iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.58929

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Comment Toxicity Tracker Using NLP with Emphasis on Machine Learning Algorithms

Ms. Shivani Kadam[1], Ms. Komal Ghatage[2], Mr. Aadesh Chaugule[3], Mr. Shubham Dilip Gajarushi[4], Prof. J.W.Bakal[5]
[1, 2, 3, 4]*Student,* [5]*Professor, IT Dept, Pillai HOC College of Engineering and Technology, Mumbai, India*

Abstract: ***The rise of online stages has driven to an uncommon volume of user-generated content, including comments on various forums, social media posts, and news articles. However, this abundance of user comments has also brought to light the issue of toxicity, where certain comments contain harmful, offensive, or inflammatory language that can negatively impact online discussions and communities. To address this issue, investigate centers on the advancement of a comment harmfulness location show utilizing Normal Dialect Handling (NLP) & Machine Learning. The proposed system leverages state-of-the-art NLP. By training these models on labelled datasets of toxic and non-toxic comments, the system learns to identify patterns and linguistic cues associated with toxic language. Key components of the system preprocessing steps to clean and tokenize the comments.Feature extraction using word embeddings or contextual embeddings, and model training using Machine Learning algorithms like neural networks, Random Forest Classifier model, etc . Evaluation measurements such as exactness, exactness, review, and F1-score are utilized to survey the execution of the prepared model.***

Keywords: Sentiment analysis, Text classification, Machine learning, Random Forest Classifier, Data Mining, Comment Toxicity.

I. INTRODUCTION

The exponential growth of online communication platforms has revolutionized the way people interact and share information. User-generated content, particularly within the shape of comments on gatherings, social media stages, and news articles, has become integral to online discourse. However, this surge in user engagement has also brought to light the pervasive issue of toxic comments, which can range from offensive language to hate speech, posing significant challenges for maintaining a healthy online environment.

Traditional methods of comment moderation, such as manual review by human moderators, are increasingly unable to keep up with the sheer volume of user comments.As a result, there is a growing need for automated systems that can efficiently and accurately identify toxic comments, thereby enabling platforms to take proactive measures to mitigate their impact. By leveraging NLP techniques, such as machine learning calculations and Machine Learning models, it is conceivable to train systems to recognize patterns and linguistic cues associated with toxic comments, enabling them to effectively differentiate between toxic and non-toxic content. The proliferation of user-generated content on online stages has driven to an uncommon volume of comments,, posing significant challenges for content moderation. Manual moderation of such a large volume of comments is not only time consuming but also prone to human error and bias. Moreover, the subjective nature of toxicity makes it troublesome to create a one size-fits-all approach for recognizing and taking care of harmful comments.In response to these challenges, researchers and developers have turned to NLP as a potential solution. By applying NLP to the task of comment toxicity detection, it becomes possible to automate the process of identifying toxic comments, thereby relieving the burden on human moderators and enabling platforms to more effectively manage their content.

II. RELATED WORK

Social media, blogs, and online news stages these days permit any web client to share his or her supposition on self-assertive substance with a wide gathering of people. The media commerce and writers adjusted to this advancement by presenting comment segments on their news stages. With increasingly political campaigning or indeed disturbance being disseminated over the Web, genuine and secure stages to talk about political themes and news in common are progressively imperative. Readers' and writers' inspirations for the utilization of news comments have been subject to investigate . Writer's inspirations are exceptionally heterogeneous and run from communicating an conclusion, inquiring questions, and rectifying real blunders, to deception with the expectation to see the response of the community. Agreeing to a study among U.S. American news commenters, the larger part (56 percent) need to specific an feeling or supposition..

©IJRASET: All Rights are Reserved

A. *Classes of Toxicity*

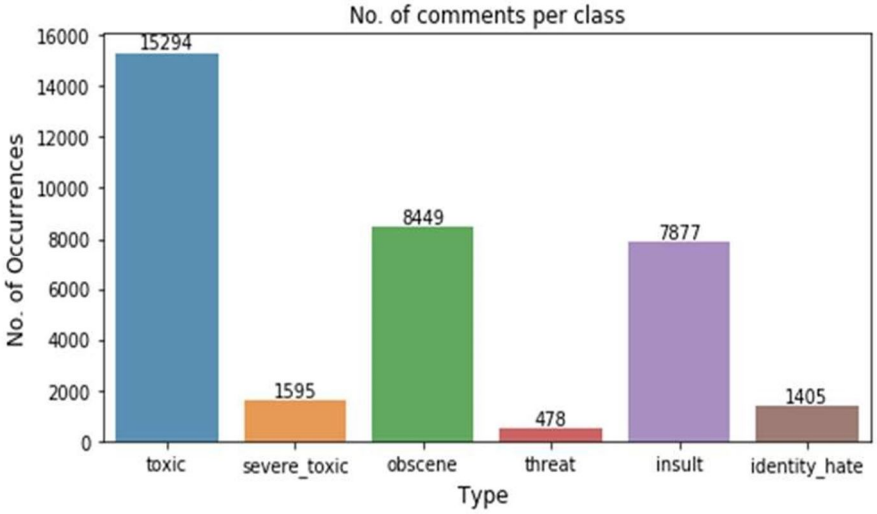
Harmfulness comes in numerous distinctive shapes and shapes. For this reason, a classification plot for harmful comments has advanced, which is propelled by explanations given in numerous datasets as depicted in Area. [2] They discover that models prepared on master comments significantly outperform models prepared on laymen comments. Within the taking after, we talk about one such classification conspire comprising of five diverse harmfulness classes. We appear cases for the distinctive classes of poisonous comments for outline.

