ELIMINATING TOXICITY: A NOVEL APPROACH TO COMMENT CLARIFICATION

Aakash Ayachit, Soham Khune, Devang Mathur, Vaidehi Mujumdar, Vedraj Belokar

Computer Engineering, Ajeenkya DY Patil SOE, Pune, MHA, India

ABSTRACT

The prevalence of toxic comments on online platforms has made automated content moderation a critical necessity. According to 2023 statistics, India has the highest rate of cyberbullying worldwide, and around 85% of the children report it. The prevalence of cyberbullying victimization increased from 3.8% to 6.4% among female respondents and 1.9% to 5.6% among male respondents over three years. About 33% of females and 16.6% of males had depressive symptoms in their young adulthood. This comprehensive review and analysis encompass many primary studies, investigating the landscape of toxic comment classification using machine learning. The research explores publication trends, dataset usage, evaluation metrics, machine learning techniques, toxicity categories, and comment languages. Online toxicity poses significant challenges, and this analysis identifies gaps in current research while offering insights into the future of automated content moderation. Leveraging machine learning algorithms, the studies reviewed aim to improve the accuracy of toxic comment classification, contributing to the creation of a safer and more respectful online environment.

Keywords: Toxic comment classification, Machine learning, Content moderation, Online toxicity, Text analysis, Classification algorithms, Ethical considerations, Automated moderation, Toxicity categories.

I. INTRODUCTION

In the contemporary digital age, online communication and discussions have ushered in a new era of global connectivity and information exchange. The internet, with its vast array of platforms and communities, offers individuals a powerful voice and a means to participate in a multitude of conversations.

Yet, within this expansive digital landscape, a sinister issue looms large – the proliferation of toxic comments.

Toxic comments, encompassing hate speech, harassment, profanity, personal attacks, and various forms of vitriol, have emerged as a potent threat to the well-being and psychological safety of online users. The unrestricted, open nature of online discussions has provided a breeding ground for toxicity, where even a single toxic comment can send shockwaves through the digital world and negatively affect young and mature minds alike.

The effects of toxic comments on individuals are profound. Online toxicity breeds fear, anxiety, and insecurity, leading to self-censorship and withdrawal from the online space. It undermines the very essence of a democratic digital world where every voice should be heard, respected, and valued. The consequences extend beyond individuals, affecting the overall health and inclusivity of digital communities.

To address this issue, this project leverages the power of Machine Learning (ML) and Natural Language Processing (NLP) to create a robust and adaptive system for classifying comments as toxic or non-toxic. This endeavor is backed by binary classification, an approach that provides a clear and effective method for distinguishing between these two categories. The primary aim is to empower online platforms and communities with a tool that can automatically detect toxic comments, ensuring the well-being and security of users.

By delving into the realms of ML and NLP, this project aspires to offer a solution that is not only effective in identifying toxicity but is also capable of adapting to evolving forms of toxic language. It is a commitment to the well-being of individuals in the online space, a testament to the belief that digital discussions should be a realm of respect, inclusion, and free expression.

In the pages that follow, we will journey through the intricacies of this project, exploring the methodologies, techniques, and technologies employed in the quest to eliminate toxicity from online discourse. We will delve into the power of ML and NLP in understanding human language, discover the nuances of binary classification, and unravel the inner workings of the system designed to foster a more secure and inclusive online environment. Together, we embark on a mission to safeguard the digital realm from the perils of toxic comments and ensure that the voices of the online world are heard, respected, and protected.

II. METHODS AND MATERIAL

NAIVE BAYES CLASSIFIER

Naive Bayesian classification is a probabilistic approach to machine learning. It is based on the Bayes Theorem.

The probability of A happening knowing that B has occurred could be calculated.

The theorem runs on the assumption that all predictors/features are independent and the presence of one would not affect the other.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P (message is toxic | message content): This is the posterior probability that a given message is toxic given its content.

P(message content | toxic) : This is the likelihood of observing the given content if the message is toxic. In the context of a Naive Bayes classifier, it's often calculated as the product of the probabilities of individual words occurring in toxic messages.

P(Toxic): This is the prior probability of a message being toxic. It represents the overall likelihood of a message being toxic without considering its content.

P(message content): This is the probability of observing the given content in any message, toxic or not. It acts as a normalization factor.

Naive Bayes classifiers often perform well in text classification tasks, including toxic comment classification.

$$P(message\ is\ toxic\ |message\ content\} = \frac{P(message\ content|toxic)P(Toxic)}{P(message\ content)}$$

Classifying toxic comments is typically approached as a binary classification problem, where WE want to determine whether a given comment is toxic (1) or non-toxic (0).

n = number of training examples

m = number of features (words or tokens in the comment)

X = feature matrix of size (n,m), where each row represents a comment and each column is a binary feature indicating the presence (1) or absence (0) of a particular word or token.

y = target vector of size (n,), where yi is the label for the i-th comment (0 for non-toxic, 1 for toxic).

