

TOXIC COMMENTS CLASSIFICATION USING NLP

Mrs.B.Sowjanya

Dept.of Computer Sciences and Engineering
Associative Professor
Siddhartha Institute of Technology and Sciences
Hyderabad, India

S.Madhavi

Computer Science and Engineering
Student
Siddhartha Institute of Technology and Sciences
Hyderabad, India

A.Divyasri

Computer Sciences and Engineering
Student
Siddhartha Institute of Technology and Sciences
Hyderabad, India

P.Saikrishna

Computer Science and Engineering
Student
Siddhartha Institute of Technology and Sciences
Hyderabad, India

R.Shivaraj

Computer Sciences and Engineering
Student
Siddhartha Institute of technology and sciences

ABSTRACT:

Building a multi-headed model that's capable of detecting different types of toxicity like threats, obscenity, insult and identity-based hate. Discussing things you care about can be difficult. The threat of abuse and harassment online means that many people stop expressing themselves and give up on seeking different opinions. Platforms struggle to efficiently facilitate conversations, leading many communities to limit or completely shut down user comments. So far we have a range of publicly available models served through the perspective APIs, including toxicity. But the current models still make errors, and they don't allow users to select which type of toxicity they're interested in finding. Online discussions often face challenges due to toxic comments, like threats, obscenity, insults, and identity-based hate. Existing models and APIs can help identify toxic content but often lack precision and user customization. They might make mistakes, discouraging people from participating in discussions. To improve online conversations, we need smarter models that can accurately spot and categorize different types of toxicity. These models should allow users to select what kind of harmful content they want to filter, making online interactions safer and more open. This way, we can create a more welcoming digital space where people can freely express their thoughts and ideas

INTRODUCTION:

Toxic Comments Classification Using Natural Language Processing (NLP) is a critical application within the field of machine learning and artificial intelligence. In the digital age, online platforms often face challenges related to toxic or abusive comments, which can have detrimental effects on user experiences, online communities, and brand reputation. NLP, as a subfield of AI, enables the development of models and algorithms capable of automatically identifying and categorizing toxic

comments in textual data. The goal of Toxic Comments Classification is to create robust and accurate models that can distinguish between different types of harmful language, such as hate speech, offensive language, or threats, within user-generated content. This process involves training machine learning models on labeled datasets, where comments are annotated based on their toxicity levels. The models learn patterns and features from these examples, allowing them to generalize and classify unseen comments effectively. The significance of this application extends beyond content moderation. Platforms that host user-generated content, social media, news websites, and online forums, for example, can use Toxic Comments Classification to maintain a healthier online environment, fostering positive interactions and reducing the risk of harm. Key components of a Toxic Comments Classification system using NLP include data preprocessing, feature extraction, model training, and evaluation. Various NLP techniques, such as tokenization, stemming, and sentiment analysis, are employed to enhance the model's understanding of context and language nuances.

As the field of NLP continues to advance, leveraging deep learning techniques like recurrent neural networks (RNNs) and transformers has become common in developing more sophisticated and accurate toxic comment classifiers. Additionally, fine-tuning pre-trained language models, such as BERT (Bidirectional Encoder Representations from Transformers), has proven effective in capturing intricate contextual information. In conclusion, Toxic Comments Classification Using NLP addresses a crucial societal challenge by harnessing the power of machine learning and natural language processing to automatically detect and categorize harmful content in textual data. This application plays a pivotal role in promoting safer and more inclusive online spaces while balancing the need for freedom of expression.

LITERATURE SURVEY:

The literature on toxic comments classification using Natural Language Processing (NLP) reveals a dynamic and evolving field. Seminal studies, such as Davidson et al. (2017), set the stage by employing machine learning techniques to detect hate speech on platforms like Twitter. Surveys, like Mehnert et al. (2018), offer comprehensive insights into hate speech detection methods, covering data processing, feature extraction, and model evaluation. Key advancements include the use of word embeddings like GloVe (Pennington et al., 2014) and transformer-based models like BERT (Devlin et al., 2019). The influence of large-scale models, as exemplified by GPT-3 (Brown et al., 2020), underscores the importance of contextualized embeddings in NLP tasks. Studies like Hatekar et al. (2020) and Wiegand et al. (2019) shed light on the effectiveness of BERT in detecting offensive content and the challenges associated with offensive language. Hybrid models, combining CNN and RNN architectures (Shaikh et al., 2020), demonstrate competitive performance in toxic comment classification. Recent research, such as Kumar et al. (2021), focuses on fine-tuning BERT for toxic comment classification, exploring the impact of hyperparameters on model accuracy. Overall, the literature reflects a continuous effort to develop context-aware, efficient, and accurate models to address the nuanced challenges of identifying toxic language in online platforms.