B. *Obscene Language/Profanity*

Case: "That rule is bullshit and ought to be ignored.". The primary lesson considers swear or revile words. Within the illustration, the single word "bullshit" comprises the toxicity of this comment.[2] Typical for this class, there's no have to be under consideration the total comment in case at slightest one debase word has been found. For this reason, basic boycotts of debase words can be utilized for location. To counter boycotts, noxious clients regularly utilize varieties or incorrect spellings of such words.

C. *Insults*

Case: "Do you know you come over as a mammoth prick?". Whereas the past lesson of comments does not incorporate articulations around people or bunches, the course "insults" does. "Insults" contain inconsiderate or hostile articulations that concern an person or a gather. Within the illustration, the comment specifically addresses another client, which is common but not necessary.



D. *Threats*

Case: "I will orchestrate to have your life terminated.". In online dialogs, a common danger is to have another user's account closed.[3] Extremely harmful comments are dangers against the life of another client or the user's family. Articulations that declare or advocate for dispensing discipline, torment, damage, or harm on oneself or others drop into this course.

E. *Hate Speech/Identity Hate*

Case : "Mate, sounds like you are Jewish. Gayness is in the air". In differentiate to insuperable, character despise points only at bunches characterized by religion, sexual introduction, ethnicity, sex, or other social identifiers. Negative traits are credited to the bunch as in the event that these qualities were universally valid. For illustration, bigot, homophobic, and sexist comments drop into the category of character despise.

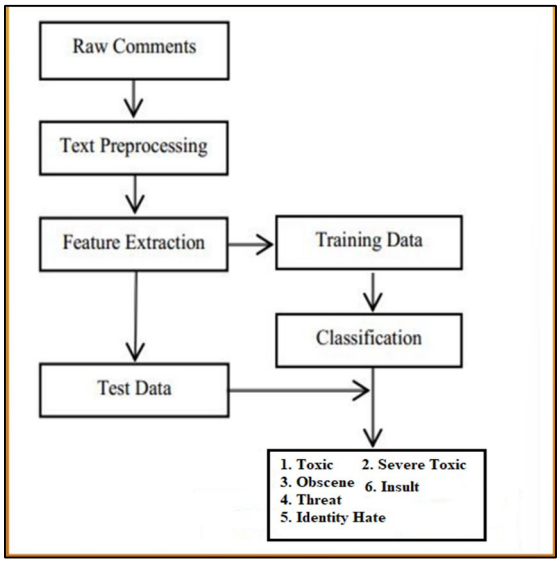
F. *Otherwise Toxic*

Example: "Bye! Don't look, come or think of coming back!". Comments that don't drop into one of the past four classes but are likely to form other clients take off a discourse are considered "toxic" without assist detail. Trolling, for illustration, by posting off-topic comments to aggravate the discourse falls into this class.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

III. RESEARCH METHODOLOGY

The code appears to follow a research methodology commonly utilized in common lingo planning (NLP) and substance classification assignments.. Here's an outline of the research methodology that we have followed:

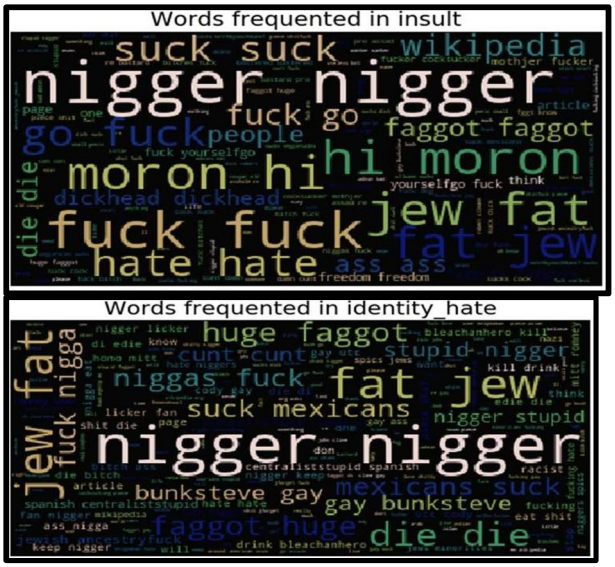


A. Problem Definition

Objective: The essential objective of this investigate is to create Machine Learning show able of classifying content comments into diverse poisonous categories. Each comment can have a place to numerous categories, such as poisonous, serious harmful, indecent, risk, offended, or personality abhor. [4] The demonstrate point to recognize and categorize poisonous substance inside the comments automatically.

