A PRELIMINARY REPORT ON

ELIMINATING TOXICITY: A NOVEL APPROACH TO COMMENT CLARIFICATION

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE
IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE

OF

BACHELOR OF ENGINEERING (COMPUTER ENGINEERING)

SUBMITTED BY

AAASH AYACHIT	Exam No : **B190884202**
SOHAM KHUNE	Exam No : **B190884275**
DEVANG MATHUR	Exam No : **B190884283**
VAIDEHI MUJUMDAR	Exam No : **B190884287**
VEDRAJ BELOKAR	Exam No : **B190884217**

Under the Guidance of
Prof. Amruta Chitari

AJEENKYA
DY Patil School of Engineering

DEPARTMENT OF COMPUTER ENGINEERING

AJEENKYA DY PATIL SCHOOL OF ENGINEERING
CHARHOLI BUDRUK, LOHEGAON, PUNE 412105

SAVITRIBAI PHULE PUNE UNIVERSITY
2023 -2024

AJEENKYA
DY Patil School of Engineering

AJEENKYA DY PATIL SCHOOL OF ENGINEERING
DEPARTMENT OF COMPUTER ENGINEERING

CERTIFICATE

This is to certify that the Project Report Entitled

"ELIMINATING TOXICITY: A NOVEL APPROACH TO COMMENT CLARIFICATION"

Submitted by

AAKASH AYACHIT	Exam No : B190884202
SOHAM KHUNE	Exam No : B190884275
DEVANG MATHUR	Exam No : B190884283
VAIDEHI MUJUMDAR	Exam No : B190884287
VEDRAJ BELOKAR	Exam No : B190884217

is a bonafide student of this institute and the work has been carried out by him/her under the supervision of **Prof. AMRUTA CHITARI** and it is approved for the partial fulfillment of the requirement of Savitribai Phule Pune University, for the award of the degree of **Bachelor of Engineering** in Computer Engineering.

Prof. Amruta Chitari
Guide
Department of Computer Engineering

Dr. Pankaj Agarkar
H.O.D
Department of Computer Engineering

Dr. F. B .Sayyad
Principal

Signature of Internal Examiner:

Signature of External Examiner:

Place : Pune
Date : 03-11-2023

ACKNOWLEDGEMENT

I would like to express my gratitude and appreciation to all those who gave me the possibility to do this project. I am deeply thankful to my project guide, Professor Amruta Chitari, for her continuous support, expert guidance, and invaluable insights that were instrumental in shaping this project.

I extend my sincere thanks to the Head of the Department, Dr. Pankaj Agarkar, for providing the necessary resources and creating an environment conducive to research and learning.

I would also like to acknowledge the Principal, Mr. F.B. Sayyad, for his unwavering support and encouragement.

Lastly, I am thankful to all the individuals, including my team members and peers, who provided their valuable input and encouragement throughout the course of this project. Their collective contributions have been invaluable in making this project a reality.

Aakash Ayachit	Exam No : B190884202
Soham Khune	Exam No : B190884275
Devang Mathur	Exam No : B190884283
Vaidehi Mujumdar	Exam No : B190884287
VEDRAJ BELOKAR	Exam No : B190884217

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.

TABLE OF CONTENTS

LIST OF ABBREVATIONS

ABBREVIATION	ILLUSTRATION
NLP	Natural Language Processing
TF-IDF	Term Frequency-Inverse Document Frequency
BOW	Bag of Words
API	Application programming interface
TP, TN, FP, FN	True Positive, True Negative, False Positive, False Negative
ROC Curve	Receiver Operating Characteristic Curve
AUC	Area Under the Curve

LIST OF FIGURES

LIST OF TABLES

01. 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.

1.1 MOTIVATION

In an age dominated by digital communication, the need for a safer online environment has never been more critical. Toxic comments not only erode the quality of discourse but also pose significant risks to vulnerable users, including children and marginalized communities. Balancing the principles of free expression with the imperative to safeguard users is the cornerstone of a healthy online space.

The impact of toxic comments resonates across various dimensions, urging us to address this growing concern through innovative means. Our motivation springs from the following imperatives:

1. Safer Internet: Toxic comment classification is pivotal in creating a safer online environment. By efficiently identifying and managing toxic content, we contribute to a digital space that safeguards individuals against harmful language and abusive behavior.

2. Vulnerable User Protection: Our commitment extends to protecting vulnerable users, including children and marginalized communities. Through effective comment classification, we aim to shield those most at risk from online harm and discrimination.

3. Balancing Speech and Safety: Striking a balance between freedom of expression and user protection is crucial. We aspire to achieve a harmonious coexistence of open dialogue and online safety. This equilibrium empowers individuals to express their views while also fostering a respectful and secure digital realm.

4. Fostering Positivity: Witnessing the harm inflicted by toxic comments upon the online community, we are compelled to act. Our project endeavors to use technology as a force for good, creating an atmosphere where healthy discussions can flourish without being overshadowed by negativity. We envision a digital space that cultivates respect and kindness, thereby enhancing the quality of our digital interactions.

Through the implementation of advanced classification techniques, we aspire to create a platform that fosters positivity, ensuring that our online interactions reflect the best of human discourse.

1.2 LITERATURE SURVEY

Paper[1] conducted a systematic review of cutting-edge methodologies in classifying toxic comments using machine learning techniques. Data was meticulously extracted from 31 pertinent primary studies. The analysis encompassed an examination of publication timelines, venues, and the level of maturity of the papers. Furthermore, each primary study was scrutinized for details on

the employed dataset, evaluation metrics, methodologies in machine learning, types of toxicity classes, and the language of comments.

In toxic comment classification, key evaluation metrics include the F1 score, accuracy, and Area under the ROC Curve (AUC ROC). The F1 score strikes a balance between precision and recall, providing a comprehensive measure of model performance. Accuracy offers a straightforward assessment of overall correctness, while AUC ROC evaluates the model's ability to distinguish between classes. These metrics collectively offer valuable insights into the model's effectiveness in classifying toxic comments.

The authors of paper[2] utilized a three-tier approach for toxic comment detection. It used three different models to predict scores for “toxicity”(which is our target), “severe toxicity”, obscenity”, “identity attack”, “insult”, “threat”. Then the ensemble methods are studied to combine multiple learning algorithms and get the best result. It commenced with the Naive Bayes-SVM baseline model, followed by exploration of deep learning with LSTM and BERT. Subsequent steps involved applying ensemble methods, contributing to an overall improvement in the project's ability to identify toxic comments. In the end, two ensemble methods were used and increased the F1 and EM scores to 84.28% and 95.14%.

In paper[3], the study leverages the XLM-RoBERTa model for the categorization of multilingual toxic remarks.XLM-RoBERTa is a transformer-based model for natural language processing. It's an extension of RoBERTa that's pre-trained on multiple languages, making it highly effective for multilingual tasks. The process involves initial training and fine-tuning of the model using training and validation datasets, followed by evaluation on test data to ascertain the final classification outcomes. Furthermore, the model's performance is assessed in comparison to alternatives like LSTM and RNN.

LSTM is a type of recurrent neural network (RNN) architecture. It's designed to handle sequences of data and can capture long-term dependencies. RNN is a class of neural networks that have connections looping back on themselves. This looped architecture allows RNNs to process sequences of inputs, making them suitable for tasks where context or order is crucial, such as language modeling or time series prediction. Experimental results affirm that the proposed model in this research demonstrates superior classification efficacy.

