iJRASET
International Journal For Research in
Applied Science and Engineering Technology

Volume: 12 Issue: IV Month of publication: April 2024

DOI: https://doi.org/10.22214/ijraset.2024.61145

Advanced Social Media Toxic Comments Detection System Using AI

Dr D Nithya[1], Nanthine K S[2], Thenmozhi S[3], Varshinipriya R [4]
[1]Assistant Professor, [2, 3, 4]UG Students, Department of Computer Science and Engineering School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, Tamil Nadu, India

Abstract: In the contemporary digital landscape, social media platforms have evolved into essential components of our daily lives, fostering connections, idea sharing, and meaningful conversations. Nevertheless, the escalating volume of online interactions has ushered in a concerning rise in toxic comments and cyberbullying, casting a shadow over the potential for a healthy onlineenvironment. Toxic comments, spanning hate speech, harassment, and offensive content, not only inflict harm on individuals but also detrimentally affect the overall user experience. This project seeks to address this pressing issue through the development of a cutting-edge Toxic Comment Detection system, leveraging the power of Natural Language Processing (NLP) and Machine Learning (ML) techniques. The primary objective of this endeavor is to create an automated system capable of identifying and flagging toxic comments in realtime across various social media platforms. By employing advanced NLP algorithms and ML models, the system aims to analyze textual content swiftly and accurately, pinpointing instances of toxicity. Once identified, the system will promptly notify moderators, enabling swift intervention and potential removal of the harmful content. By implementing this technological solution, the project aspires to contribute significantly to fostering a safer and more inclusive online environment where users can engage without fear of encountering toxic behavior. Through the fusion of NLP and ML, this endeavor aims to exemplify the transformative potential of technology in mitigating the challenges posed by toxic comments in the digital age.

Keywords: Natural Language Processing, Machine Learning, K-nearest neighbour.

I. INTRODUCTION

In the contemporary era, the omnipresence of digital communication and social media platforms has redefined the way individuals connect, share ideas, and engage in conversations. This digital revolution has undoubtedly brought about a myriad of benefits, fostering global connectivity and democratizing information dissemination.However, as the volume of online interactions continues to surge, an insidious challenge has emerged: the proliferation of toxic comments and cyberbullying. This dark underbelly of the digital age threatens not only individual well-being but also the overall health of online communities. Recognizing the gravity of this issue, this project embarks on the development of a Toxic Comment Detection system, leveraging the power of Natural Language Processing (NLP) and Machine Learning (ML) techniques. The advent of social media platforms has ushered in an era where individuals across the globe can seamlessly connect, share experiences, and participate in diverse conversations.

Platforms like Facebook, Twitter, Instagram, and others have become integral to daily life, providing spaces for social interaction, information dissemination, and community building. The positive aspects of these platforms are undeniable, from facilitating long-distance relationships to serving as catalysts for social and political movements. Toxic comments encompass a spectrum of harmful behaviors, including hate speech, harassment, and the dissemination of offensive content. This phenomenon not only poses a direct threat to the mental and emotional well- being of individuals but also has broader implications for the overall user experience on these platforms. As individuals face a barrage of toxic comments, the quality of online discourse deteriorates, hindering the platforms' original mission to facilitate positive interactions. This project is centered on the development of a Toxic Comment Detection system that harnesses the capabilities of NLP and ML. The primary goal is to empower social media platforms with the ability to automatically identify and flag toxic comments, enabling timely intervention by moderators. The system is designed not only to enhance the well-being of individual users but also to cultivate a healthier online environment conducive to positive interactions.

II. LITERATURE SURVEY

Rabia Rachidi Et all [1] proposed that cyberbullying takes its place in social media and has increased throughout the past few years.
The damage that cyberbullying has on the users is undeniable they get attacked either on their appearances, ethnicities, religions, and even their thoughts and personal opinion.