B. Data Collection

The dataset is obtained from a CSV file named 'train.csv'.[1] This file likely contains columns such as 'id', 'comment text', and columns representing binary labels for different toxic categories. Each row in the dataset corresponds to a comment, and the labels indicate whether the comment belongs to specific toxic categories.



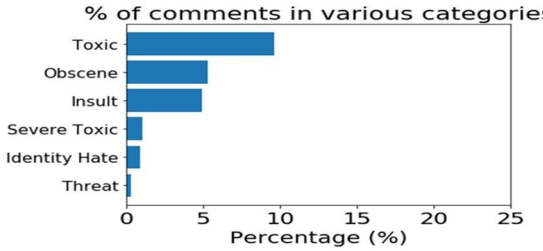
©IJRASET: All Rights are Reserved

C. Data Exploration:

In the data exploration phase, the research begins by gaining a comprehensive understanding of the dataset. Firstly, the data.info() and data.head() functions are employed to obtain basic information about the dataset, such as the data types, non-null counts, and the initial rows.[11] This aids in grasping the structure of the dataset and identifying any potential issues, such as missing values or data types that need adjustment. Subsequently, descriptive statistics are generated to explore the distribution of comments across different toxic categories. Bar charts are created to visually represent the prevalence of each category, offering insights into the imbalance or balance within the dataset.[5] Additionally, a sample of comments is inspected to gain qualitative insights into the language and content of the comments .The combination of quantitative and qualitative analyses at this stage forms a foundational understanding essential for effective feature engineering and model development in later stages of the research.

D. Data Preprocessing

1) *Text Cleaning:* Firstly, the comments undergo text cleaning processes such as tokenization and lowercasing, ensuring consistency and ease of analysis. Punctuation is also removed to streamline the text.
2) *Stop word Removal:* Subsequently, common stopwords, which are words with minimal semantic meaning like 'and' or 'the', are eliminated from the comments to focus on more meaningful content.



3) *Vectorization:* Following the preprocessing steps, the comments are subjected to vectorization using the TF-IDF (Term Frequency-Inverse Record Recurrence) procedure. This change changes over the content information into numerical features.
4) *Exploratory Data Analysis (EDA):* Moving on to the Exploratory Data Analysis (EDA) stage (Step 5), visualizations, including bar charts and word clouds, are employed to delve deeper into the characteristics and distribution of comments across different toxic categories. Bar charts provide a quantitative overview of the prevalence of each category, while word clouds visually highlight the most frequent words, offering qualitative insights into the content. This combination of preprocessing and EDA lays the groundwork for subsequent model training and evaluation, ensuring the input data is appropriately processed and understood before deploying machine learning algorithms.

E. Model Selection:

In the model comparison stage, the F1 scores obtained from evaluating different machine learning models are visualized to facilitate a comparative analysis of their performanceF1 score may be a that equalizations exactness and review, making it a suitable measure for tasks with imbalanced datasets, such as toxic comment classification. this helps us to select and finalize the required model for our comment toxicity model.

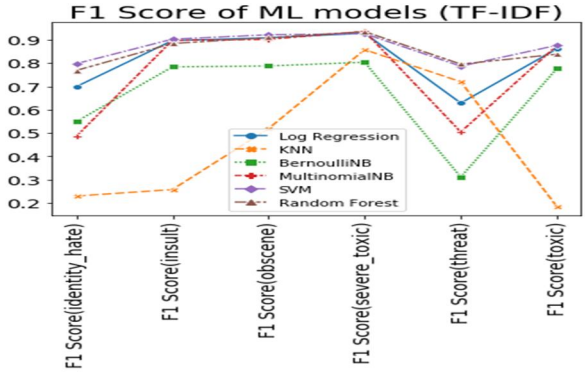


Fig. Model Comparison and Selection

We choose Random Forest instead of Linear SVM although the latter performs well, as RDF has predicting probability function and LinearSVM does not.

F. Model Interpretation:

In the model interpretation stage, the results obtained from model training and evaluation is analyzed to derive meaningful insights into the effectiveness of different models within the setting of poisonous comment classification. This involves interpreting key performance metrics, such as F1 scores, precision, and recall, for model.

IV. PROBLEM STATEMENT AND CHALLENGE

A. Problem statement

Online platforms face a significant challenge in managing toxic comments. These comments can be harmful, offensive, or abusive, leading to negative experiences for users and potentially damaging the platform's reputation. Manual moderation of comments is time consuming and may not be feasible when dealing with a large volume of user-generated content. Therefore, the problem is to develop an automated system that can classify comments as toxic or non-toxic, enabling platforms to take appropriate actions to maintain a safe and respectful online environment.