This research[4] endeavors to leverage Natural Language Processing techniques to effectively categorize various forms of toxic content, including but not limited to obscene, identity hate, threat, toxic, insult, and severe toxic remarks. The algorithm takes comments from online platforms, distinguishing between toxic and non-toxic content. The model's primary goal is to predict the toxicity class. The study progresses in two phases: Phase I focuses on assessing comment toxicity using techniques like TFIDF and Spacy, enabling a nuanced understanding of word categorization within specific toxic classes. In Phase II, data is further scrutinized to categorize comments as toxic or non-toxic, enhancing the ability to discern potentially harmful content.

TFIDF (Term Frequency-Inverse Document Frequency) is a statistical measure used in Natural Language Processing to evaluate the importance of a term in a document relative to a collection of documents.

Spacy is a popular library in NLP that provides pre-trained models and tools for processing and understanding text.

NLP stands for Natural Language Processing, a field of artificial intelligence focused on enabling machines to interpret, understand, and generate human language. LSTM (Long Short-Term Memory) is a type of recurrent neural network architecture that is adept at processing sequences of data, making it particularly useful for tasks involving natural language.

The study[5] introduces a deep learning-based approach for detecting toxic content and 'trolls' in shared posts, stories, and memes. A machine learning solution is proposed to identify toxic images based on the text content embedded within them, leveraging GloVe word embeddings for enhanced predictive capabilities. The methodology involves implementing Long Short-term Memory (LSTM) Gated Recurrent Unit models, both in their standard and Bidirectional variants. Comparative analysis with related works highlights notable enhancements. The experiments establish that the Bidirectional LSTM model stands out, achieving an impressive testing accuracy of 0.92, an inference accuracy of 0.88, along with F1-scores of 0.92 and 0.88. The process comprises two key steps: Image Text Extraction and Text Classification. For accurate extraction, Python-tesseract is employed, primarily designed for English text. Advanced models like Bidirectional LSTM and Bidirectional GRU are used for robust Natural Language Processing in the text classification phase. Additionally, it may be necessary to implement platform-specific data scraping for comprehensive production deployment.Python-tesseract is an Optical Character Recognition (OCR) tool for extracting text from images. Bidirectional LSTM (Long Short-term Memory) and GRU (Gated Recurrent Unit) are advanced neural network architectures capable of processing sequential data in both forward and backward directions. GloVe (Global Vectors for Word Representation) word embeddings are pre-trained word vectors that capture semantic relationships between words in a text corpus. These embeddings are used to enhance the model's understanding of the textual content.

The presented work[6] addresses the pressing issue of a notable increase in negative and toxic comments, leading to a decline in online discourse and an uptick in abusive behavior by deploying machine learning algorithms including Logistic Regression, Random Forest, and Multinomial Naive Bayes. To assess the data, ROC (Receiver Operating Characteristic) and Hamming scores are employed. The results are presented as percentile rates and through graphical representation, shedding light on various categories such as toxic, severely toxic, obscene, threat, insult, and identity hate. This initiative is poised to significantly mitigate online bullying and harassment experienced by educators and learners, fostering a safer and more conducive learning environment. This shift in focus towards learning, unhampered by discouraging comments, will undoubtedly lead to a more enriching educational experience for all.

Logistic Regression: A statistical method used for analyzing a dataset in which there are one or more independent variables that can be used to predict the outcome.

Random Forest: An ensemble learning method used for classification, regression, and other tasks, which operates by constructing a multitude of decision trees at training time.

Multinomial Naive Bayes: A variant of Naive Bayes algorithm suitable for classification tasks with discrete features, particularly effective for text classification.

ROC (Receiver Operating Characteristic) score: A graphical representation of the model's ability to discriminate between positive and negative classes, commonly used in binary classification tasks.

Hamming score: A metric used for evaluating the similarity between two sets by comparing their respective elements. In this context, it's used to assess the performance of the classification model.

This study[7] delves into the analysis and categorization of comments, aiming to discern whether a given comment falls within the toxic or non-toxic spectrum through the utilization of diverse machine learning techniques. Leveraging six distinct attributes and employing vectorization, a dictionary is constructed from an established vocabulary (Dataset) to train the ML model. Given the presence of multiple attributes, the model undergoes iterative training against each trait, facilitating the identification of the most proficient algorithm for discerning various types of toxicities. The study reveals the commendable performance of the Random Forest algorithm across all traits, yielding an accuracy of 85% with a precision of 91%. Initial investigations indicate a prevalent focus on demographic and local languages in prior research endeavors. In this context, the researchers sought to develop a classifier tailored for the English language.

K Nearest Neighbors: A machine learning algorithm used for classification and regression tasks, which makes predictions based on the k-nearest data points.

Random Forest: An ensemble learning method used for classification, regression, and other tasks, which operates by constructing a multitude of decision trees at training time.

Decision Tree: A decision support tool that uses a tree-like model of decisions and their possible consequences, widely used in classification tasks.

Naïve Bayes: A probabilistic machine learning algorithm based on the Bayes theorem, often used for text classification.

Logistic Regression: A statistical method used for analyzing a dataset with one or more independent variables, commonly employed in binary classification tasks.

This study[8] addresses this issue that has arisen from the advent of social media that has revolutionized both digitalization and communication. While it offers a global platform for expressing views and engaging with a wide audience, it has also become a breeding ground for

harmful behavior, including cyberbullying, offensive language, and personal attacks. The issue is addressed by assessing toxicity levels in Google's Jigsaw dataset, comprising six distinct classes: toxic, severe_toxic, obscene, threat, insult, and identity_hate. Employing multilabel classification, a single comment may be categorized under multiple classes. The impact of Multinomial Naive Bayes, Logistic Regression, and Support Vector Machine with TF-IDF on toxicity identification is examined. These models are trained on the provided dataset and subsequently tested. Results indicate that Logistic Regression outperforms other models in terms of accuracy and hamming loss.

Multinomial Naive Bayes is a probabilistic machine learning algorithm based on the Bayes theorem, often used for text classification. Logistic Regression is a statistical method used for analyzing a dataset with one or more independent variables, commonly employed in binary classification tasks. Support Vector Machine SVM is a supervised learning algorithm used for classification tasks, particularly effective for high-dimensional data. Hamming Loss is a metric used for evaluating the similarity between two sets by comparing their respective elements. In this context, it's used to assess the performance of the classification model.

The researchers of this paper[9] delve into the realm of contemporary online interactions, where users actively engage in discussions across diverse social platforms, news websites, and forums. Unfortunately, within this discourse, a subset of comments manifests as harmful or abusive. Recognizing the impracticality of manual oversight at such scale, numerous systems turn to machine learning models for the automated identification of such content. While extensive efforts have been dedicated to English, the classification of toxic comments in the Bangla language remains relatively underexplored. In response, this study pioneers a multi-label classification approach tailored to categorizing Bangla comments into six distinct classes: toxic, severe toxic, obscene, threat, insult, and identity hate, all while gauging their severity. For their experiment, the researchers meticulously compiled a dataset of 12,634 comments from a spectrum of sources, including the popular platform TikTok. Drawing from insights garnered from prior studies, the proposed classification model harnesses the power of deep learning techniques, particularly Long Short-Term Memory (LSTM) and Long Short-Term Memory-Convolutional Neural Networks (LSTM-CNNs), synergized with word embeddings. This sophisticated approach culminated in an impressive accuracy rate of 83.66% and a Mean Squared Error (MSE) loss of 1.61 for the severity measure.

