International Journal of Research in Engineering, Science and Management
Volume 7, Issue 4, April 2024
https://www.ijresm.com | ISSN (Online): 2581-5792

Toxic Comments Detection and Classifier

Adarsh Vinod[1], K. V. Adithyan[2*], M. Manoranjan[3], Ramsha Riyaz[4], N. Arul[5]
[1,2,3,4]*Student, Department of Computer Science and Engineering, A. J. Institute of Engineering and Technology, Mangalore, India*
[4]*Assistant Professor, Department of Computer Science and Engineering, A. J. Institute of Engineering and Technology, Mangalore, India*

***Abstract*: This project proposes a novel approach to detecting and managing toxic comments online. It detects harmful content effectively using a smart machine learning system. Users play an important role by providing an easy reporting system and quick actions to hide or block toxic comments. The platform is intended to empower users by providing customizable filters, an education hub, and a reward system that encourages positive online behaviour. Transparency is a top priority, with users receiving detailed moderation histories and real-time alerts. Additional features, such as content dispute resolution, inclusive language suggestions, and collaborative moderation tools, aim to make the online environment safer, more inclusive, and enjoyable. This project also looks into user-friendly admin tools, personalized content filters, and even blockchain for transparency. And also, by keeping things simple and effective, our machine learning-focused approach aims to redefine content moderation, creating a safe, collaborative, and enjoyable online environment.**

***Keywords*: toxic comments, toxicity, personal assaults, hate speech.**

1. Introduction

In today's world of online communication, dealing with harmful comments has become a big challenge. This project is all about creating a smart system or a model that can automatically find and categorize these toxic comments. Internet is an open discussing space for everyone to freely express their opinions. However, harassment and abuse are discouraging people from sharing their ideas and disturbing the internet environment. By using a diverse dataset, we're ensuring that our system can handle various types and intensities of toxic language. It's not just about spotting harmful comments but also categorizing them based on how severe they are. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments if it's toxic. Motivated by this problem, we want to build technology to protect voices in conversation by machine learning models that can identify toxicity in online conversations, where toxicity is defined as anything rude, disrespectful or otherwise likely to make someone leave a discussion. For our model, the input will be a comment. Then the ensemble methods are studied to combine multiple learning algorithms and get the best result. The system is designed not only to identify toxic content but also to educate users about the criteria used for classification, fostering awareness and promoting responsible online communication. As we delve into the details of data collection, model development, and web application deployment, this project seeks to provide a comprehensive solution to the pervasive issue of toxic comments, contributing to the creation of digital spaces that prioritize positive interactions and community well-being. By combining technical expertise, ethical considerations, and a commitment to user education, this project endeavors to contribute to a safer, more inclusive online environment where individuals can freely express themselves while mitigating the negative consequences of toxic communication.

2. Problem Formulation

Toxic comments are a common problem in online communication that make it difficult to maintain the inclusive and positive atmosphere of digital interactions. The creation of an automated system for the accurate detection and categorization of harmful language in user-generated content is the main issue this study attempts to solve. Beyond simple identification, this system needs to be able to classify harmful remarks into more complex categories like hate speech, personal assaults, or derogatory language. The careful reduction of biases present in the classification model is a crucial aspect of the problem formulation. Acknowledging the possible impact of variables such as gender and race on the model's forecasts, the study attempts to put policies in place that guarantee impartial and equitable results. In addition, the dynamic character of online language demands a model that can adjust to changing linguistic variances, colloquialisms, and trends in order to maintain accuracy over time. Using real-time detection mechanisms is just another crucial component of the problem formulation process. In order to enable prompt intervention and moderation, this calls for the smooth integration of the toxic comment detection model into active online environments. Concurrently, the study recognizes the importance of user feedback in improving the accuracy of the model. Through the implementation of a feedback loop designed to incentivize users to report misclassifications, the system embraces a collaborative methodology for ongoing enhancement.