B. Challenge

Building a comment toxicity classification model presents several challenges. Firstly, the subjectivity and ambiguity inherent in language make it difficult to precisely identify toxic comments, as they may involve sarcasm or cultural references. Data bias is another issue, where a lack of diversity in training data can result in a skewed model that struggles with certain comment types or demographics. The rapid evolution of online language poses a challenge, requiring continuous updates to keep the model relevant. Handling multimodal data, including text, images, and videos, introduces additional complexities. Imbalanced datasets, where toxic comments are a minority, can lead to poor generalization.

C. Applications and Uses :

A comment toxicity detection system serves a crucial role in various online platforms to maintain a healthy and respectful online environment. Here are some applications and uses for such a system:

Social Media Platforms:Content Moderation: Automatically identify the toxic comments, hate speech, and harassment, maintaining a healthier online community.

User Safety: Enhance user safety by identifying and addressing potentially harmful comments, reducing the risk of cyber bullying.

E-commerce Platforms:Customer Reviews Moderation: Identify and filter out toxic reviews, ensuring that the product review section remains informative and trustworthy.

Enhancing Customer Experience: Promote a positive online shopping experience by preventing the dissemination of offensive content in product reviews and discussions.

Educational Platforms:Student Safety: Protect students from cyberbullying and inappropriate content in online discussions and forums.

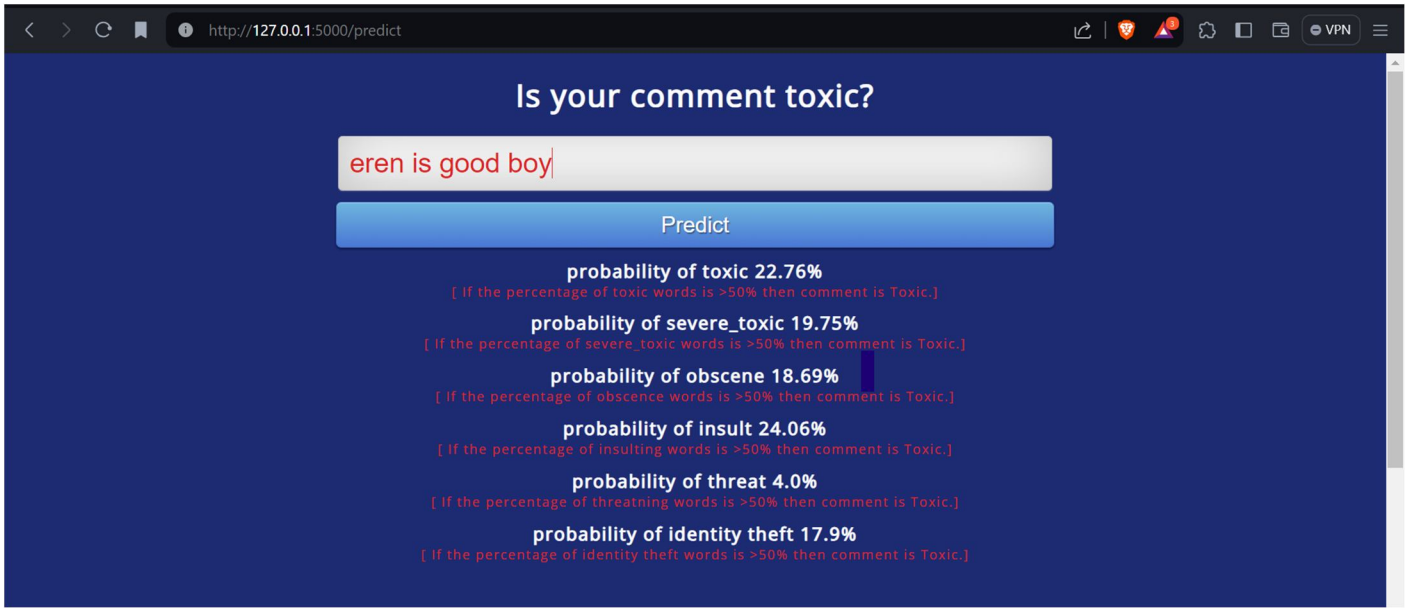
V. MODEL EVALUATION & RESULT

The result of our analysis and work shows the proper evaluation of all the toxic comments according to the categories. this algorithm or model we can say that gives the accurate results if we look the percentages of each category. The train and test data that we used is fully satisfied with all the comment categories like toxic and non-toxic and a mixture of both classes.