θ = parameter vector of size (m,), where θj represents the weight for the j-th word or token.

The logistic regression model can be represented as:

Hypothesis Function (Logistic Function):

The hypothesis function is used to predict the probability that a given comment is toxic (class 1):

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Here, x is the feature vector for a particular comment, and is the transpose of the parameter vector

Cost Function (Logistic Loss):

The logistic loss function is used to measure the error between the predicted probabilities and the actual labels: This is a binary cross-entropy loss function.

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(h_\theta(x_i)) + (1 - y_i)\log(1 - h_\theta(x_i))]$$

Parameter Update (Gradient Descent):

To minimize the cost function and learn the optimal parameters 0, you can use gradient descent. The update rule for gradient descent is as follows:

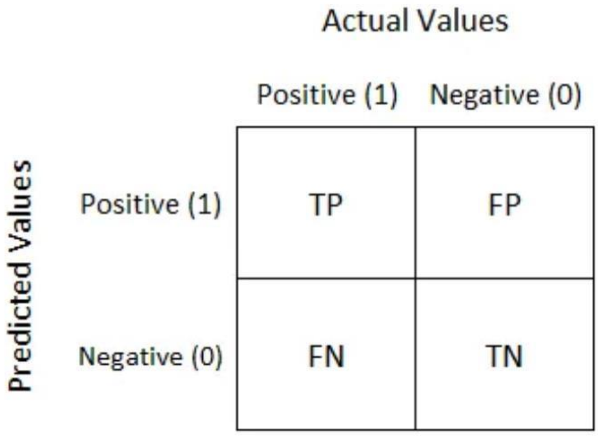
$$\theta_j = \theta_j - \alpha\frac{1}{n}\sum_{i=1}^{n}(h_\theta(x_i) - y_i)x_{ij}$$

Where alpha is the learning rate, and is the value of the -th feature for the -th comment.

METRICS - Confusion Matrix

Used to evaluate performance of the algorithms.

Useful for measuring Recall, Precision, Accuracy, AUC-ROC curves etc



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Accuracy (all correct / all) = TP + TN / TP + TN + FP + FN

Misclassification (all incorrect / all) = FP + FN / TP + TN + FP + FN

Precision (true positives / predicted positives) = TP / TP + FP

Sensitivity aka Recall (true positives / all actual positives) = TP / TP + FN

Specificity (true negatives / all actual negatives) =TN / TN + FP

F-1 Score (harmonic mean of precision and recall) = 2 x P x R / P + R

III. RESULTS AND DISCUSSION

A. **Figures**

Fig.1: Dataset Distribution

This figure illustrates the distribution of comments across different toxicity levels in the dataset used for training and testing the classification model.

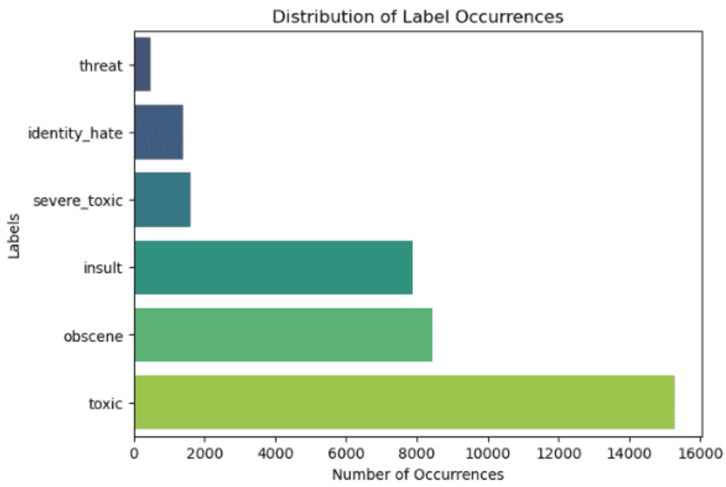


Figure 1: Distribution of comments across toxicity levels in the dataset.

Fig.2: UML Diagram for Toxic Comment Classification System

This figure presents a Unified Modeling Language (UML) diagram illustrating the key components and their relationships in the toxic comment classification system.

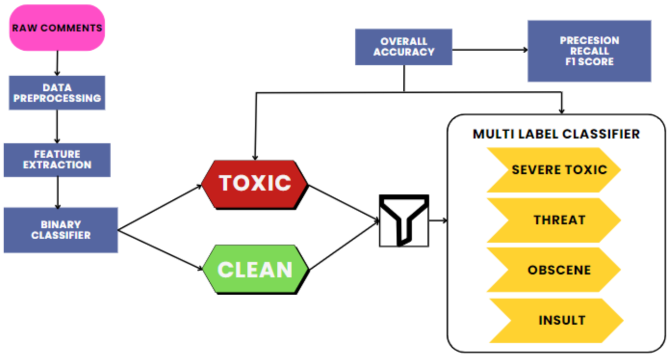


Figure 2: UML diagram depicting the components and relationships in the toxic comment classification system.