This paper presents a cyberbullying detection system in the Moroccan dialect on an Instagram-collected dataset. The experiment results gave accuracies of around 77% to 91% from both the ML and DL algorithms. The LSTM model gave the best outcome by 91.24% outperforming the ML models. Mohammed Taleb Et all [2] proposed that the measures we used to evaluate our methods are accuracy, recall, and Fl-score. Our experiments showed that deep learning models performed unquestionably in the task of detecting toxic comments. The LSTM models with the Globe representation and LSTM with Fast Text were able to produce a higher F1 and accuracy compared to the other models used. For Toxic spans detection, the higher scores were obtained when combining LIME with classifier LSTM(GloVe) with an accuracy of 98% to identify the toxic spans.

Krishna Dubey Et all [3] proposed that the model that we have developed is able to classify given comments as toxic ornontoxic with 94.49% precision, 92.79% recall and 94.94% Accuracy score.

Nayan Banik Et all [4] proposed as our research finding, we demonstrate that both the deep learning-based models have outperformed other classifiers by 10% margin where Convolutional Neural Network achieved the highest accuracyof 95.30%.

KGSSV Akhil Kumar Et all [5] proposed that the following study uses 6 different traits, with the help of u25a1 vectorization a dictionary will be created out of known vocabulary (Dataset) to train the ML model. Since Multiple Traits are present the ML model has to get trained multiple times against each trait. It was identified that the Random Forest algorithm performed well against all types of traits which gave us a good accuracy of 85% with a precision of 91%.

Varun Mishra Et all [6] proposed as the results of the experiments show that the suggested model performs where we are presenting the fresh modelling CNN approach to detect the toxicity of textual content present on the social media platforms and we categorized the toxicity into positive and negative impact on our society.

Ali Salehgohari Et all [7] proposed that the toxic comment dataset is utilized in this research to train a deep learning model that categorizes comments into the following categories: severe toxic, toxic, threat, obscene, insult, and identity hatred. To categorize comments, use a bidirectional long short-term memory cell (Bi-LSTM).

Nitin Kumar Singh Et all [8] proposed that the impact of Multinomial NB, Logistic Regression, and Support Vector Machine with TF-IDF on identifying toxicity in text. These models were trained using the training data and after training were tested on the test data provided in the dataset. Experimental results show that Logistic Regression trumps the other models in terms of accuracy and hamming loss.

Felix Museng Et all [9] proposed to achieve this, we have compiled the datasets of research papers and analyse the algorithm used. The findings indicate that Long Term Short Memory is the most routinely mentioned deep learning model with 8 out of26 research papers. LSTM has also repeatedly yielded high accuracy results with above 79% for around 9000 data which could be adjusted depending on the pre- processing method used. There have been attempts to combine more than one deep learning algorithms, however these hybrid models might not result in a better accuracy thanan original model. Furthermore, the most frequent sources of datasets came from Kaggle and Wikipedia datasets and a total of 13 researchers that used Wikipedia's talk page edits as their dataset.

Nayan Banik Et all [10] proposed that in this scholarly manuscript, we provide a comparative analysis of five supervised learning models (Naive Bayes, Support Vector Machines, Logistic Regression, Convolutional Neural Network, and Long Short-Term Memory) to detect toxic Bengali comments from an annotated publicly available dataset. As our research finding, we demonstrate that both thedeep learning-based models have outperformed other classifiers by 10% margin where Convolutional Neural Network achieved the highest accuracy of 95.30%.

