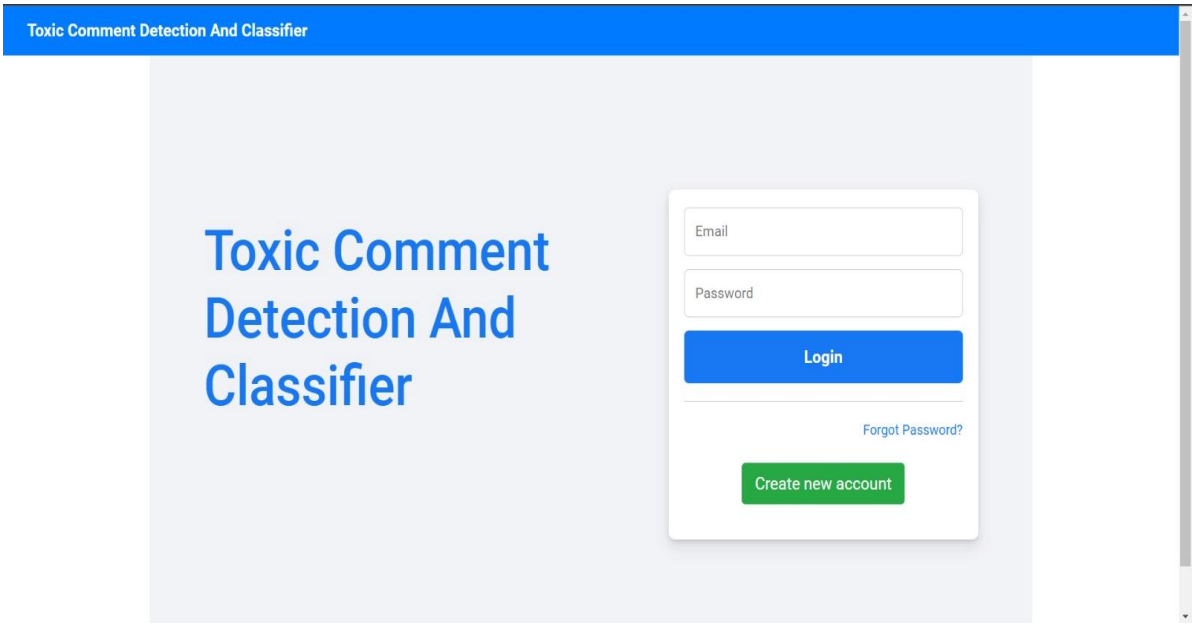


Figure 1

In Figure 2 it shows signup page where new user can have their own account.