	F1 Score (toxic)	F1 Score (severe_toxic)	F1 Score (obscene)	F1 Score (insult)	F1 Score (threat)	F1 Score (identity_hate)
Log Regression	0.861234	0.92787	0.908655	0.896599	0.6288	0.6990
KNN	0.185120	0.857	0.5190	0.257992	0.7200	0.2301
BernoulliNB	0.776521	0.80372	0.787830	0.7837	0.3188	0.5492
MultinomialNB	0.874958	0.936170	0.901463	0.8974	0.5047	0.4858
SVM	0.876133	0.92600	0.921378	0.9062	0.7867	0.7975
Random Forest	0.838055	0.934874	0.909091	0.8839	0.7955	0.7684

The Random Forest Classifier model also performs well, indicating that it may be a good choice for simpler classification tasks. Accuracy can be increased by adding more dense layers in the neural network model and performing bagging. this model shows maximum accuracy than the four models that we discussed.

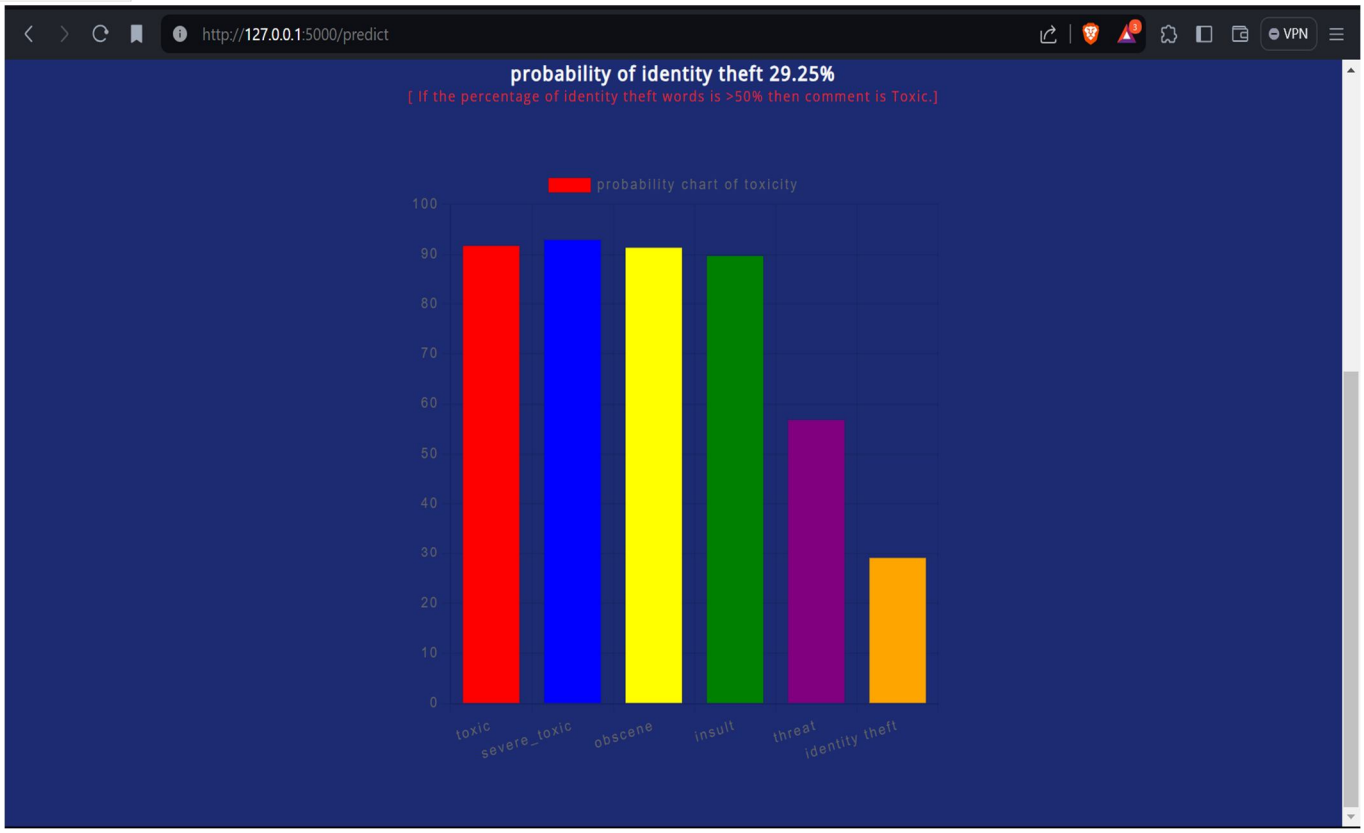
VI. RESULTS



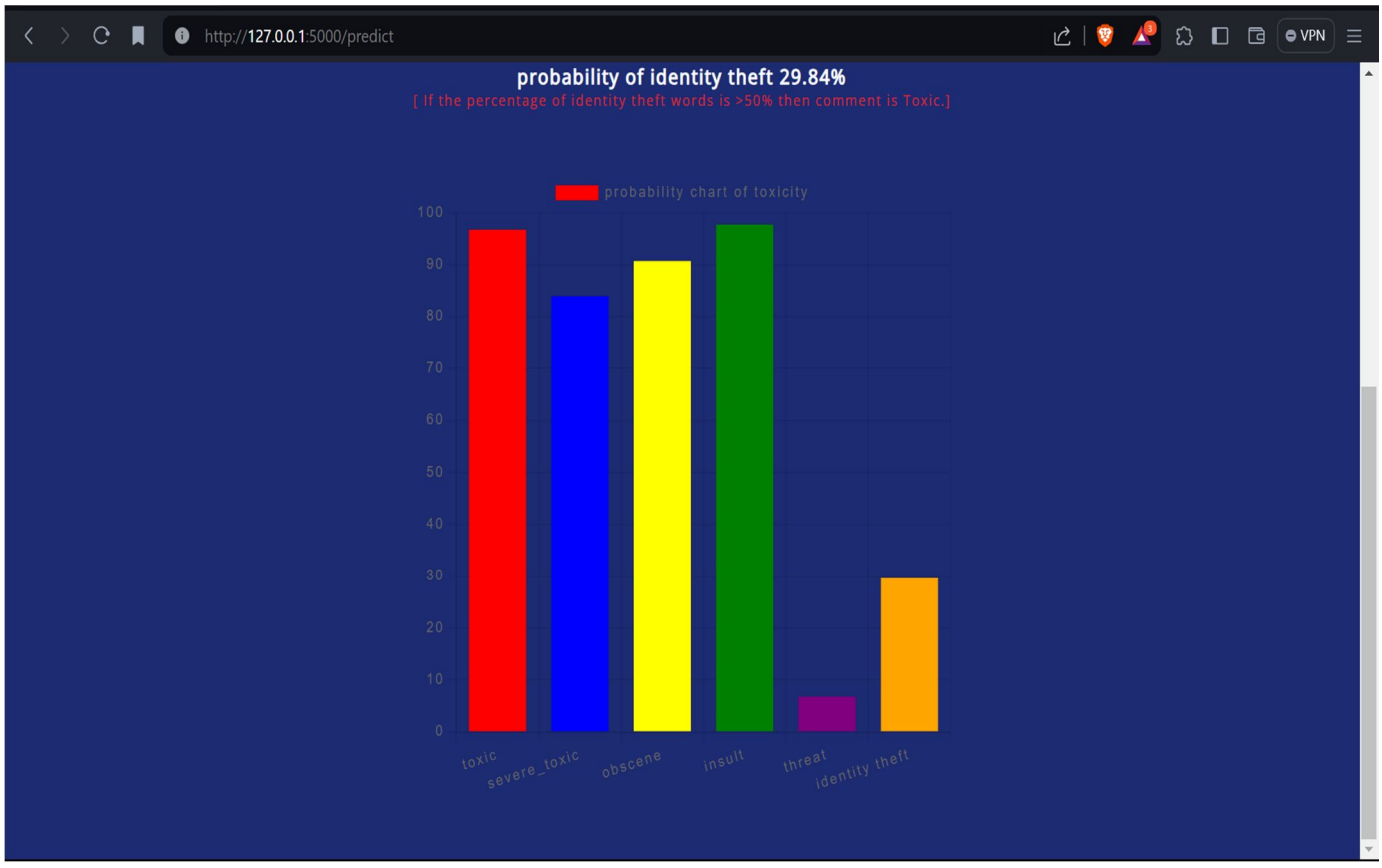
PREDICTING NON-TOXIC COMMENT



PREDICTING SEVERE-TOXIC COMMENT

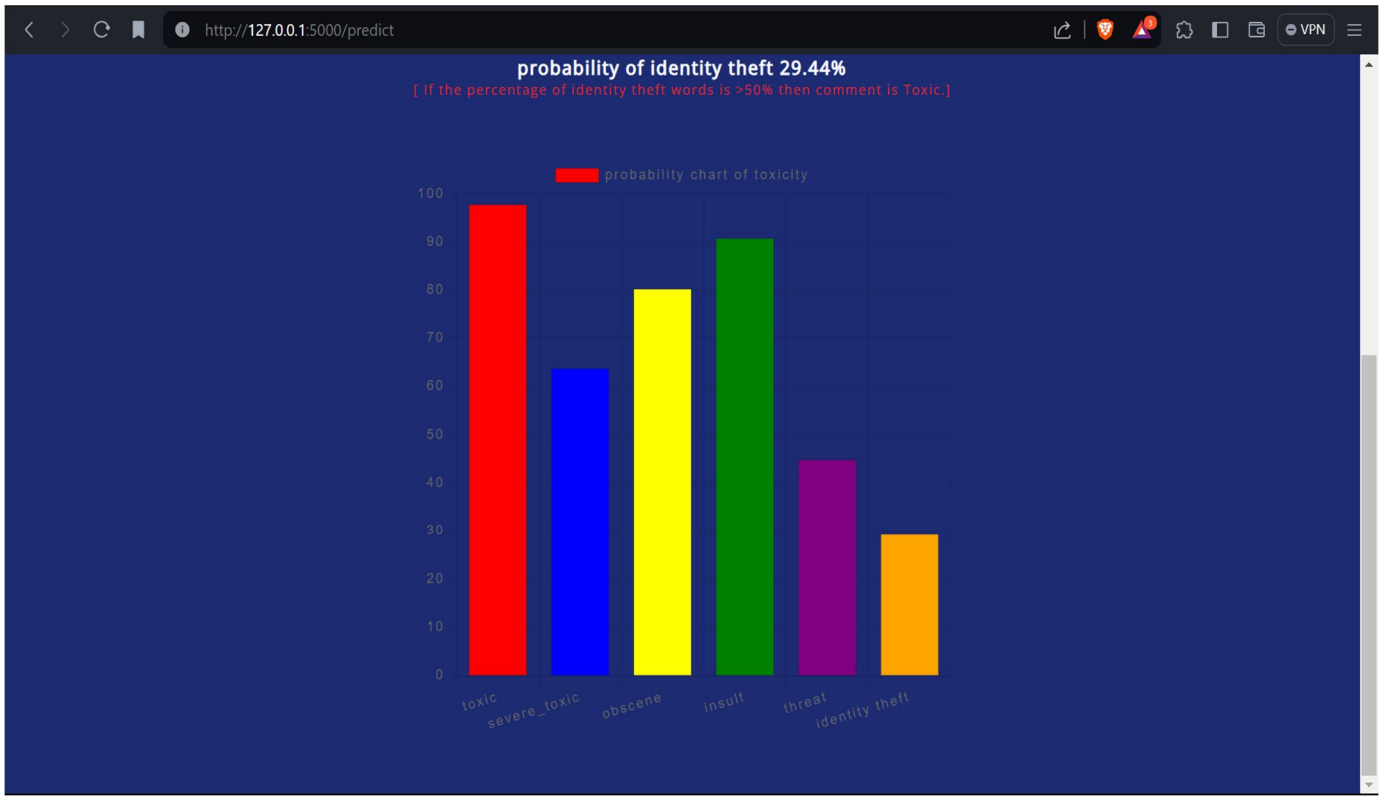


