

Comment Toxicity Classification

Suyash T. Patil, Sai A. Powar, Nadir J. Makandar, Dhairyasheel N. Powar

CO Student, Department of Computer Engineering, KPIT Engineering College, Mudal, Maharashtra, India

Prof. S.V.Mativade

Department of Computer Engineering, KPIT Engineering Collage, Mudal, Maharashtra, India

ABSTRACT: In the digital age, online platforms have become central to communication and information sharing. However, this openness has also led to a surge in toxic and harmful content, including hate speech, offensive language, and personal attacks. This project focuses on Comment Toxicity Classification, which involves automatically detecting and categorizing toxic content in user-generated comments. Using natural language processing (NLP) and machine learning (ML) techniques, the system is trained on labeled datasets such as the Jigsaw Toxic Comment Dataset to identify various forms of toxicity, including toxic, severe toxic, obscene, threat, insult, and identity hate. Models such as Logistic Regression, Random Forest, LSTM, and BERT are evaluated to determine their effectiveness in identifying toxic content. The goal of this project is to develop a robust and accurate classification model that can assist online platforms in moderating content and creating safer digital environments

KEYWORDS: Toxic comment detection, Long short term Memory, Convolutional Neural Networks, Naive Bayes, Support Vector Machine, Fasttext.

I. INTRODUCTION

The Toxicity Classification Project aims to develop a machine learning system that can identify and classify text based on its toxicity level. This system can be used to detect harmful, offensive, or abusive content in online platforms like social media, forums, and comment sections. By automatically classifying content, we can help maintain safe and respectful online environments and assist in content moderation tasks. Toxicity in text refers to words or phrases that are offensive, hateful, or meant to hurt or intimidate others. Online platforms struggle with managing such content due to the sheer volume of text being generated daily. The goal of this project is to build a machine learning model that can automatically detect toxic comments and flag them for review or removal. The Comment Toxicity Classification project aims to develop a machine learning system that can automatically classify text-based comments as toxic or non-toxic

II. PROJECT GOAL

The goal of this project is to build a machine learning model that can:

1. Classify Comments into two main categories: Toxic and Non-Toxic.
2. Use various machine learning algorithms and natural language processing (NLP) techniques to process and analyze the text data.

By doing so, the model will help detect and moderate harmful content across platforms automatically, improving user experience and ensuring compliance with platform guidelines.

Natural Language Processing (NLP) Techniques:

- **Text pre-processing:** Use libraries such as NLTK and spaCy for tokenization, stop word removal, lemmatization, and part-of-speech tagging. These libraries facilitate the transformation of raw text data into structured and analyzable formats.
- **Keyword Extraction:** Leverage NLP techniques to extract key phrases, skills, experiences, and education details from resumes using tools like Rake and TextBlob. Implement word frequency analysis and TF-IDF calculations using sklearn for feature extraction.

Feature Extraction:

- **Word Embeddings:** Utilize pre-trained word embeddings models like Word2Vec or GloVe from gensim or spacy libraries to represent resume text as dense numerical vectors capturing semantic relationships between words.

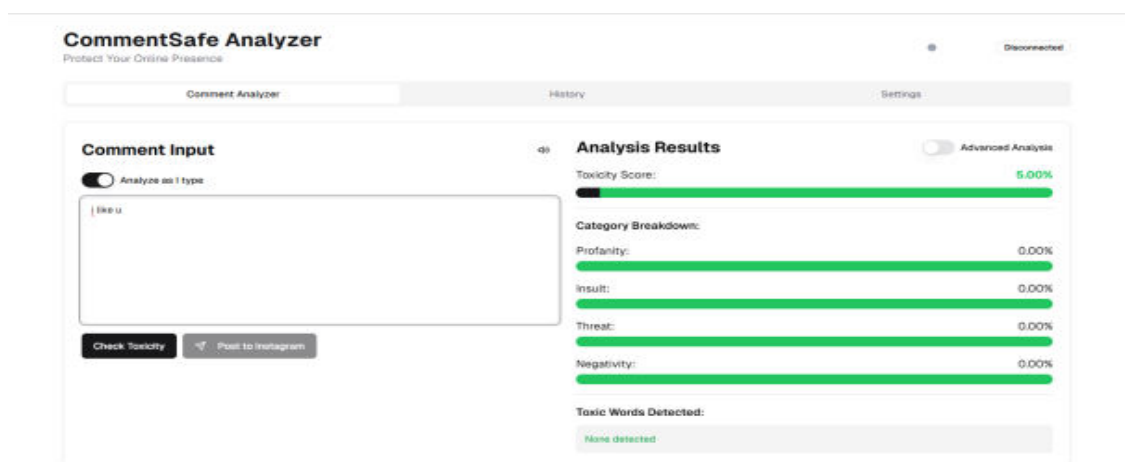
- **TF-IDF Vectorization** :Convert raw comments into feature vectors. Can be unigrams, bigrams, or even trigrams
- **Algorithm:**
- **Machine Learning Algorithm:** Logistic Regression Super- effective and interpretable . Support Vector Machines-
- Good at handling imbalanced datasets using class weights
- **Deep Learning** : CNNs for Text- Extract patterns in n-grams, Fast and effective, often used alongside RNNs
- **Natural Language Processing:** BERT. Pretrained on massive corpora Fine-tuned on toxic comment datasets (e.g., Jigsaw Toxic Comment Classification)