The paper[10] analyzes toxic comments, denoting those remarks in online forums that are characterized by rudeness, offensiveness, or unfairness, often leading to the exit of numerous users from the conversation. The prevailing threat of bullying and abuse on the internet significantly hampers the open exchange of ideas, thereby constraining the expression of diverse viewpoints. This has prompted many websites to grapple with the dilemma of either severely restricting or entirely disabling user comments. The research endeavors to comprehensively explore the landscape of online abuse, categorizing it into distinct labels to enable a precise assessment of toxicity through the application of various machine learning algorithms. Each algorithm is tailored with a unique methodology to ensure optimal results. Notably, the neural-based model demonstrated superior performance over non-neural models, particularly in terms of F1 score and

accuracy. The study meticulously employs five distinct machine learning algorithms, namely Multinomial Naïve Bayes, Random Forest Classifier, Bernoulli Naïve Bayes, Nearest Centroid, and Ridge Classifier, to ensure the attainment of optimal results.

The researchers employed three distinct methods—Binary Relevance, Classifier Chain, and Label Power Set—across various machine learning algorithms. The results for each algorithm are presented, accompanied by the evaluation metrics of Log-Loss, Accuracy, and Hamming-Loss.

02 PROBLEM DEFINITION AND SCOPE

2.1 PROBLEM STATEMENT

In the contemporary digital landscape, online discussions have become a cornerstone of communication, providing a platform for individuals from diverse backgrounds to share their views, engage in dialogue, and form communities. These discussions, often hosted on social media platforms, forums, and comment sections, have transformed the way we exchange ideas, foster connections, and participate in public discourse. However, the unfiltered and open nature of online discussions has given rise to a pervasive and urgent issue - the proliferation of toxic comments

Toxic comments, which encompass a wide spectrum of harmful content, including hate speech, harassment, profanity, personal attacks, and various forms of online vitriol, have far-reaching consequences. They not only erode the emotional well-being of individuals but also hinder the promise of online platforms as inclusive and constructive spaces for interaction. Toxicity, when left unaddressed, poses a significant threat to the vitality of digital communities, fostering a culture of fear and exclusion.

The challenge at hand revolves around the pressing need to enhance the classification of comment toxicity in online discussions. The existing systems and tools for comment toxicity detection exhibit inherent limitations that demand careful consideration. Chief among these limitations is the lack of nuance in toxicity assessment. Current systems predominantly offer binary classification, categorizing comments as either toxic or non-toxic, a framework that often fails to capture the complex and multifaceted nature of online toxicity. Many comments exist along a gradient of toxicity, and classifying them as purely toxic or non-toxic does not do justice to the subtleties of human communication.

Moreover, users' autonomy and empowerment in shaping their online experience are currently constrained by a lack of customization options in comment filtering. To truly reflect the diverse preferences and sensitivities of users, an advanced system should allow individuals to set their own toxicity thresholds, enabling them to curate their digital environment to align with their personal standards.

Beyond these considerations, a robust comment toxicity classification system should take into account the contextual nuances of comments. Toxicity is not solely dependent on the words used but also on the context in which these words are employed. A sophisticated system should be equipped to evaluate the intent behind a comment and determine whether it is harmful or benign in its specific context.

Additionally, the online world is marked by a multitude of toxicity forms, each requiring distinct handling. Current systems often lack the ability to differentiate between these different types of toxicity, from hate speech and profanity to threats and personal attacks. It is imperative to advance

the classification process to discern the precise nature of toxicity, enabling more targeted and effective moderation.

Furthermore, user reputation and historical behavior play a pivotal role in determining the likelihood of posting toxic comments. A comprehensive system should integrate user reputation scores into the classification process, allowing for a tailored approach to comment assessment. Users who have a history of posting toxic content should be subject to different scrutiny than those who consistently engage positively.

One of the most pressing challenges that confronts any attempt to address online toxicity is the dynamic and ever-evolving nature of the problem. Toxicity, like language, adapts and changes over time, presenting new challenges and manifestations. Hence, the solution cannot be static but must be a dynamic model that continually learns and adapts. This adaptation requires an integration of user feedback, real-time data, and the capacity to recognize emerging forms of toxicity. Users should be active participants in this process, as their feedback is invaluable in fine-tuning the system and ensuring its relevance and effectiveness.

2.2 OBJECTIVE

Multi-Class Classification: The primary objective of this project is to develop a multi-class classification system capable of categorizing comments into multiple levels of toxicity. This classification includes distinguishing between mild, moderate, and severe toxicity, offering a more nuanced assessment of comments and their impact on online discourse

Threshold Customization: A key objective is to empower users to customize their online experience by setting their preferred toxicity thresholds. This allows individuals to tailor their comment filtering criteria to match their unique sensitivities and preferences, thereby promoting a more user-centric approach to content moderation.

Sentiment Analysis: An essential objective is the integration of sentiment analysis to differentiate between toxic comments, neutral comments, and positive comments. This not only considers the words used but also takes into account the overall tone and intent of the content, providing a more holistic view of comment sentiment

Content-Specific Classification: To address the diversity of toxicity forms in online discussions, another goal is the implementation of content-specific classification. This involves distinguishing between various types of toxicity, such as hate speech, profanity, threats, and personal attacks. This granular classification ensures that different forms of toxicity receive appropriate attention and moderation.

User Reputation Integration: Integrating user reputation scores into the classification process is a vital objective. This integration considers the historical behavior of users, which can be indicative

of their likelihood to post toxic comments. Users with a track record of posting harmful content should be subject to a different level of scrutiny.

Contextual Analysis: The project aims to incorporate contextual analysis to evaluate comments within their specific context. This analysis determines whether certain language or terms are used in a toxic or non-toxic manner, acknowledging that the same words can have different implications in distinct situations.

Historical User Behavior: Understanding the historical behavior of users is a critical objective. By analyzing a user's comment history and interactions, the system aims to predict the likelihood of a user posting toxic comments. This proactive approach enables targeted moderation and intervention.

Feedback Loop: Establishing a feedback loop is a core objective of the project. This loop enables users to actively participate in the system's improvement by reporting comments and providing feedback on comment classifications. User feedback is invaluable for continuous model enhancement.

Machine Learning Algorithms: The project aims to implement advanced machine learning algorithms, including neural networks and ensemble methods, for precise and robust toxicity level classification. These algorithms are central to achieving high accuracy in comment assessment

2.3 SCOPE

The scope of the project 'Enhancing Comment Toxicity Classification in Online Discussions' is thoughtfully defined to encompass a broad yet well-delineated range of considerations. It outlines the parameters within which the project operates, ensuring that the project's objectives are pursued effectively. Here are the key aspects of the project's scope:

Geographical and Linguistic Reach: This project is designed to cater to a diverse and global user base. It takes into account the linguistic and regional diversity of online discussions, thereby providing a solution that is accessible to a wide array of users.