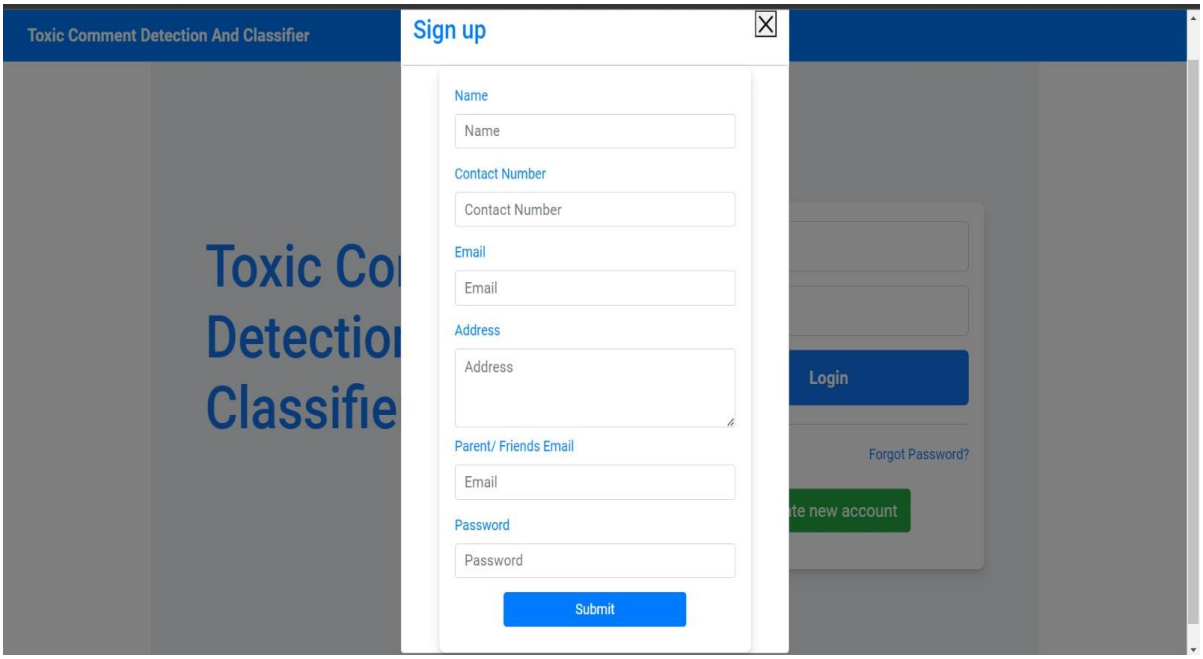


Figure 2

In Figure 3 we could see the dashboard of our project. This is a locally created social media website

