IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

CLASSIFYING HARMFUL COMMENTS

[1]Mr.Dharmendra Kumar Roy, [2]P. Tejaswini, [3]P. Umesh Chandra, [4]T. Jyothi Latha, [5]M. Deeksha

[1]Associate Professor, [2]Student, [3]Student, [4]Student, [5]Student
[1]Computer Science and Engineering,
[1]Hyderabad Institute of Technology and Management, Hyderabad, India

Abstract: The harmful comments classification system enables users to input text via a web form. Upon submission, the system processes the comment by employing a pre-trained TF-IDF Vectorizer (tfv) and subsequently applies a Random Forest model for toxicity prediction. The toxicity levels are categorized as "Not Toxic," "Moderately Toxic," or "Highly Toxic" based on a predefined classification map. To enhance user comprehension, the application computes and displays the confidence level linked to the prediction, determined by comparing the top two probability scores for the predicted classes. The surge of online communities and social platforms has granted individuals a potent voice, yet it has also exposed them to potentially harmful and offensive content. In addressing this concern and striving to establish safer online spaces, this research centers on developing a machine learning-driven system for the classification of harmful comments.

Index Terms **–** Random Forest Model, Confidence level, TF-IDF Vectorizer, Toxicity prediction.

I. INTRODUCTION

The evolution of internet communication has brought new challenges, especially with the rise of social media. The detection of harmful and toxic content, such as hate speech, is crucial for maintaining a safe online environment. Developing reliable solutions for identifying and preventing the spread of such content is an ongoing challenge to ensure the well-being of internet users. The rise of online participation brings both constructive feedback and challenges. Filtering toxic comments is crucial for maintaining a positive online environment, considering the diverse audience. Governments worldwide are grappling with cyberbullying and its consequences, emphasizing the need for responsible content creation. The democratization of content creation has led to an uncontrollable influx, impacting various aspects of society. Detecting toxic comments remains a significant challenge, hindering healthy public interaction and discouraging participation in online forums. Addressing this issue is essential for fostering a positive online space accessible to all.

The impact extends beyond individuals, influencing businesses, political landscapes, family dynamics, and societal norms. Researchers and developers are actively exploring ways to detect and mitigate toxic comments, recognizing the importance of fostering an environment that encourages constructive engagement. Striking a balance between freedom of expression and the responsibility of content creators is imperative in shaping a digital landscape that promotes positive interaction for all.

1.1 MOVITATION:

The swift evolution of computer science and technology has given rise to the remarkable innovation of the 21st century – the Internet. This interconnected realm enables global communication effortlessly through smartphones and internet connectivity. During the early internet phase, communication centered around emails, posing the challenge of distinguishing between authentic messages and spam.

As time advanced, the landscape of internet communication underwent a significant transformation, notably with the emergence of social media platforms. The ascent of social media underscored the importance of categorizing content as positive or negative to safeguard society and address antisocial behavior.

In recent times, authorities have taken action against individuals based on the harmful and toxic nature of their social media posts. Evaluating these toxic comments involves a systematic process, integrating classification techniques to ensure the precision of the obtained results.

1.2 OBJECTIVE OF THE PROJECT:

The objective of the project is to develop a machine learning-driven system for classifying harmful comments submitted through a web form. This involves utilizing a pre-trained TF-IDF Vectorizer and a Random Forest model to predict toxicity levels, categorizing them as "Not Toxic," "Moderately Toxic," or "Highly Toxic." The system aims to enhance user comprehension by computing and displaying confidence levels associated with the predictions, ultimately contributing to the creation of safer online spaces in response to the prevalence of harmful and offensive content in online communities and social platform.

II. LITERATURE REVIEW

Mujahed A. Saif and Alexander N. Medvedev introduces four models for online abusive comments classification are proposed, which are: logistic regression model and three neural networks models convolutional neural network (Conv), long short-term memory (LSTM), and Conv+ LSTM. Based on the obtained results, one can conclude that the most effective is the combined model Conv+ LSTM, which provides the best accuracy: 0.9820 and 0.9645 when testing on 0.1 and 0.33 of the training data set respectively.

H Fan et al. This paper explores the application of deep learning for classifying social media toxicity, specifically focusing on real-world data related to the UK Brexit. It tested the effectiveness of the BERT-based model in analyzing toxicity in user-generated content, particularly tweets related to the UK Brexit, from two different periods. The results demonstrated the model's efficiency in identifying and analyzing toxic tweets. Deep learning techniques, specifically the adoption of BERT, were employed in toxicity detection within social media user-generated data, specifically tweets.