GRAPH REPRESENTATION OF COMMENT CATEGORIES





COMMENT TOXICITY PREDICTION VISULIZATION



CHECKING COMMENT RESULT



USER INPUT FIELD AND PREDICTION DATA

VII. CONCLUSIONS

This research paper has delved into the critical realm of comment toxicity detection utilizing Natural Language Processing (NLP) techniques. Through an extensive exploration of various methodologies and models, it becomes evident that the application of NLP in identifying and mitigating toxic comments is a promising avenue for fostering healthier online discourse. Additionally, the study has highlighted the challenges inherent in the development of accurate and unbiased toxicity detection systems. Issues like context sensitivity, cultural nuances, and evolving linguistic trends pose ongoing challenges that warrant continual refinement of detection models. The fusion of NLP and comment toxicity detection holds immense promise for creating a more inclusive and respectful online space.

REFERENCES

[1] Samer Hassan, Rada Mihalcea, and Carmen Banea. 2007. Random walk term weighting for improved text classification. International Journal of Semantic Computing, 1(04):421–439.

[2] Toxicity detection in online Georgian discussions Nineli Lashkarashvili , Magda Tsintsadze a Department of Computer Science, San Diego State University/ Iv. Javakhishvili, Tbilisi State University, Tbilisi, Georgia b Department of Computer Science, Iv. Javakhishvili Tbilisi State University, Tbilisi, Georgia International Journal of Information Management Data Insights 2 (2022) 100062

[3] Detecting abusive comments at a fine-grained level in a low-resource language☆Bharathi Raja Chakravarthi a,⁎, Ruba Priyadharshini b , Shubanker Banerjee c Manoj Balaji Jagadeeshan d, Prasanna Kumar Kumaresan e, Rahul Ponnusamy f, Sean Benhur g,John Philip McCraee

[4] Aken, V., Risch, B., Kestrel, J., Löser, R., Alexander, 2018. Challenges for toxic comment classification: an in- depth error analysis. In: Proceedings of the 2nd Workshop on Abusive Language