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134174

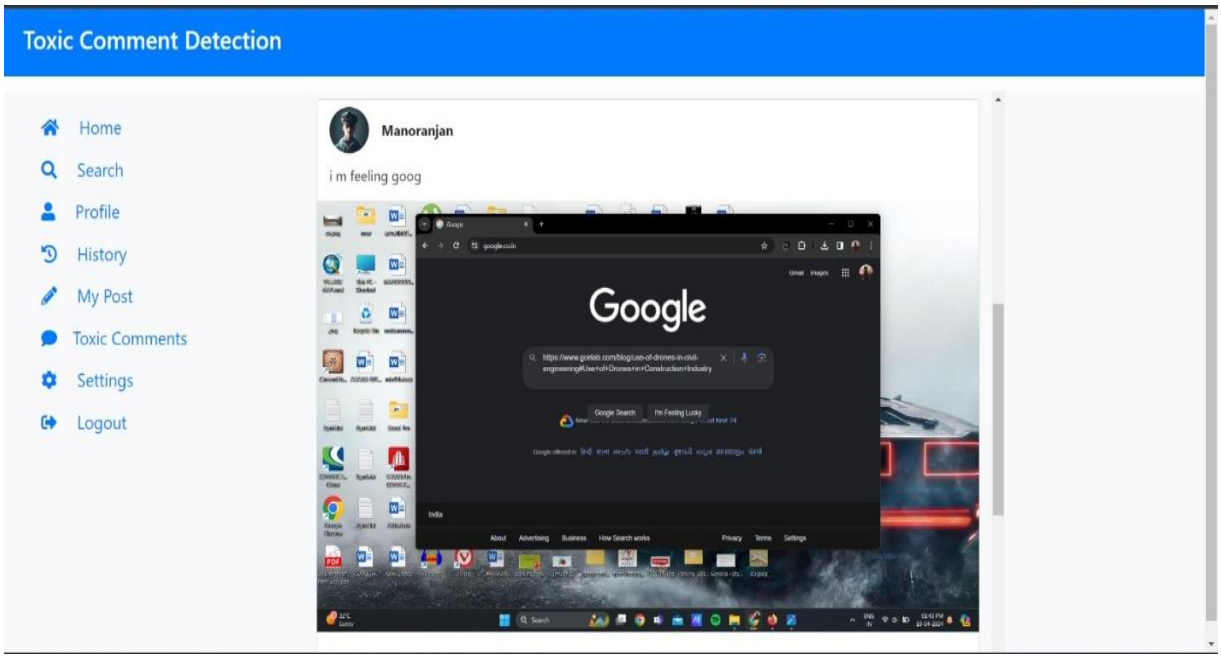


Figure 3

In Figure 4 this page detects the toxic comments to the user who posted

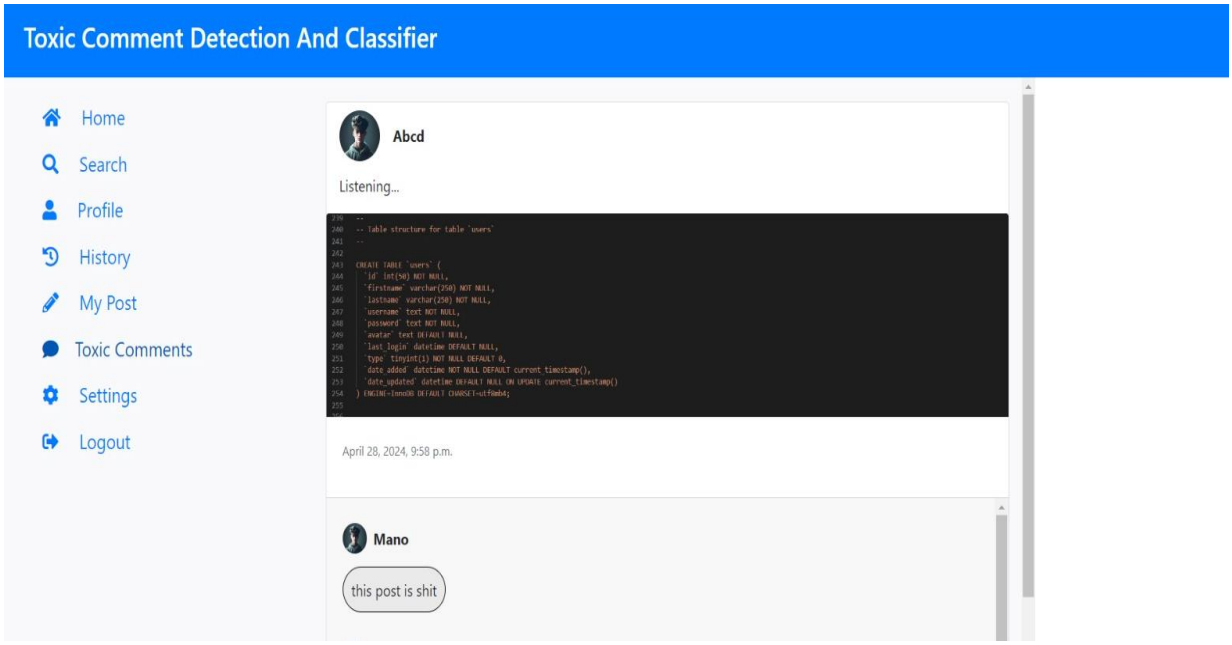


Figure 4

In figure 5 we can identify the classification result using http methods which is seen in the terminal