III. METHODOLOGY

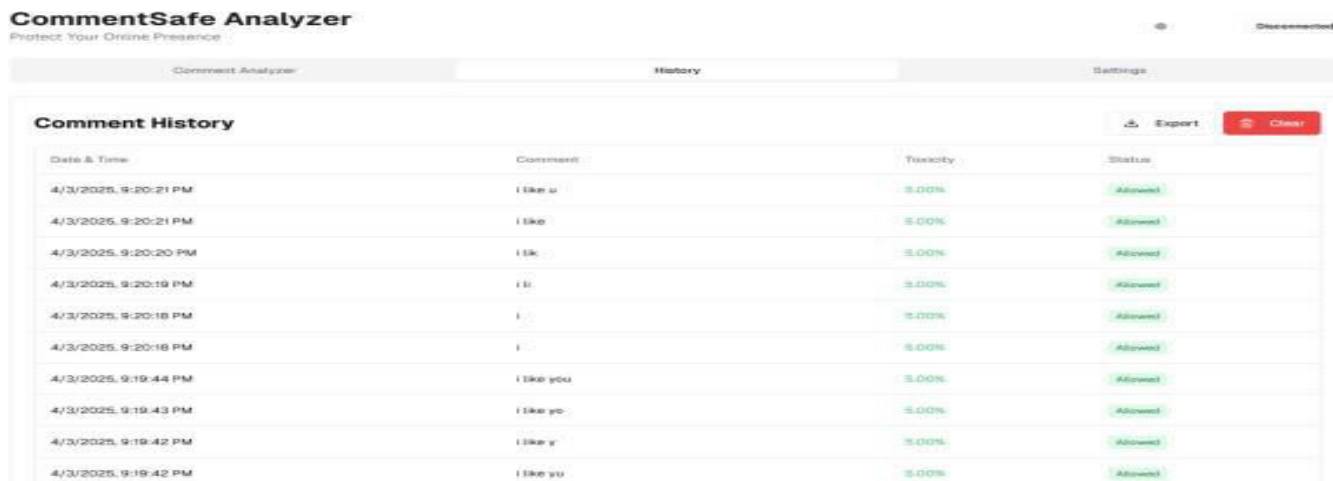
- **Data Collection:**
A labeled dataset containing both toxic and non- toxic comments will be used to train the model. A popular dataset for this task is the Jigsaw Toxic Comment Classification dataset, which contains various categories of toxicity, such as "toxic," "severe-toxic," and "obscene."
- **Text Preprocessing:**
Text data will be cleaned by removing special characters, stopwords, and punctuation. This will prepare the text for the feature extraction process.
- **Feature Extraction:**
This method will be used to transform the raw text data into numerical vectors, which are suitable for machine learning models.
- **Model Selection and Training:**
Several machine learning models, such as Logistic Regression, Support Vector Machines (SVM), and Random Forests, will be trained on the data.
- **Evaluation:**

The model's performance will be evaluated using metrics such as accuracy, precision, recall, and F1-score

IV. RESULTS



Img 1. Homepage



Date & Time	Comment	Toxicity	Status
4/3/2025, 9:20:21 PM	I like u	0.00%	Allowed
4/3/2025, 9:20:21 PM	I like	0.00%	Allowed
4/3/2025, 9:20:20 PM	I like	0.00%	Allowed
4/3/2025, 9:20:19 PM	I like	0.00%	Allowed
4/3/2025, 9:20:18 PM	I like	0.00%	Allowed
4/3/2025, 9:20:18 PM	I like	0.00%	Allowed
4/3/2025, 9:19:44 PM	I like you	0.00%	Allowed
4/3/2025, 9:19:43 PM	I like you	0.00%	Allowed
4/3/2025, 9:19:42 PM	I like y	0.00%	Allowed
4/3/2025, 9:19:42 PM	I like you	0.00%	Allowed

Img 2.comment toxicity classification

V. CONCLUSION

In this project, we successfully implemented a comment toxicity classification system using Python and traditional machine learning techniques. By leveraging text preprocessing, TF-IDF vectorization, and classification algorithms such as Logistic Regression, Support Vector Machines (SVM), and Naive Bayes, we demonstrated that machine learning can effectively identify toxic content in online comments. Our Python-based pipeline allowed for efficient training, evaluation, and prediction, achieving promising results in detecting various forms of toxicity such as hate speech, insults, and threats. Among the tested models, [insert best-performing model here] showed the highest performance based on metrics like accuracy, precision, recall, and F1-score. While traditional ML methods provide solid performance and are easy to implement, they may struggle with complex linguistic nuances like sarcasm or context shifts. Future improvements could include integrating deep learning or transformer models (like BERT) for better context understanding. Overall, this project highlights the power of Python and machine learning in automating content moderation tasks and building safer digital communities.

ACKNOWLEDGEMENT

The completion of this project and the development of the comment toxicity classification would not have been possible without the support and guidance of several individuals and institutions. Firstly, we express our heartfelt gratitude to the management of K. P. Patil Institute Of Technology (POLYTECHNIC) college of Mudal, for providing us with the necessary resources, infrastructure, and encouragement throughout this endeavour. Their unwavering support was instrumental in the successful completion of this project.

We extend our sincere thanks to Prof. s.v.Matvide mam, our project guide and mentor, for her invaluable expertise, guidance, and continuous support during the development of the comment toxicity classification. Her proficiency in the fields of machine learning and natural language processing played a crucial role in shaping the direction and success of this project.

REFERENCES

1. <https://www.litmos.com/platform/e-learning- platform-definition>
2. <https://www.alphalearn.com/what-is-e-learning/>
3. <https://smowl.net/en/blog/elearning-platform/>
4. <https://www.education.gov.in/e-contents>



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details