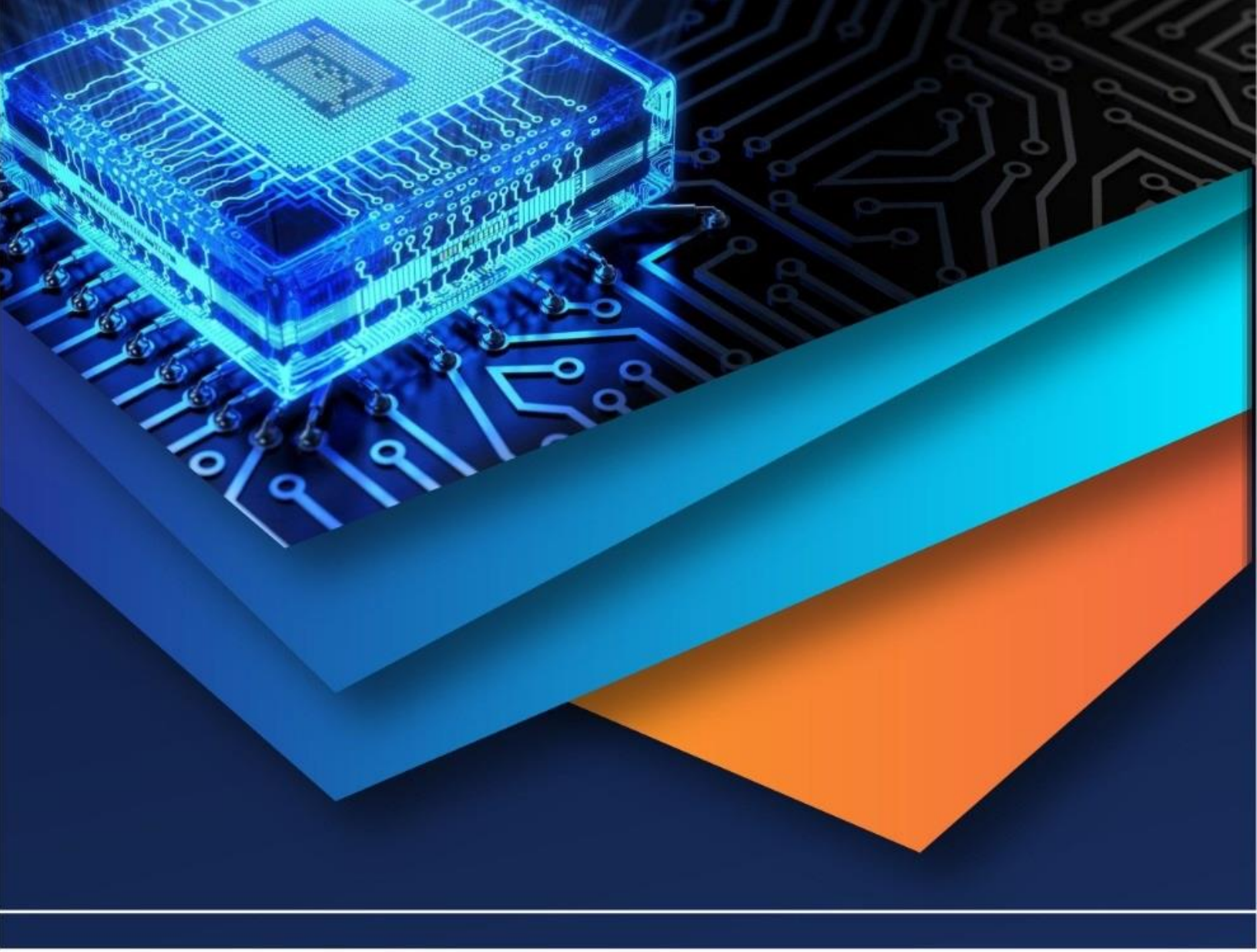
[5] Tekiroğlu, S.S., Chung, Y.L., Guerini, M., 2020. Generating counter-narratives against online hate speech: Data and strategies. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics,
pp. 1177–1190.

[6] Zhilu Z., Mert R. S. 2018. Generalized cross-entropy loss for training deep neural networks with noisy labels. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 8792– 8802.
https://dl.acm.org/doi/10.5555/3327546.3327555

[7] Naseeba, B., Sai, P.H.R., Karthik, B.V.P., Chitteti, C., Sai, K., Avanija, J. (2023). Toxic Comment Classification. In: Abraham, A., Hong, TP., Kotecha, K., Ma, K., Manghirmalani Mishra, P., Gandhi, N. (eds) Hybrid Intelligent Systems. HIS 2022. Lecture Notes in Networks and Systems, vol 647. Springer, Cham.
https://doi.org/10.1007/978-3-031-27409-1_80

[8] Kiran Babu Nelatoori and Hima Bindu Kommanti
Journal: Language Resources and Evaluation, 2024
DOI: 10.1007/s10579-023-09708-6

[9] Amirita Dewani, Mohsin Ali Memon, Sania Bhatti, Adel Sulaiman, Mohammed Hamdi, Hani Alshahrani, Abdullah Alghamdi and Asadullah Shaikh
Journal: Applied Sciences, 2023, Volume 13, Number 4, Page 2062
DOI: 10.3390/app13042062
[10] Marcos Zampieri, Tharindu Ranasinghe, Diptanu Sarkar and Alex Ororbia
Journal: Journal of Intelligent Information Systems, 2023, Volume 60, Number 3, Page 613
DOI: 10.1007/s10844-023-00787-z
[11] Rajnish Pandey and Jyoti Prakash Singh
Journal: Journal of Intelligent Information Systems, 2023, Volume 60, Number 1, Page 235
DOI: 10.1007/s10844-022-00755-z



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)