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134174

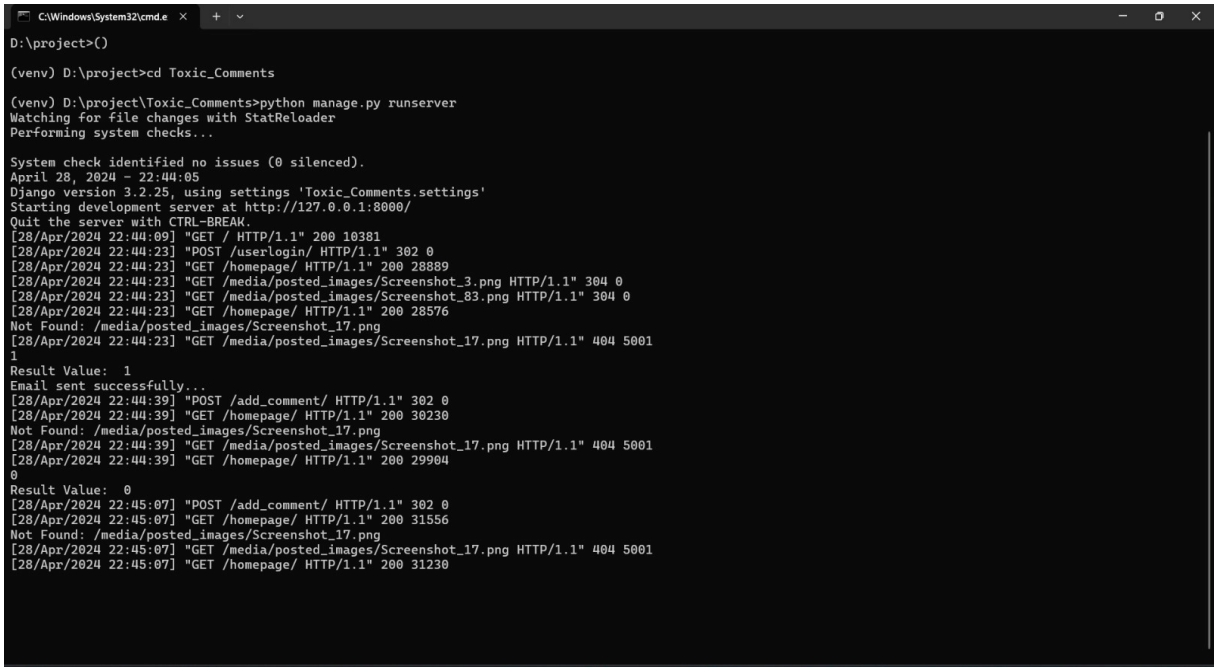


Figure 5

In Figure 6 it represents the Gmail template send to user for every toxic comment detected with helpline services