Data Sources: The project leverages publicly available data from various online platforms, forums, and communities. It adheres to ethical data usage and content usage terms and conditions, ensuring that the data sources are legally accessible and compliant with privacy standards.

Regulatory Compliance: In an era of evolving regulations and legal standards, the project is committed to compliance. It operates within the framework of legal and regulatory requirements governing content moderation and user data protection. This entails ensuring that all project activities align with relevant legal and ethical standards.

Project Timeline: The project's timeline encompasses distinct phases, including development, testing, and potential future iterations. This timeline is subject to the constraints of time and resources, allowing for a well-structured and efficient execution.

Exclusions: It is essential to define what falls outside the scope of the project. This includes data that is inaccessible due to privacy concerns, legal restrictions, or proprietary constraints. The project also refrains from actions that would be deemed illegal, unethical, or contrary to the principles of responsible content moderation.

By defining these parameters, the project ensures that its efforts are focused and effectively aligned with the overarching objectives. The geographical and linguistic inclusiveness guarantees that the solution will be accessible to a wide user base, irrespective of their language or location. Regulatory compliance and ethical considerations underscore the importance of conducting project activities in a responsible and lawful manner. Additionally, the project's well-structured timeline ensures efficient execution, allowing for the attainment of the project's goals and objectives within the defined constraints. Importantly, the clear demarcation of exclusions ensures that the project adheres to legal and ethical standards, avoiding actions that could compromise the project's integrity or legality.

2.4 Methodologies of Problem Solving and Efficiency issues

- Root Cause Analysis: When issues arise in the authenticity of the food supply chain using blockchain, it's important to identify the root causes. This can involve investigating where and why discrepancies occur, whether they are related to data entry errors, tampering, or other factors.

- Data Validation and Verification: Implement data validation and verification processes at various points in the supply chain. This can include mechanisms to ensure that data entered into the blockchain is accurate and consistent with the physical products being tracked.

- Continuous Monitoring: Use real-time monitoring and analytics to track the performance of the blockchain system. Implement alerts and notifications for any discrepancies or anomalies in the data, allowing for swift responses.

- Quality Assurance Testing: Regularly conduct quality assurance testing to identify and rectify any issues within the blockchain system that may affect its accuracy and efficiency.

- Feedback Loops: Establish feedback loops with stakeholders, including producers, distributors, and consumers. Encourage them to report any issues they encounter with the blockchain system and use this feedback to make improvements.

- Regulatory Compliance Audits: Regularly audit the blockchain system to ensure it complies with relevant food safety and traceability regulations. Address any compliance issues promptly.

03 PROJECT PLAN

3.1 PROJECT ESTIMATION

ITEM	QTY	RATE(RS)	AMOUNT(Rs)
Laptop	**4**	**90,000**	**360,000**
Total	**4**		**360,000**

3.2 RISK MANAGEMENT

Risk Identification

User Input: The app prompts the user to input relevant information, the user can submit fake information or fake credentials and can manipulate the data stores present in block nodes.

Risk Calculation: The scalability of this Supply chain management is very low. If this platform scales in a larger number the data can become manipulative by the majority of the people. Risk Identification

3.3 PROJECT SCHEDULE

Project Task Set

T1 :- Planning the Adim, Distributor, User modules.

T2 :- Making the algorithm.

T3 :- Writing pseudo codes.

T4 :- Making functionality of each module.

T5:- Making the front end of the website. Developing user interface i.e web page using HTML.

T6 :- Creating a connection between the front-end and back-end.

T7 :- Connecting with back-end SQL server.

T8 :- Documentation and project report.

T9 :- Testing of the entire module.

04 REQUIREMENT SPECIFICATIONS

4.1 DATABASE

- **Database: MySQL**

  MySQL serves as a robust relational database management system, which is integral for data storage and retrieval in this project. It enables efficient organization and management of the datasets used for training and testing machine learning models. MySQL's scalability and performance make it a reliable choice for handling large volumes of data, ensuring smooth data operations throughout the project.

4.2 SOFTWARE

- **Operating System: Windows 10/11**

  Windows 10/11 provides a stable and widely used platform for software development. It offers seamless compatibility with a wide range of tools and libraries, making it an ideal choice for implementing and testing machine learning models. Additionally, it ensures that the developed solution is accessible to a broad audience, as Windows is one of the most prevalent operating systems globally.

- **IDE: Visual Studio Code**

  It is a versatile and lightweight integrated development environment. Its intuitive interface and extensive plugin ecosystem make it an excellent choice for this project. VS Code supports various programming languages, including Python, enabling efficient code development, debugging, and version control. It also provides a platform for easy collaboration via extensions, enhancing productivity throughout the development lifecycle.

- **Python Libraries: NumPy, pandas, scikit-learn, NLTK, and spaCy**

  - NumPy: Essential for numerical computations, optimizing operations for algorithms.
  - pandas: Streamlines data manipulation, simplifying tasks like cleaning, transformation, and analysis.
  - scikit-learn: Comprehensive library for machine learning, providing a wide array of algorithms and tools for model development and evaluation.

- NLTK and spaCy: Vital for text preprocessing, offering tools for tasks like tokenization, stemming, and advanced NLP operations.

- **Version Control: GitHub**

  GitHub serves as a powerful version control platform, allowing for seamless collaboration and tracking of changes throughout the project's development. It enables multiple contributors to work on different aspects of the project simultaneously, ensuring smooth integration of code and easy management of different versions.

4.3 HARDWARE

- **CPU: Minimum 4-core Processor (Intel/AMD)**

  The project necessitates a CPU with a minimum of four cores, either from Intel or AMD. This ensures that the system can efficiently handle the computational demands of machine learning tasks. With multiple cores, the CPU can execute multiple tasks concurrently, enhancing the performance of training and running models.

- **GPU: Minimum Integrated Graphics Card**

  While a discrete GPU is not mandatory, a system with at least an integrated graphics card is recommended. This aids in tasks that can benefit from GPU acceleration, such as certain deep learning operations. While it may not be as powerful as dedicated GPUs, an integrated card still contributes to smoother processing.