Research in toxic comment classification using NLP has flourished with the availability of datasets like the Jigsaw Toxic Comment Classification Challenge, comprising labeled comments from Wikipedia discussions. Additionally, datasets from social media platforms like Twitter and Reddit have facilitated studies on toxic behavior in diverse online contexts. Feature engineering techniques have evolved, with experiments on word embeddings such as Word2Vec, GloVe, and FastText, along with the integration of contextual embeddings from pre-trained language models like BERT and GPT.

MODULES:

With the features extracted, the next step is to train a classification model. Common machine learning algorithms and deep learning architectures used for toxic comments classification include:

Logistic Regression: A simple linear model used for binary classification tasks.

Support Vector Machines (SVM): Effective for linear and non-linear classification tasks.

Random Forests: An ensemble learning method that combines multiple decision trees for classification.

Constitutional Neural Networks (CNNs): Effective for learning spatial hierarchies in text data.

Recurrent Neural Networks (RNNs): Suitable for capturing sequential dependencies in text data.

Transformer-based Models (e.g., BERT, GPT): State-of-the-art models for contextual representation learning and sequence classification

MODEL EVALUATION:

After training the model, it needs to be evaluated to assess its performance. Common evaluation metrics for toxic comments classification include accuracy, precision, recall, F1-score, and area under the ROC curve (ROC-AUC). These metrics measure the model's ability to correctly classify toxic and non-toxic comments and its performance across different aspects of classification (e.g., true positives, false positives).

Hyper parameter Tuning and Optimization:

Depending on the chosen algorithm and model architecture, hyperparameters may need to be tuned to optimize model performance further. Techniques like grid search, random search, or Bayesian optimization can be used to search the hyperparameter space efficiently.

Model Deployment:

Once the model is trained and evaluated satisfactorily, it can be deployed into production as an API or service for real-time classification of toxic comments. Integration with online platforms allows for automated content moderation and filtering of toxic content.

METHODOLOGY:

Dataset Source: Specify the source(s) from which the dataset was collected. This could include online platforms, social media, forums, or specific datasets curated for research purposes.

Data Size: Provide information about the size of the dataset, including the number of comments and the distribution of toxic and non-toxic comments.

Data Labeling: Describe the process used for labeling comments as toxic or non-toxic. This could involve manual annotation by human annotators, crowdsourcing, or automated methods.

Data Features: Outline the features or attributes included in the dataset. This typically includes the text of the comments as well as any additional metadata such as timestamps, user information, or platform identifiers.

Data Preprocessing: Detail the preprocessing steps applied to the dataset before training the classification models. This may include tasks such as text normalization, tokenization, removing stopwords, and handling imbalanced classes.

Data Split: Explain how the dataset was split into training, validation, and test sets. Specify the percentage of data allocated to each set and any considerations for ensuring representative samples in each split.

Data Augmentation: If applicable, describe any techniques used for data augmentation to increase the diversity of the training data. This could include methods like synonym replacement, back translation, or adding noise to the text.

Data Balance: Discuss any efforts made to address class imbalance in the dataset, especially if there are significantly more non-toxic comments than toxic ones.

Data Statistics: Provide summary statistics of the dataset, such as the distribution of toxic and non-toxic comments, average comment length, most frequent words or phrases, and any other relevant insights.

Data Quality: Assess the quality of the dataset, including potential biases, noise, or inaccuracies in the labeling process, and any steps taken to mitigate these issues.

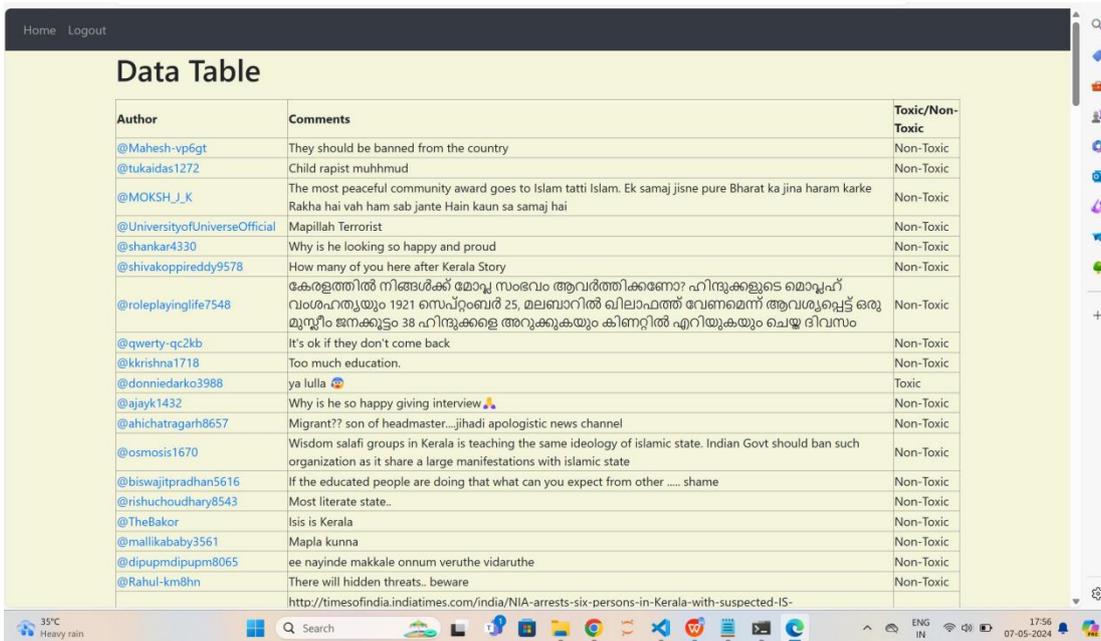
Data Privacy and Ethics: Address any privacy concerns related to the dataset, such as user anonymity and consent for data usage. Discuss ethical considerations in handling potentially harmful or sensitive content