J. Moon, W. I. Cho focused on creating a dataset for toxic speech detection in online news comments in the Korean language. This initiative aims to compile a comprehensive dataset to train and evaluate models for identifying and handling toxic language within the context of Korean online discussions.

Sulke A. L., and A. S. Varude This paper addresses sorting harmful online comments, using machine learning (k-nearest neighbor, SVM, logistic regression, and decision tree) on Wikipedia's talk page dataset. It emphasizes the importance of accurate classification amidst growing online interactions, achieving higher accuracy through meticulous data preparation, and concluding that support vector machine (SVM) is the preferred algorithm for this task.

EXISTING SYSTEM:

Automated Cyberbullying Activity Detection using Machine Learning Algorithm. It uses smart technology to spot cyberbullying and has become much better at it. It watches online actions in real-time, quickly catching and handling bullying situations. The system learns from various examples of bullying and non-bullying behavior to get smarter over time. It pays attention to the words people use and understands the feelings behind them. The system can handle a lot of online information and regularly updates itself to stay sharp against new bullying trends. By using this technology, it helps make the online world safer by identifying and stopping cyberbullying fast. The methodology involves training the system on diverse datasets containing examples of cyberbullying and non-bullying interactions. Feature extraction techniques are applied to convert raw text data into meaningful features, capturing patterns indicative of cyberbullying. The model is trained using supervised learning algorithms, such as natural language processing and sentiment analysis, to recognize subtle nuances in language that signify harassment. The smart technology in this project involves automated machine learning. This means the system can learn and improve on its own without explicit programming. It uses techniques like natural language processing and sentiment analysis to understand the language used online and identify patterns indicative of cyberbullying. Essentially, it's a computer system that gets better at recognizing and stopping online harassment by learning from various examples and adapting to new situations over time.

PROPOSED SYSTEM:

The proposed approach aims to offer a user-friendly and interpretable tool for content moderation, empowering both users and platform administrators to identify and address harmful content effectively. By leveraging advanced machine learning techniques, the system seeks to contribute to the creation of a safer and more respectful online space. We propose the development and implementation of a machine learning-based application for classifying harmful comments. Users will interact with a web form to submit text comments, and the system will employ a pre-trained TF-IDF Vectorizer (tfv) to capture the textual features. Subsequently, a Random Forest model will be utilized for toxicity prediction, categorizing comments into "Not Toxic," "Moderately Toxic," or "Highly Toxic" based on a predefined classification map.

To enhance user understanding and system transparency, the application will calculate and present a confidence level associated with each toxicity prediction. This confidence level will be derived by comparing the top two probability scores for the predicted classes, providing insights into the reliability of the system's assessments.

III. METHODOLOGY

1. Define Scope

- Identify specific types of harmful comments the system aims to classify (e.g., hate speech, threats, insults).
- Consider the intended users and platforms for implementing the system.

2. Set Objectives

- Establish measurable goals, such as achieving a certain accuracy level in toxicity prediction.
- Determine how the system's effectiveness will be assessed over time.

3. Identify Data Sources

- Specify the sources from which the training data will be collected (e.g., social media platforms, forums).
- Ensure the diversity and representativeness of the data to enhance model generalization.

4. Data Cleaning

- Address issues like missing data, handle outliers, and remove irrelevant information.
- Implement techniques like stemming or lemmatization to standardize text data.

5. Text Vectorization

- Explore the choice of TF-IDF Vectorizer and understand its impact on feature representation.
- Experiment with other text vectorization methods to compare their effectiveness.

6. Model Selection and Training

- Fine-tune hyperparameters of the Random Forest model for optimal performance.
- Consider ensembling or combining multiple models to improve overall accuracy.

7. User Interface Development

- Prioritize user experience by designing an intuitive and responsive web form.
- Include informative visualizations to help users understand the toxicity predictions and confidence levels.

8. Testing and Evaluation

- Implement a robust testing strategy, including cross-validation and splitting data into training and testing sets.
- Use evaluation metrics such as precision, recall, and F1 score to assess model performance.

IV.IMPLEMENTATION

Implementation includes building the wen application on classifying harmful comments with the required features. Implementation includes building app.py.

Various Steps in Implementation:

The major step in the implementation includes the following steps. They are as follows :-

1. Building the front-end (Basic UI) using Flask.
2. Adding the functionality of Data Classification.
3. Integrating both to get our final GUI.