Fig.3: Model Architecture

This figure outlines the architecture of the toxic comment classification model, depicting the layers and connections within the network.

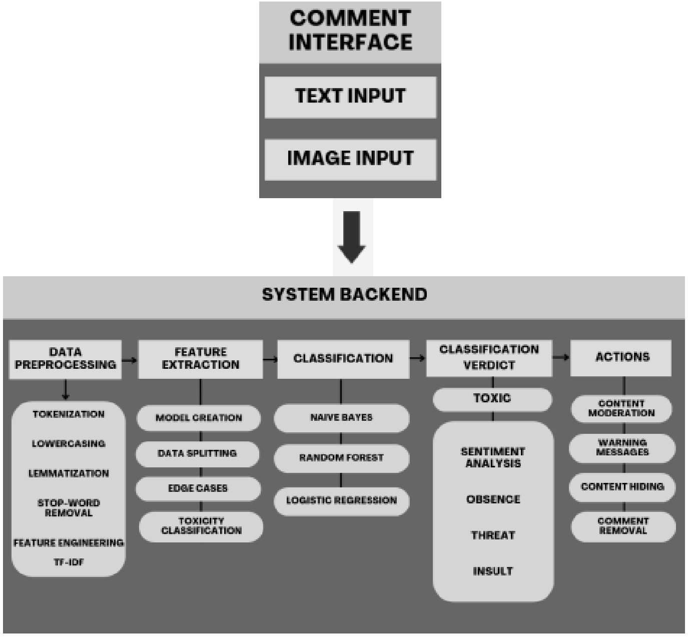


Figure 3: Model architecture for toxic comment classification.

IV. CONCLUSION

By developing an automated system capable of identifying and categorizing toxic comments, the project has the potential to create safer and more inclusive online spaces, improve user experiences, and reduce the burden on human content moderators.

The project's advantages include enhanced online safety, user protection, and the ability to efficiently moderate content at scale.

It also provides valuable data-driven insights, contributes to legal compliance, and helps build a positive public image for platforms.

However, it is essential to acknowledge the project's limitations, including the potential for false positives, sensitivity to context, algorithmic biases, and challenges in keeping up with evolving forms of toxicity.

Future Work may involve reducing false positives and negatives, further research and development are needed to minimize them in toxic comment classification. This involves improving the model's understanding of context and intent. New models and algorithms will also be used to improve model performance. The project will also be able to adapt to rapidly changing online environments and emerging forms of toxic behaviour through real-time updates and continuous learning.

REFERENCES

[1] "Machine learning methods for toxic comment classification: a systematic review" by Darko Androcec (University of Zagreb) DOI:10.2478/ausi-2020-0012

[2] Toxic Comment Detection and Classification, by Hao Li, Weiquan Mao, Hanyuan Liu.

[3] Multilingual Toxic Text Classification Model Based On Deep Learning by Wenji Li; Anggeng Li, JiangXi Agricultural University, Nanchang, Tianqi Tang; Yue Wang; Zejian Fang DOI: 10.1109/ICBAIE56435.2022.9985930

[4] Classification of Toxicity in Comments using NLP and LSTM by Anusha Garlapati; Neeraj Malisetty; Gayathri Narayanan DOI: 10.1109/ICACCS54159.2022.9785067

[5] A Novel Preprocessing Technique for Toxic Comment Classification by Muhammad Husnain; Adnan Khalid; Numan Shafi DOI: 10.1109/ICAI52203.2021.9445252

[6] Toxic Comment Analysis for Online Learning by Manaswi Vichare; Sakshi Thorat; Cdt. Saiba Uberoi; Sheetal Khedekar; Sagar Jaikar DOI: 10.1109/ACCESS51619.2021.9563344

[7] Analysis of Multiple Toxicities Using ML Algorithms to Detect Toxic Comments by KGSSV Akhil Kumar; B. Kanisha DOI: 10.1109/ICACITE53722.2022.9823822

[8] Machine Learning-based Multilabel Toxic Comment Classification by Nitin Kumar Singh; Satish Chand DOI: 10.1109/ICCCIS56430.2022.10037626

[9] Bangla Toxic Comment Classification and Severity Measure Using Deep Learning by Naimul Haque; Md. Bodrul Alam; Abdullah Ath Towfiq; Mehorab Hossain DOI: 10.1109/ICRPSET57982.2022.10188551

[10] Classification of Online Toxic Comments Using Machine Learning Algorithms by Rahul; Harsh Kajla; Jatin Hooda; Gajanand Saini DOI: 10.1109/ICICCS48265.2020.9120939.

[11] H. M. Saleem, K. P . Dillon, S. Benesch, and D. Rut hs, "A Web of Hate: Tackling Hateful Speech in Online Social Spaces," 2017, [Online]. Available: http://arxiv.org/abs/1709.10159.

[12] M. Duggan, "Online harassment 2017," Pew Res., pp. 1 – 85, 2017, doi: 202.419.4372.