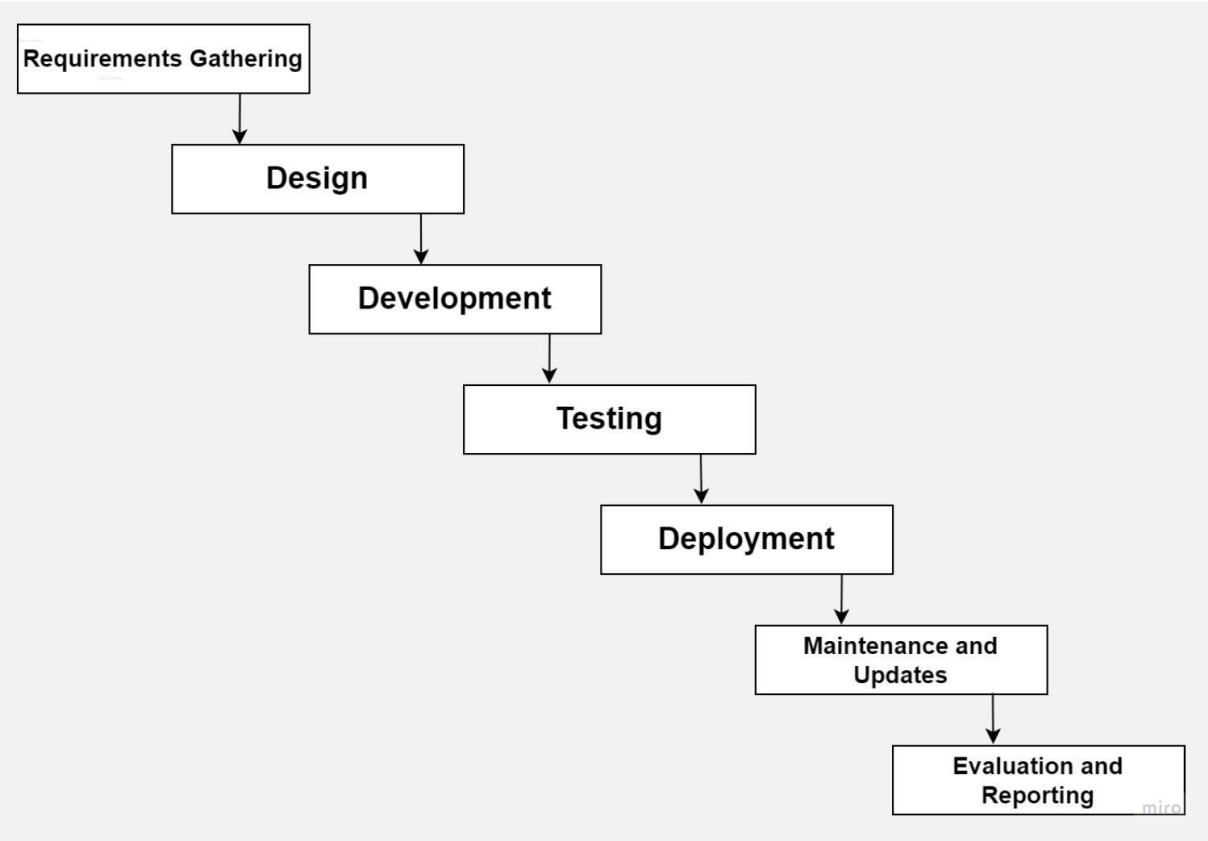
- **RAM: 8GB**

  With a minimum of 8GB RAM, the system can effectively manage the memory requirements of various processes. This ensures smooth execution of tasks, especially when dealing with large datasets or running complex algorithms.

- **Storage: 256GB HDD/SSD**

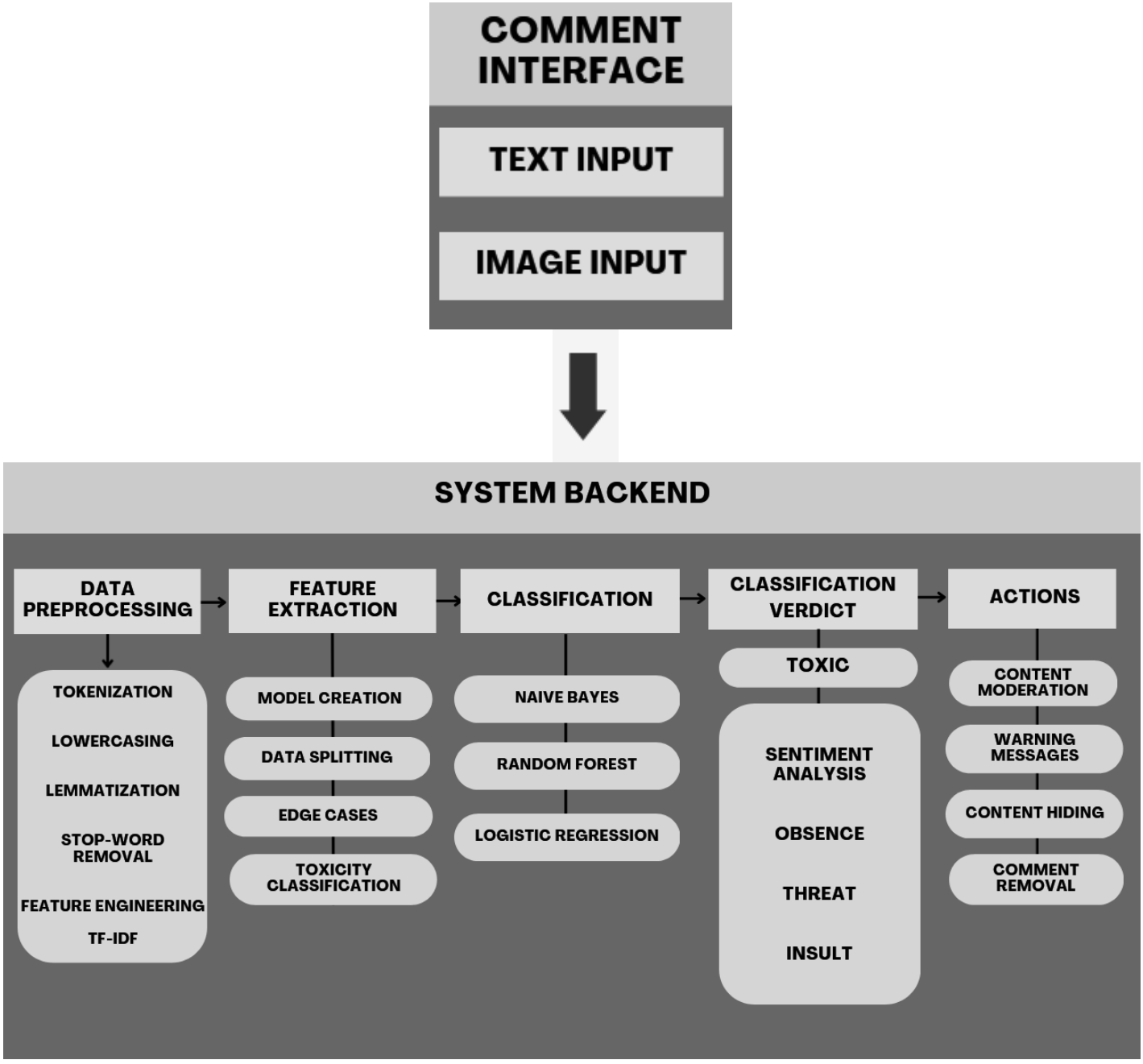
  A minimum of 256GB of storage, whether on an HDD or SSD, provides ample space for storing datasets, code, and any additional software or libraries required for the project. An SSD is preferred due to its faster read/write speeds, which can lead to quicker data access and model training times.

4.4 ANALYSIS MODEL - SDLC Model



05 SYSTEM DESIGN

5.1 SYSTEM ARCHITECTURE



5.2 MATHEMATICAL MODEL

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:

1. Hypothesis Function (Logistic Function):
   The hypothesis function $h_\theta(x)$ 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 θ^T is the transpose of the parameter vector θ

2. Cost Function (Logistic Loss):
   The logistic loss function is used to measure the error between the predicted probabilities and the actual labels:
   $$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(h_\theta(x_i)) + (1 - y_i)\log(1 - h_\theta(x_i))]$$
   This is a binary cross-entropy loss function.

3. 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 $\propto$ is the learning rate, and x_{ij} is the value of the j -th feature for the i-th comment.