III. PROPOSED METHODOLOGY

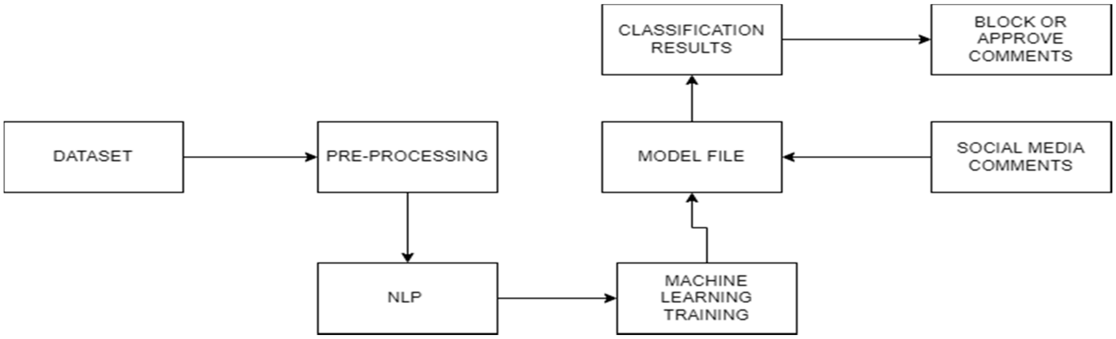


Fig 3.1 System Model Block Diagram

A. Data Collection and Preprocessing

The first phase involves gathering labeled social media comments, a crucial step for training and evaluating the machine learning model. Datasets containing comments with labels indicating toxicity are collected. Following this, the text data undergoes thorough preprocessing to ensure optimal model performance. This includes tasks such as removing irrelevant characters, handling misspellings, and addressing abbreviations to create a clean and standardized dataset.

Preprocessing is a crucial stage in your project's workflow, where raw text data undergoes several transformations to make it suitable for analysis. These steps typically include tokenization, where text is divided into individual words or tokens, followed by lowercasing to standardize text case. Punctuation removal and stop word removal help filter out unnecessary noise from the text. Stemming or lemmatization may also be applied to reduce words to their base forms, further simplifying the text representation. Through preprocessing, your system prepares the data for classification, ensuring that machine learning algorithms can effectively discern toxic comments from benign ones.

B. NLP (Natural Language Processing) in our Project

NLP is the backbone of your project, enabling your system tounderstand, interpret, and respond to human language. It involves various techniques and algorithms designed to process and analyze text data. In your project, NLP algorithms preprocess raw text data, transforming it into a format suitable for machine learning models. This preprocessing involves tasks like tokenization, removing stop words, and converting text into numerical representations. Additionally, NLP facilitates tasks such as sentiment analysis, topic modeling, and named entity recognition, enhancing your system's ability to detect toxic commentsaccurately.

In our project, NLP plays a crucial role in understanding and analyzing the textual data present in online comments. NLP techniques allow your system to comprehend the nuances of language, detecting patterns indicative of toxic behavior such as hate speech or harassment. Through various NLP processes, your system can preprocess and interpret text, enabling it to effectively identify and flag potentially harmfulcomments in real-time.

C. Feature Engineering and ML Model (K-Nearest Neighbors - KNN)

Feature engineering is essential to transform the text data intoa format suitable for machine learning. Text-based features such as word frequency, TF-IDF scores, or word embeddings are extracted. In the case of K-Nearest Neighbors (KNN), the algorithm relies on the similarity between data points, makingTF-IDF scores a suitable choice for feature representation. KNN is a non-parametric algorithm that classifies a data point based on the majority class of its k-nearest neighbors in the feature space.

Training the KNN model involves feeding it the labeled dataset, and the algorithm 'learns' the patterns in the feature space associated with toxic and non-toxic comments. The choice of the parameter 'k,' representing the number of neighbors to consider, is crucial and should be determined through model validation.

D. Real-time Analysis and Notification

To enable real-time analysis of our project, the trained KNN model is implemented into a system that continuously monitors incoming comments. As new comments are posted, their features are extracted, and the KNN model predicts their toxicity. A notification system is implemented to alert moderators or administrators when a comment surpasses a certain toxicity threshold. It ensures timely intervention in addressing potentially harmful content.

E. Comment Management

The final phase involves the incorporation of mechanisms forcomment management based on platform policies. Dependingon the severity of toxicity predicted by the KNN model, actions such as comment removal or hiding may be taken. The system's performance is continuously evaluated, and improvements are made through iterations. Regular model updates and retraining on new data help adapt to evolving patterns of toxic behavior in online comments. The methodology employs K-Nearest Neighbors as the machine learning algorithm for toxic comment detection, leveraging features extracted through preprocessing and feature engineering. The real-time analysis and comment management components enhance the system's effectiveness in creating a safer online environment.