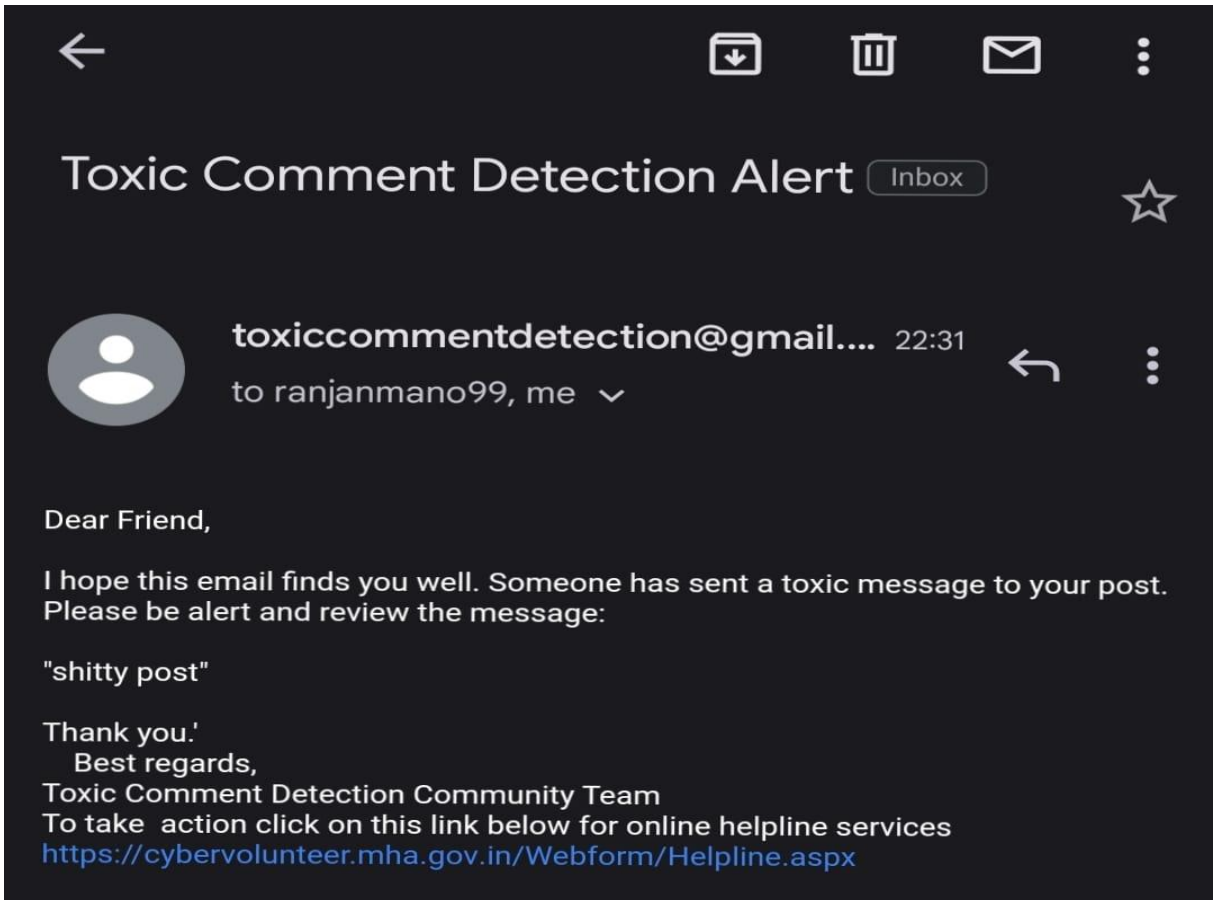


Figure 6

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134174

VII. CONCLUSION

In conclusion, a big step toward building a more secure and welcoming online community has been made with the creation of the locally hosted social media platform equipped with integrated machine learning capabilities for the detection of toxic comments. We have effectively developed a platform that promotes user safety and well-being while encouraging constructive interactions and community engagement by utilizing Python, Django and other technologies.

By utilizing machine learning models for real-time comment moderation, users may quickly identify and flag offensive comments, giving them the power to take proactive steps to uphold a kind and encouraging online community. Furthermore, by integrating an email notification system, users can take immediate action to moderate information when it is uploaded on their accounts and are promptly notified to any potentially hazardous content. We have confirmed the dependability and efficiency of the platform in identifying and reducing harmful behavior through thorough testing and validation. The platform's functionality and user experience have been improved through user and stakeholder feedback, which has also increased the platform's overall efficacy.

Future-focused maintenance and updates will be necessary to meet changing user needs and problems. The platform's significance and influence in the dynamic world of online social interaction will depend on its ability to continuously improve and adapt to new technologies and user patterns. In general, the creation of this initiative serves as a reminder of how crucial it is to use technology to advance online safety and cultivate a polite and respectful culture in social media settings. We've come a long way in making the internet experience more enjoyable and fulfilling for everyone by fusing creative thinking with user-centered design principles.

VIII. FUTURE WORK

In future rounds of the research, using an AI sentiment tracker instead of traditional machine learning (ML) models may improve comment moderation efficacy. The platform may be able to more accurately identify toxic conduct by utilizing AI techniques like sentiment analysis and natural language understanding (NLU) to better understand the complex context and emotive tone of user comments. This strategy might produce better moderating results and a more flexible platform that can change to accommodate changing online interaction patterns. By using AI sentiment tracking, communities may develop a more compassionate and understanding culture and users may gain a more sophisticated understanding of their online interactions.

REFERENCES

[1]. M. Husnain, A. Khalid, and N. Shafi, “A Novel Preprocessing Technique for Toxic Comment Classification,” in Proc. ICAI, Online, Apr. 2021, pp. 22–27.
[2]. H. Almerekhi, H. Kwak, J. Salminen, and B. J. Jansen, "Predicting Triggers of Toxicity in Online Discussions," in Proc. of The Web Conf. 2020, Taipei, Taiwan, Apr. 2020, pp. 3033-3040
[3]. S. Zaheri, J. Leath, and D. Stroud, "Toxic Comment Classification," SMU Data Sci. Rev., vol. 3, no. 1, Art. 13, 2020.
[4]. Rahul, and H. Kajla, "Classification of Online Toxic Comments Using Machine Learning Algorithms," in Proc. ICICCS 2020, 2020
[5]. D. A. Coc, "Machine learning methods for toxic comment classification: a systematic review," Acta Univ. Sapientiae Inform., vol. 12, no. 2, pp. 205-216, 2020.
[6]. N. Frank and G. Simmons, "Comparative Analysis of Machine Learning Algorithms for Online Harassment Detection," J. Inform. Technol. Appl., vol. 22, no. 3, pp. 324-340, 2021.
[7]. B. Thompson and Y. Choi, "Text Classification Techniques for Detecting Hate Speech," Data Sci. J., vol. 19, no. 1, pp. 101-110, 2018.
[8]. R. Gomez and J. Patel, "Using NLP and Machine Learning to Combat Cyberbullying," Comput. Secur. Rev., vol. 15, no. 4, pp. 290-298, 2019.
[9]. L. Harper and A. Johnson, "Deep Learning Approaches for Detecting Offensive Language in Social Media," Soc. Netw. Anal. Min., vol. 10, no. 1, pp. 55-65, 2020.
[10]. E. Brown and R. Clarke, "Enhancing Toxic Comment Classification with Deep Neural Networks," J. Data Sci., vol. 9, no. 2, pp. 180-195, 2020.