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.

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

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

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

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.

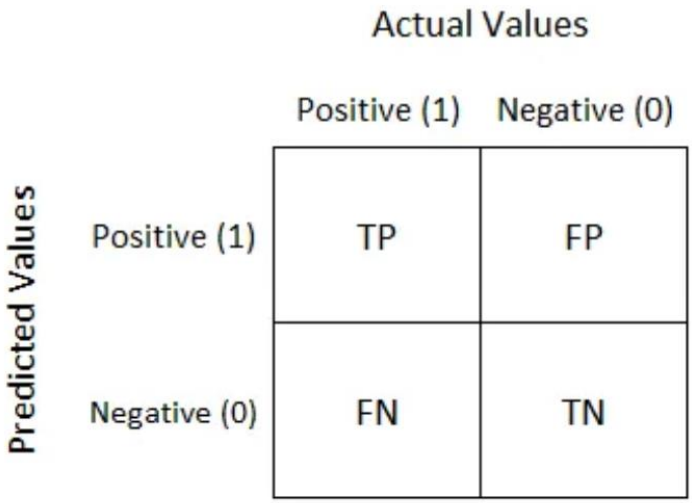
P(message content): This is the probability of observing the given content in any message, toxic or not. It acts as a normalization factor.

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

METRICS - Confusion Matrix

Used to evaluate performance of the algorithms.

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



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

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

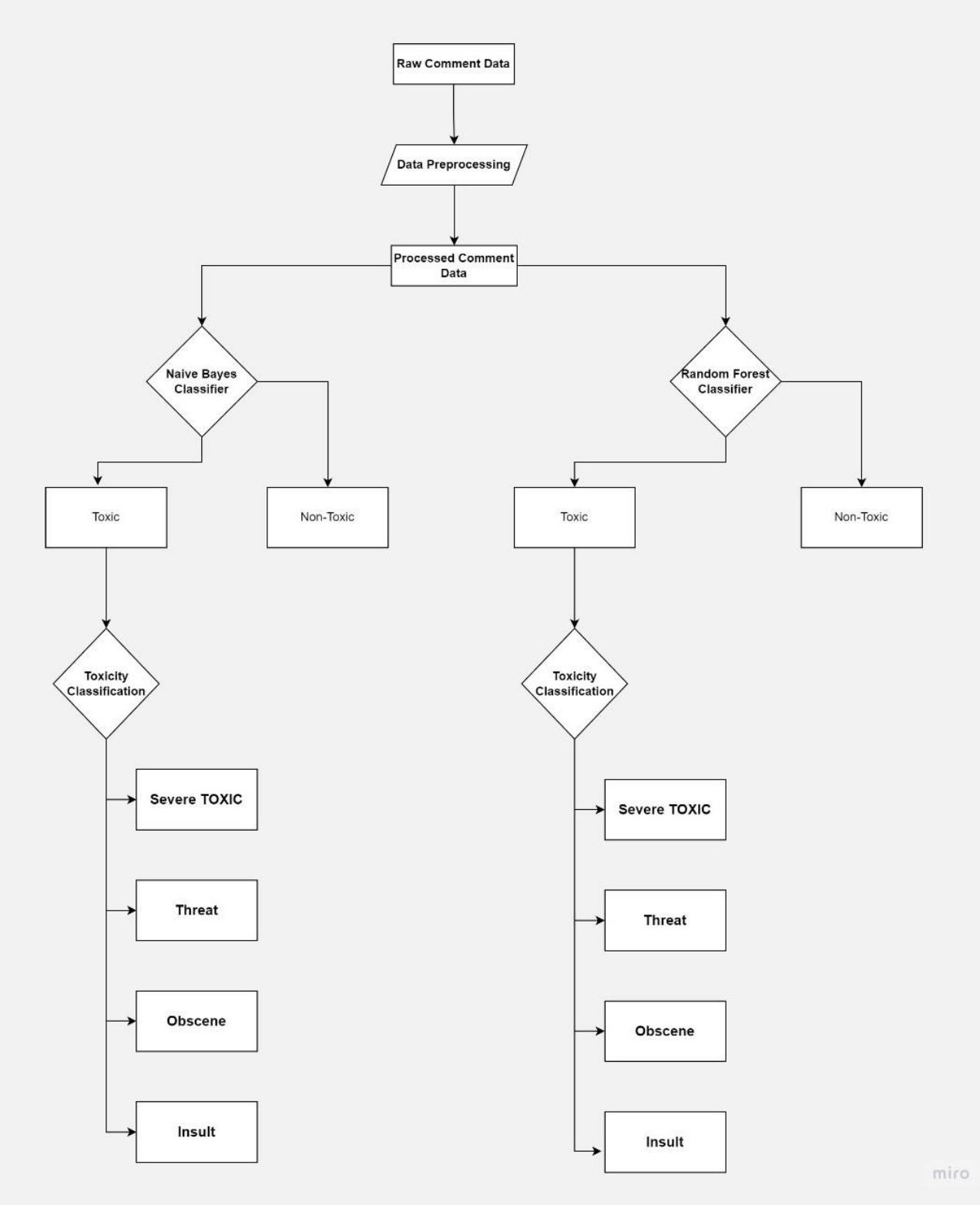
Precision (true positives / predicted positives) = TP / TP + FP

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

5.3 DATA FLOW DIAGRAM



The data flow diagram shows a comment classification system. The system takes raw comment data as input and outputs the toxicity classification of the comment as the output.

The system first preprocesses the raw comment data. This may involve steps such as tokenization, stopword removal, stemming, and lemmatization. The preprocessed data is then fed to two machine learning classifiers: a Naive Bayes classifier and a Random Forest classifier.

Each classifier outputs a toxicity score for the comment. The final toxicity classification of the comment is determined by combining the scores from the two classifiers. The system also outputs the severity of the toxicity, if the comment is classified as toxic.

Here is a more detailed description of each step in the data flow:

1. Raw Comment Data: This is the input to the system. It can be text data from any source, such as social media posts, online forums, or customer reviews.

2. Data Preprocessing: This step prepares the raw comment data for the machine learning classifiers. It may involve steps such as tokenization, stopword removal, stemming, and lemmatization.

3. Naive Bayes Classifier: This classifier uses Bayes' theorem to classify comments as toxic or non-toxic. It is a simple but effective classifier that is often used for text classification tasks.

4. Random Forest Classifier: This classifier uses an ensemble of decision trees to classify comments as toxic or non-toxic. It is a more complex classifier than the Naive Bayes classifier, but it can achieve higher accuracy.

5. Toxicity Classification: The final toxicity classification of the comment is determined by combining the scores from the Naive Bayes classifier and the Random Forest classifier.