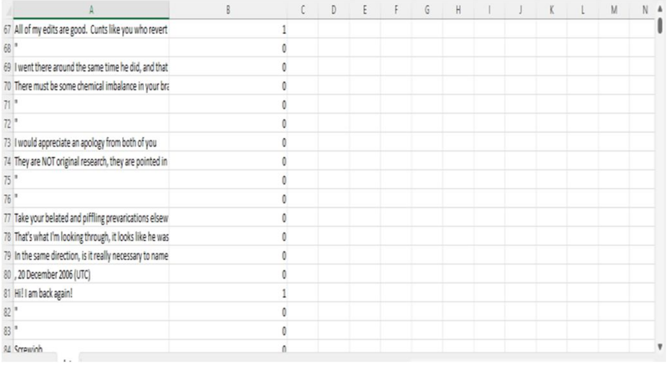
How KNN (K-Nearest Neighbors) Works in our Project: KNN is a machine learning algorithm used for classification tasks, including toxic comment detection in your project.KNN works by classifying unseen data points based on themajority class of their nearest neighbors in the feature space. In your context, after preprocessing the comments, the system represents each comment as a feature vector.

When a new comment is received, KNN calculates its distance to the existing comments in the feature space. The comment is then classified based on the class of its nearest neighbors, where the "k" nearest neighbors opinions are considered. This approach allows your system to make real-time predictions regarding the toxicity of incoming comments, aiding in prompt intervention and moderation to maintain a positive online environment.

F. KNN Steps

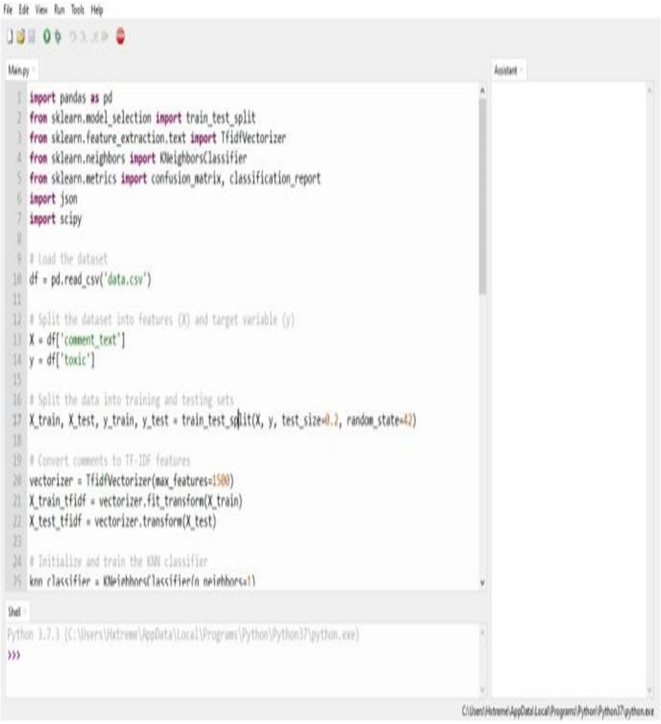
KNN (K-Nearest Neighbors) is a machine learning algorithm employed in your project to classify comments based on their similarity to existing data points. In KNN, each comment is represented as a point in a multi-dimensional space, with its position determined by its feature values. When a new comment arrives, the system calculates its distance to the existing comments and selects the "k" nearest neighbors. The class labels of these neighbors are then used to determine the classification of the new comment. This approach allows your system to make real-time decisions regarding the toxicity of incoming comments, enabling timely intervention and moderation to maintain a healthy online environment.