3. Methodology

The flowchart in figure 1 outlines the process of toxic comments detection and classification.

*Corresponding author: adithyankv77@gmail.com

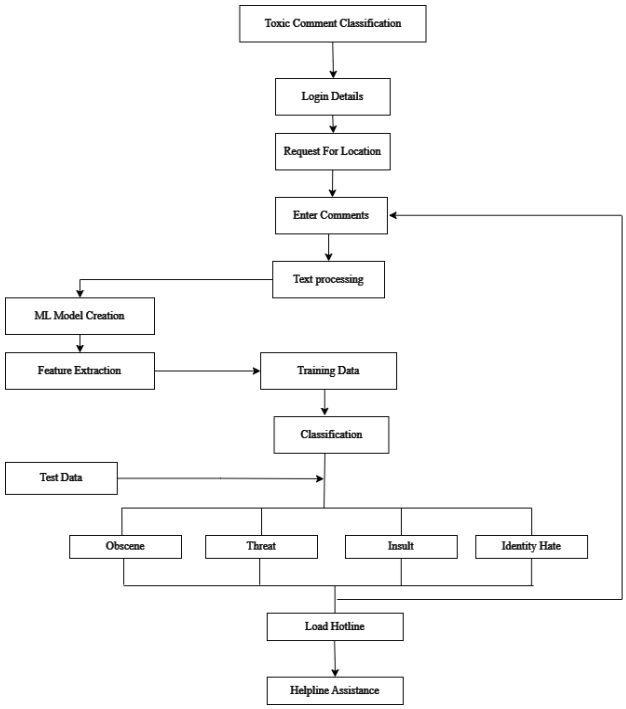

Fig. 1. Flow diagram

1. *Login Details:* Here the user must provide their name and Email id.
2. *Request for Location:* The web page pops an alert box for location request.
3. *Enter Comments:* User needs to enter the comments.
4. *Text Processing:* Here the user comment is decoded into separate keywords.
5. *ML Model creation and Feature Extraction:* This stage includes collecting datasets from various sources.
6. *Training Data:* In this phase the datasets are trained.
7. *Classification:* Here we provide keywords for the ML model to classify the comment on basis of this keyword.
8. *Test Data:* This is the dataset used to test the accuracy of our ML model.
9. *Toxicity Level:* The comments are classified based on the level of toxicity such as Obscene, Threat, Insult, Identity Hate.
10. *Load Hotline:* The website provides different helpline links based on the toxicity level.
11. *Helpline Assistance:* This the last landing page for user assistance.

In summary, the flowchart describes the about our project 'Toxic comments and detection', means it aims a good online environment. This flowchart aims to enhance understanding among developers and stakeholders by illustrating the sequential steps, decision points, and interactions within the system. The main aim is to present a clear and comprehensive visual guide to the systematic process involved in identifying and managing toxic comments within an online platform.

A. Algorithm Used

The algorithms used in this project: MultinominalNB

1) MultinominalNB

The Multinomial Naive Bayes algorithm is a Bayesian learning approach popular in Natural Language Processing (NLP). The program guesses the tag of a text, such as an email or a newspaper story, 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.

4. Conclusion

The conclusion emphasizes the importance of having tools that can find and stop mean comments online. It's like having a friendly guardian for the internet. These tools are like superheroes that protect us from online bullies and make sure our online chats are safe and enjoyable. When these tools are in action, they reduce the number of hurtful comments, making the internet a better place. This not only makes people happier while chatting but also helps everyone feel more comfortable expressing themselves online. Imagine these tools as a shield that keeps bad stuff away, allowing us to have more positive and friendly interactions. As more and more people join the online world, these tools need to keep improving to handle new challenges and keep the internet a cool and welcoming space for everyone. In short, using tools to find and stop mean comments is like having a friendly guide for our online adventures, making sure we all have a great time while staying safe and happy in the digital world.

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.