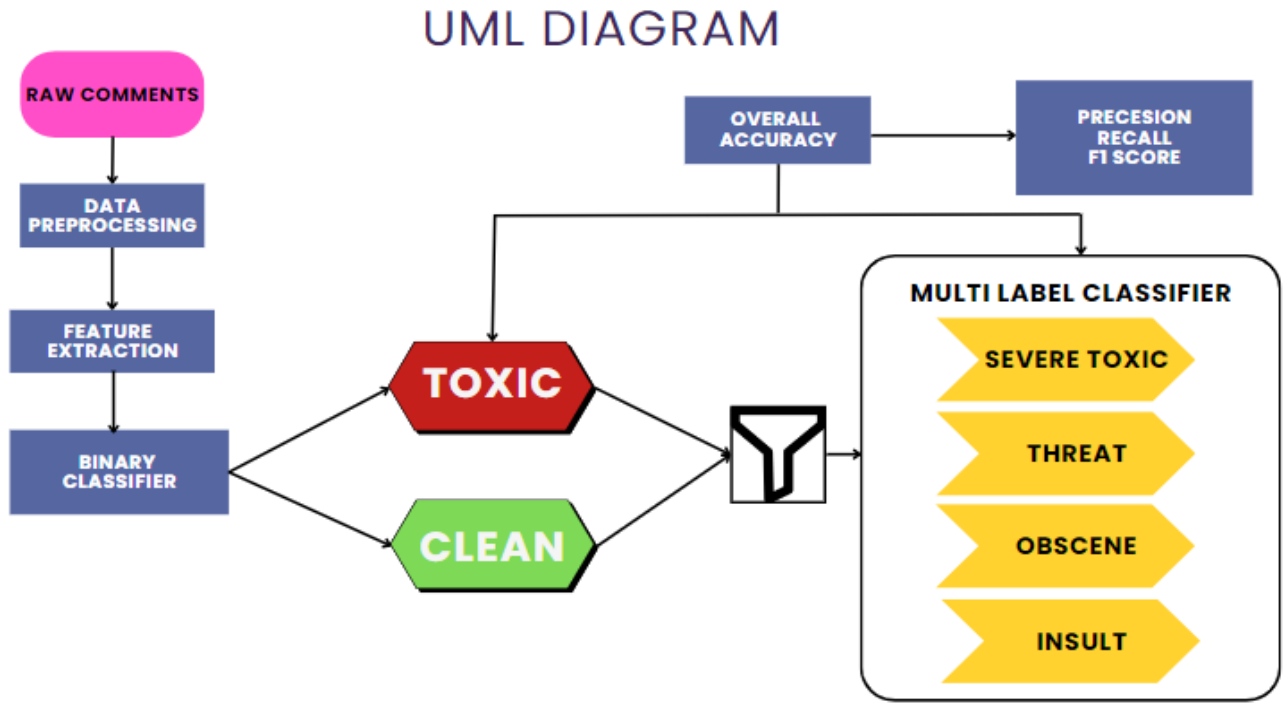
6. Severity of Toxicity: If the comment is classified as toxic, the system also outputs the severity of the toxicity. This may involve classifying the comment as obscene, threatening, or insulting.

The data flow diagram also shows the different types of toxicity that the system can classify. These include:

- Severe TOXIC
- Threat
- Obscene
- Insult

The system can be used to identify and remove toxic comments from online platforms. This can help to create a more positive and inclusive online environment.

5.5 UML DIAGRAM



The Unified Modeling Language (UML) diagram shows a multi-label classification system for classifying chemicals based on their toxicity level. The system consists of the following components:

- Data preprocessing: This component prepares the data for the multi-label classifier by cleaning and transforming the data.
- Feature extraction: This component extracts features from the preprocessed data that are relevant to the classification task.
- Multi-label classifier: This component takes the extracted features as input and predicts a set of labels for each chemical.
- Binary classifiers: The multi-label classifier is typically implemented using a set of binary classifiers, one classifier for each label. Each binary classifier predicts whether or not the chemical has a given label.

The UML diagram also shows the following performance metrics for the multi-label classification system:

- Overall accuracy: This metric measures the percentage of chemicals that are correctly classified by the system.
- Precision: This metric measures the percentage of predicted labels that are correct.

- Recall: This metric measures the percentage of actual labels that are correctly predicted.
- F1 score: This metric is a harmonic mean of precision and recall.

The UML diagram also shows the following comments about the multi-label classification system:

- RAW COMMENTS: These comments may contain additional information about the system, such as the specific machine learning algorithms used or the performance of the system on different datasets.
- OVERALL: This comment provides a summary of the system's performance on the overall dataset.
- PRECESION: This comment provides a summary of the system's precision for each label.
- RECALL: This comment provides a summary of the system's recall for each label.
- F1 SCORE: This comment provides a summary of the system's F1 score for each label.

Overall, the UML diagram provides a high-level overview of the multi-label classification system and its performance. The diagram can be used to understand the system's architecture, identify areas for improvement, and compare the system to other multi-label classification systems.

06 OTHER SPECIFICATIONS

6.1 ADVANTAGES

This project offers several advantages, which can be summarized in a point-by-point description:

Improving Online Safety:

One of the primary advantages of a toxic comment classifier is its role in making online spaces safer for users. By automatically detecting and flagging toxic comments, it helps in reducing the harm caused by online harassment and abuse.

Enhancing User Experience:

Users can have a more positive and enjoyable experience in online communities when they are protected from offensive or harmful content. The classifier helps in maintaining a healthy and respectful online environment.

Time and Resource Savings:

Content moderators and community managers often spend significant time manually reviewing and moderating comments. A toxic comment classifier can reduce the workload and free up human moderators to focus on more complex issues.

Scalability:

The classifier can handle a high volume of comments and scale with the growth of online communities. It ensures that all comments are screened for toxicity without the need for a corresponding increase in moderation staff.

Consistency:

Automation through the classifier ensures a consistent application of community guidelines and content policies. It eliminates the potential for human bias or oversight.

Real-time Detection:

Toxic comment classifiers are often implemented to provide real-time detection. This means harmful comments can be identified and addressed immediately, preventing them from causing further harm.

Data-driven Insights:

The data generated by the classifier can be analyzed to gain insights into the prevalence of toxic comments, the types of toxic behavior, and trends over time. This information can inform content policies and community management strategies.

Customization and Adaptability:

Many toxic comment classifiers can be fine-tuned and customized to suit the specific needs of different online platforms. This adaptability allows the classifier to be more effective in identifying context-specific toxicity.

User Trust and Retention:

A safer online environment encourages users to remain active in the community and trust the platform. Users are more likely to continue engaging with content and participating in discussions when they feel protected from toxic content.

Legal and Regulatory Compliance:

In some jurisdictions, there are legal requirements for online platforms to address and remove harmful content promptly. A toxic comment classifier helps in complying with these regulations and avoiding potential legal issues.

Reducing Liability:

By proactively identifying and addressing toxic comments, platforms can reduce their liability in cases where users are harmed due to such content. This can lead to fewer legal disputes and associated costs.

Positive Public Image:

Online platforms that effectively address toxic content and prioritize user safety can build a positive public image, which can lead to increased user acquisition and retention.

Crisis Management:

In the event of a crisis, such as a coordinated harassment campaign, a toxic comment classifier can swiftly detect and mitigate the impact by identifying and blocking toxic comments.

Feedback Loop for Users:

Some classifiers provide feedback to users whose comments are flagged as toxic. This can be an educational tool, encouraging users to reflect on their online behavior and potentially leading to more constructive interactions.

Reduced Stress for Moderators:

Content moderators often face psychological stress from exposure to toxic content. Using a classifier can reduce this stress by sparing moderators from the need to review highly disturbing content.

6.2 DISADVANTAGES

While toxic comment classifiers offer numerous benefits, they also come with certain disadvantages and limitations.