IV. RESULTS



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
67	All of my edits are good. Cunts like you who revert	1												
68	"	0												
69	I went there around the same time he did, and that	0												
70	There must be some chemical imbalance in your bra	0												
71	"	0												
72	"	0												
73	I would appreciate an apology from both of you	0												
74	They are NOT original research, they are pointed in	0												
75	"	0												
76	"	0												
77	Take your belated and piffling prevarications elsew	0												
78	That's what I'm looking through, it looks like he was	0												
79	In the same direction, is it really necessary to name	0												
80	, 20 December 2006 (UTC)	0												
81	Hi! I am back again!	1												
82	"	0												
83	"	0												
84	Screwjob	0												

Fig 4.1 Dataset (0-Normal,1-Toxic)

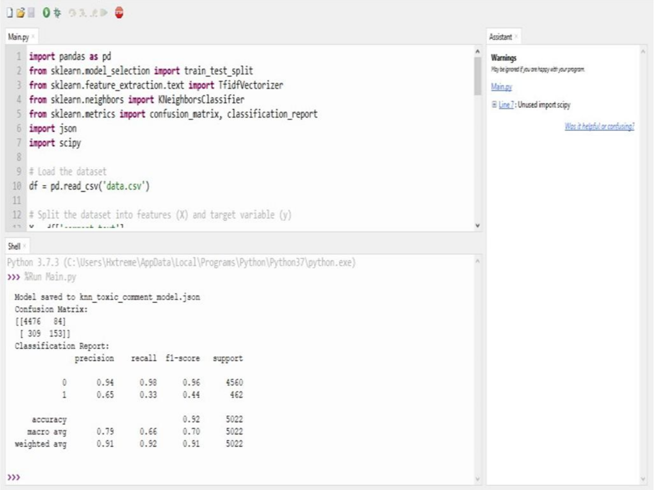


```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import confusion_matrix, classification_report
import json
import scipy

# Load the dataset
df = pd.read_csv('data.csv')

# Split the dataset into features (X) and target variable (y)
X = df['comment_text']
y = df['toxic']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Convert comments to TF-IDF features
vectorizer = TfidfVectorizer(max_features=1500)
X_train_tfidf = vectorizer.fit_transform(X_train)
X_test_tfidf = vectorizer.transform(X_test)

# Initialize and train the KNN classifier
knn_classifier = KNeighborsClassifier(n_neighbors=1)
```

```
Python 3.7.3 (C:\Users\Hxtreme\AppData\Local\Programs\Python\Python37\python.exe)
>>>
```

```
C:\Users\Hxtreme\AppData\Local\Programs\Python\Python37\python.exe
```

Fig 4.2 KNN Toxic Comment Classification Code

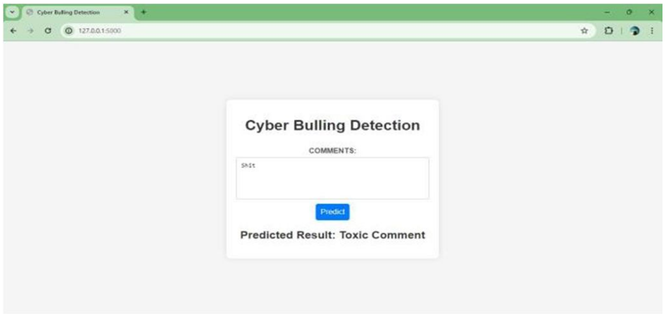


```
1  import pandas as pd
2  from sklearn.model_selection import train_test_split
3  from sklearn.feature_extraction.text import TfidfVectorizer
4  from sklearn.neighbors import KNeighborsClassifier
5  from sklearn.metrics import confusion_matrix, classification_report
6  import json
7  import scipy
8
9  # Load the dataset
10 df = pd.read_csv('data.csv')
11
12 # Split the dataset into features (X) and target variable (y)
```

```
Python 3.7.3 (C:\Users\Hxtreme\AppData\Local\Programs\Python\Python37\python.exe)
>>> %Run Main.py

 Model saved to knn_toxic_comment_model.json
 Confusion Matrix:
 [[4476   84]
  [ 309  153]]
 Classification Report:
               precision    recall  f1-score   support

           0       0.94      0.98      0.96      4560
           1       0.65      0.33      0.44       462

    accuracy                           0.92      5022
   macro avg       0.79      0.66      0.70      5022
weighted avg       0.91      0.92      0.91      5022


>>>
```

Warnings
May be ignored if you are happy with your program.
Main.py
Line 7 : Unused import scipy
Was it helpful or confusing?