RESULT AND DISCUSSION

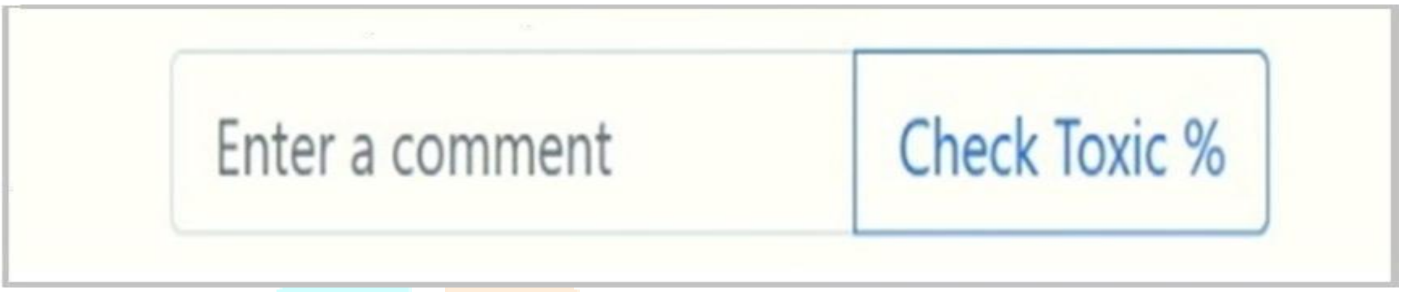


Figure: User can enter a comment to check the initial classification and the chances of word becoming toxic.

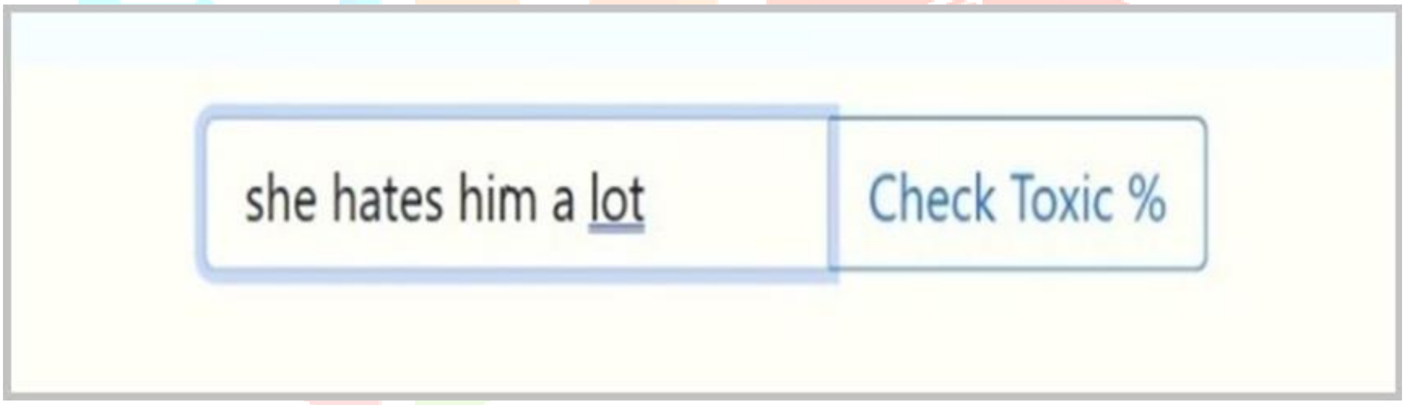
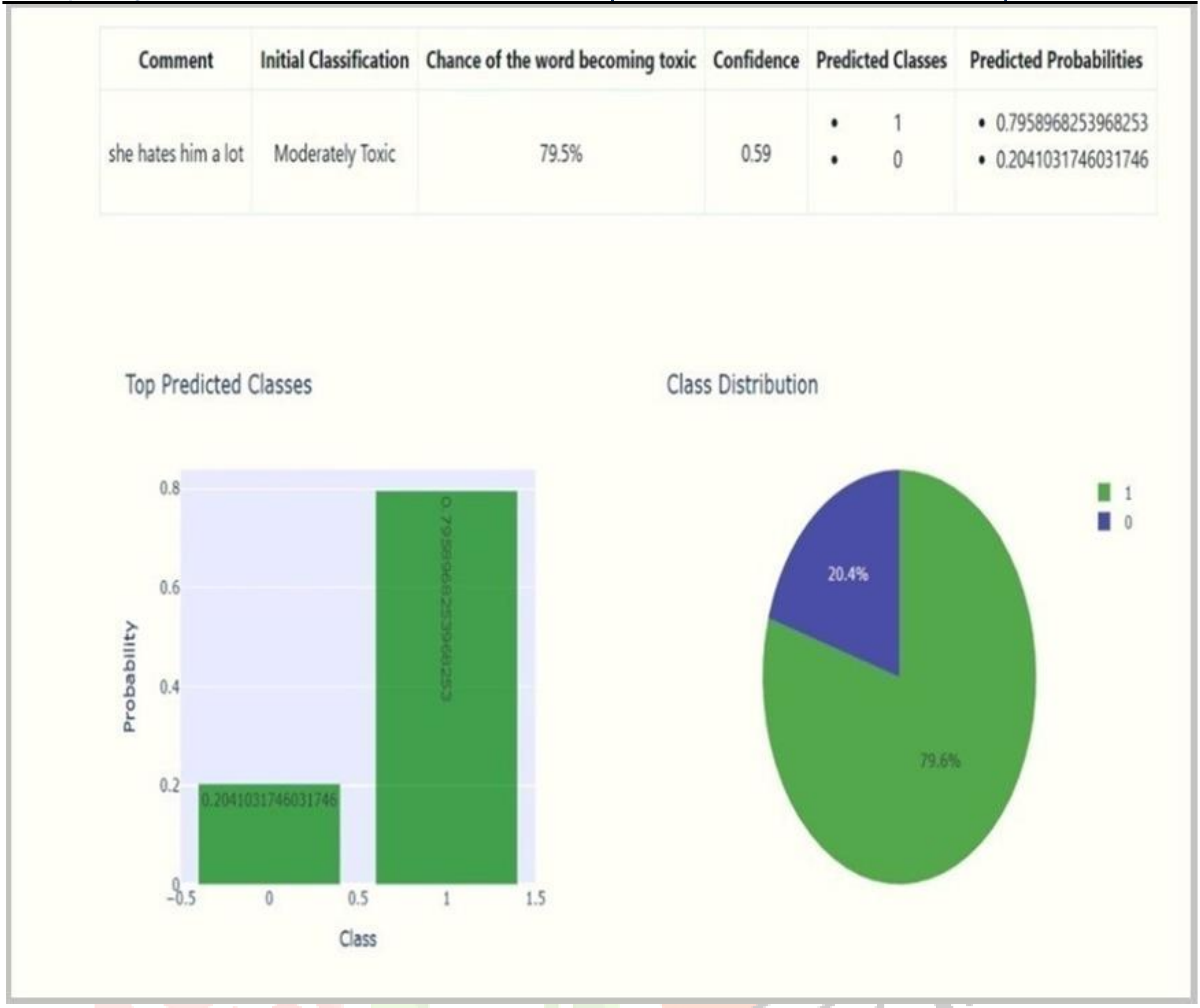