False Positives:

Toxic comment classifiers may incorrectly flag non-toxic comments as harmful, leading to the removal or moderation of innocent content. This can frustrate users and stifle free expression.

False Negatives:

Conversely, the classifier may miss some toxic comments, allowing harmful content to remain visible, which could harm users or damage the platform's reputation.

Difficulty in Context Understanding:

Toxicity can be highly context-dependent, and classifiers may struggle to understand the nuances and subtleties that determine whether a comment is genuinely harmful or not.

Sensitivity Settings:

Adjusting the sensitivity of the classifier to reduce false positives may result in more false negatives, and vice versa. Finding the right balance can be challenging.

Bias and Fairness Concerns:

Toxic comment classifiers can inherit biases from their training data, potentially leading to discriminatory outcomes or the unfair treatment of certain groups.

Constant Model Updates:

Toxicity evolves, and new forms of harmful language and behavior emerge over time. Constant model updates and maintenance are necessary to stay effective.

Resource Intensive:

Developing, training, and maintaining a toxic comment classifier can be resource-intensive, requiring significant computational power and expertise.

Algorithmic Opacity:

The inner workings of some classifiers, particularly deep learning models, can be complex and difficult to interpret, making it challenging to understand their decision-making processes.

Adversarial Attacks:

Malicious users can attempt to manipulate the classifier by employing techniques to evade detection, such as using subtle language or code words.

Inability to Address Novel Toxicity:

Toxic comment classifiers may not recognize entirely new forms of toxicity or hate speech, leaving users vulnerable to emerging threats.

Language and Cultural Sensitivity:

Toxic comment classifiers may struggle with understanding and identifying toxic comments in languages and cultures they were not specifically trained on, limiting their effectiveness in diverse contexts.

Impact on User Freedom:

Overzealous moderation can unintentionally limit freedom of speech and curb open discourse, leading to concerns about censorship and stifling of unpopular opinions.

Lack of Contextual Understanding:

Classifiers may misinterpret humor, sarcasm, or satire as toxic content, leading to unnecessary moderation actions and misunderstandings.

Resource Allocation:

Over Reliance on toxic comment classifiers might lead to reduced human moderation, which could be detrimental in situations where human judgment and context are essential.

Ethical Dilemmas:

Deciding what content should be considered toxic or harmful can pose ethical dilemmas and challenges in balancing freedom of speech with user safety.

Privacy Concerns:

The process of content moderation through toxic comment classifiers may involve analyzing user-generated data, raising privacy concerns if not handled properly.

User Backlash:

Users who feel their content is unfairly moderated or flagged as toxic may become dissatisfied and leave the platform, affecting user retention and engagement.

6.3 APPLICATION

Content Moderation:

Online platforms, such as social media websites, forums, and comment sections, use toxic comment classifiers to automatically identify and filter out harmful and offensive content, maintaining a safe and respectful online environment.

User Engagement and Retention:

Toxic comment classifiers help create a more positive user experience by reducing the presence of toxic comments. This, in turn, encourages user engagement and retention on online platforms.

Public Opinion Analysis:

Researchers and organizations can use toxic comment classifiers to analyze public sentiment and identify trends in toxic behavior on social media, enabling them to gain insights into public opinion.

Brand and Reputation Management:

Companies and brands can employ these classifiers to monitor social media and online discussions, identifying and addressing toxic comments that may harm their reputation and brand image.

Customer Support:

Toxic comment classifiers can assist in automating the detection of abusive language and harassment in customer support channels, allowing companies to provide a safer and more respectful customer service experience.

News and Media Outlets:

News websites and media outlets can utilize toxic comment classifiers to filter out offensive or inappropriate comments on articles, ensuring a more civil and constructive discussion.

Educational Platforms:

Toxic comment classifiers can be applied in educational settings to maintain a respectful and safe online learning environment, where students can engage in discussions without the fear of bullying or harassment.

Government and Regulatory Bodies:

Governments and regulatory bodies can employ these classifiers to monitor and enforce online safety laws, addressing hate speech and harmful content in public online spaces.

Crisis Management:

In times of crisis or emergencies, toxic comment classifiers can quickly identify and mitigate harmful content, helping to manage situations more effectively and reduce panic or misinformation.

Mental Health Support:

Mental health support platforms can use these classifiers to identify and flag potentially harmful comments and provide users with resources and assistance when they encounter toxic behavior.

Community Management:

Online communities, including gaming platforms and discussion forums, can use toxic comment classifiers to maintain a respectful and enjoyable atmosphere, thereby fostering a sense of belonging and community.

Custom Content Filtering:

Platforms can use toxic comment classifiers to allow users to customize their content filters, giving individuals control over the type of content they wish to see or avoid.

Research and Social Studies

Researchers can use toxic comment classifiers to study and analyze online behavior, online harassment, and hate speech as part of sociological and psychological research.

Legal Compliance:

Online platforms can use toxic comment classifiers to comply with legal requirements to remove or moderate toxic content, reducing legal liabilities.

Preventing Cyberbullying:

Toxic comment classifiers can be employed in apps and websites used by children and teenagers to prevent cyberbullying and ensure a safer online environment for young users.

Public Safety and Security:

Law enforcement agencies can use toxic comment classifiers to monitor online communications for threats and potential indicators of criminal activity.

AI Chatbots and Virtual Assistants:

AI chatbots and virtual assistants can use toxic comment classifiers to ensure that their interactions with users remain respectful and free from abusive language.

06 CONCLUSION AND FUTURE WORK

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:

Reducing False Positives and Negatives: Further research and development are needed to minimize false positives and false negatives in toxic comment classification. This involves improving the model's understanding of context and intent.

Bias Mitigation: Addressing algorithmic bias is critical. Ongoing efforts should focus on creating fairer and more equitable classifiers that do not discriminate against specific groups or viewpoints.

Multilingual and Multicultural Support: Enhancing the project's ability to handle different languages and cultures will be crucial to its global applicability.

Nuanced Classification: Developing classifiers that can recognize and categorize different types of toxicity (e.g., hate speech, cyberbullying, harassment) with higher granularity will provide more refined results.

Real-time Adaptation: The project should be able to adapt to rapidly changing online environments and emerging forms of toxic behavior through real-time updates and continuous learning.

Transparency and Explainability: Improving the transparency and explainability of classifier decisions will be vital for users to understand and trust the moderation process.

User Feedback Mechanisms: Implementing effective feedback mechanisms for users whose content is flagged as toxic can promote user education and improve classifier accuracy.

Privacy and Data Protection: Ensuring that the project adheres to strict privacy and data protection regulations while analyzing and processing user-generated data is paramount.

Customization Options: Providing users with the ability to customize the sensitivity and filtering settings of the classifier according to their preferences.

Education and Awareness: Collaborating with educational institutions and organizations to raise awareness about online toxicity and promote responsible online behavior.

Cross-platform Integration: Developing APIs and solutions that enable different online platforms to integrate the toxic comment classifier seamlessly.

Crisis Management Features: Enhancing the project's capabilities for managing and mitigating toxic content during online crises, such as viral misinformation or harassment campaigns.

Ethical Considerations: Conducting ongoing discussions and research on the ethical implications of content moderation, including the balance between free speech and user safety.

07 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