Fig 4.3 KNN Evaluation



Cyber Bulling Detection

COMMENTS:

Shit

Predicted Result: Toxic Comment

Fig 4.4 Comment Prediction

V. CONCLUSION

In conclusion, this project represents a crucial step toward mitigating the pervasive issue of toxic comments on social media platforms. By leveraging Natural Language Processing and Machine Learning, the developed Toxic Comment Detection system showcases the potential for automated identification and flagging of harmful content in real time. The significance of this endeavor lies not only in enhancing individual user experiences but also in contributing to the broader goal of fostering a healthier online environment. As digital communication continues to evolve, the imperative to address the darker aspects of online discourse becomes more pronounced. This project underscores the viability of technological solutions in curbing the impact of toxic comments, paving the way for more positive and inclusive digital interactions.

VI. FUTURE SCOPE

Looking ahead, the future scope of this project extends to refining and expanding the capabilities of the Toxic Comment Detection system. Continued research could explore the integration of advanced deep-learning techniques for an even more nuanced identification of toxic behavior. Additionally, collaboration with social media platforms could facilitate the real-world implementation and testing of the system on a larger scale. Further developments may also involve addressing the ethical considerations surrounding content moderation and the delicate balance between free speech and the need for a safe online environment. The project sets the foundation for ongoing efforts to combat toxic behavior in the dynamic landscape of digital communication.

REFERENCES

[1] Rabia Rachidi,Mohamed Amine Ouassil,Mouaad Errami,Hassan Silkan,"Social Medias Toxic Comments Detection Using Artificial Intelligence Techniques",2023 3rd International Conference on Innovative Research in Applied Science Engineering and Technology

[2] Mohammed Taleb,Alami Hamza,Mohamed Zouitni,Nabil Burmani,Said Lafkiar,Noureddine En-Nahnahi,"Detection of toxicity in social media based on Natural Language Processing methods",2022 International Conference on Intelligent Systems and Computer Vision

[3] Krishna Dubey,Rahul Nair,Mohd. Usman Khan,Prof. Sanober Shaikh,"Toxic Comment Detection using LSTM",2020 Third International Conference on Advances in Electronics Computers and Communications

[4] Nayan Banik,Md. Hasan Hafizur Rahman,"Toxicity Detection on Bengali Social Media Comments using Supervised Models",2019 2nd International Conference on Innovation in Engineering and Technology

[5] KGSSV Akhil Kumar,B. Kanisha,"Analysis of Multiple Toxicities Using ML Algorithms to Detect Toxic Comments",2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering

[6] Varun Mishra,Monika Tripathi,"Detecting Toxic Comments Using Convolutional Neural Network Approach",2022 14th International Conference on Computational Intelligence and Communication Networks

[7] Ali Salehgohari,Mina Mirhosseini,Hamed Tabrizchi,Annamaria Varkonyi Koczy,"Abusive Language Detection on Social Media using Bidirectional Long-Short Term Memory",2022 IEEE 26th International Conference on Intelligent Engineering Systems

[8] Nitin Kumar Singh,Satish Chand,"Machine Learning- based Multilabel Toxic Comment Classification",2022 International Conference on Computing Communication and Intelligent Systems

[9] Felix Museng,Adelia Jessica,Nicole Wijaya,Anderies Anderies,Irene Anindaputri Iswanto,"Systematic Literature Review: Toxic Comment Classification",2022 IEEE 7th International Conference on Information Technology and Digital Applications

[10] Nayan Banik,Md. Hasan Hafizur Rahman,"Toxicity Detection on Bengali Social Media Comments using Supervised Models",2019 2nd International Conference on Innovation in Engineering and Technology

[11] D. Nithya and S. Sivakumari, “Categorizing online news articles using Penguin search optimization algorithm”, International Journal of Engineering and Technology, Vol. 7, pp. 2265-2268, 2018.

[12] D. Nithya and S. Sivakumari, “Fuzzy Based Latent Dirichlet Allocation in Spatio-Temporal and Randomized Least Angle Regression”, In International Conference on Sustainable Communication Networks and Application ,pp. 497-508, 2019.