she hates him a lot

Figure: The user has entered a comment and by clicking on check toxic % we can see the result.



Comment	Initial Classification	Chance of the word becoming toxic	Confidence	Predicted Classes	Predicted Probabilities
she hates him a lot	Moderately Toxic	79.5%	0.59	• 1 • 0	• 0.7958968253968253 • 0.2041031746031746

Figure: User finds the above comment as moderately toxic and the chances of becoming toxic is 79.5% and the application calculates a confidence level associated with each toxicity prediction by comparing the top two probability scores for the predicted classes.

CONCLUSION

In conclusion, our suggested machine learning-based application offers a user-friendly tool for comment classification, thereby addressing the crucial problem of online toxicity. Through the use of a Random Forest model and a pre-trained TF-IDF Vectorizer, the system is able to precisely classify comments into three toxicity levels. By enabling users and administrators to recognize and remove hazardous content, this method helps create a safer online environment. The web form approach makes involvement easy, and adding a confidence level makes toxicity forecasts more transparent. Our technology uses cutting-edge machine learning techniques to support the larger objective of creating a more polite online community. Long-term efficiency is ensured by the iterative refinement process, which allows for adaptation to changing harmful patterns.

REFERENCES

[1]. Saif, Mujahed A., et al. "Classification of online toxic comments using the logistic regression and neural networks models." AIP conference proceedings. Vol. 2048. No. 1. AIP Publishing, 2018.

[2]. Fan H, Du W, Dahou A, Ewees AA, Yousri D. Social media toxicity classification using deep learning: real-world application UK Brexit. Electronics. 2021 Jun 1;10(11):1332.

[3]. Moon, J., Cho, W.I. and Lee, J., 2020. BEEP! Korean corpus of online news comments for toxic speech detection. arXiv preprint arXiv:2005.12503.

[4]. Sulke, A. L., and A. S. Varude. "Classification of Online Pernicious Comments using Machine Learning." IJSRD-International Journal for Scientific Research & Development (2019).

[5]. Van Aken, Betty, Julian Risch, Ralf Krestel, and Alexander Löser. "Challenges for toxic comment classification: An in-depth error analysis." *arXiv preprint arXiv:1809.07572* (2018).

[6]. Androcec, Darko. "Machine learning methods for toxic comment classification: a systematic review." *Acta Universitatis Sapientiae, Informatica* 12, no. 2 (2020): 205-216.