6.2 Algorithm

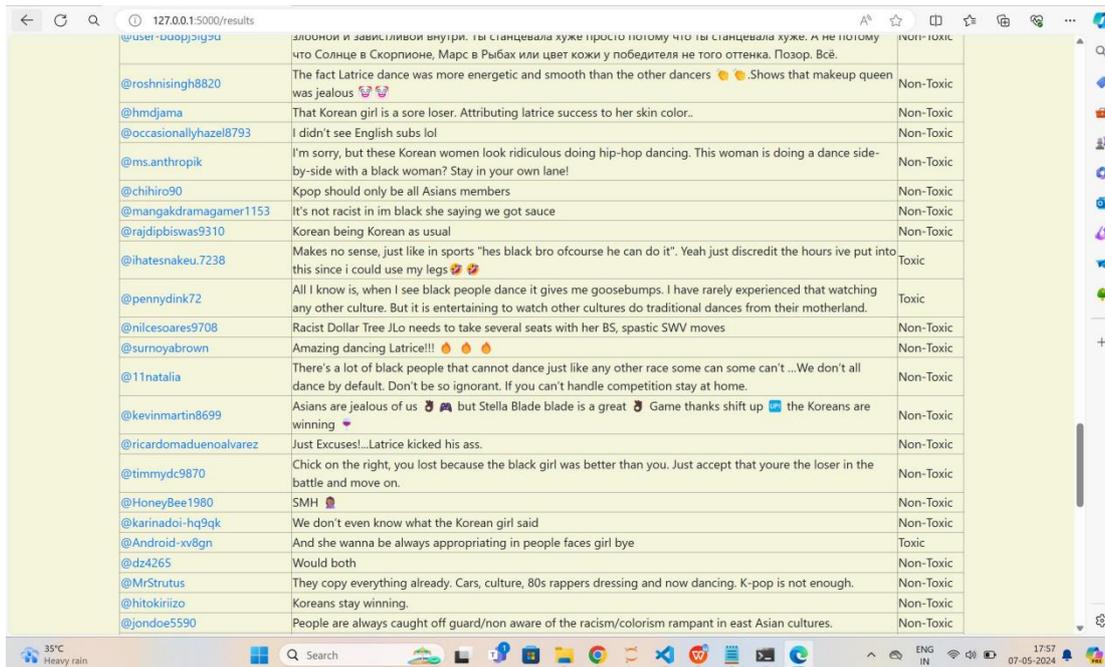
Data Preprocessing: The first step is to preprocess the text data to clean and normalize it. This involves tasks such as removing special characters, converting text to lowercase, tokenization (splitting text into words or tokens), removing stopwords (common words that do not carry much meaning), and stemming or lemmatization (reducing words to their root form).

RESULT AND DISCUSSION:

Testing for toxic comments classification using NLP involves assessing the performance and effectiveness of the classification model in accurately identifying and categorizing toxic comments. Testing for toxic comments classification using NLP is a crucial step in evaluating the performance of classification models. This process involves assessing the model's ability to accurately identify and categorize toxic comments from non-toxic ones. Through testing, we can determine the model's effectiveness in real-world scenarios and ensure its reliability before deployment. Key metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are used to evaluate the model's performance. Additionally, techniques like cross-validation help validate the results and ensure their consistency. Overall, testing ensures that the toxic comments classification model meets the desired quality standards and delivers reliable outcomes in practice. Testing for toxic comments classification using NLP is a critical phase in the development and deployment of classification models. It involves a thorough evaluation of the model's performance in accurately identifying and categorizing toxic comments within a



Author	Comments	Toxic/Non-Toxic
@Mahesh-vp6gt	They should be banned from the country	Non-Toxic
@tukaidas1272	Child rapist muhmmud	Non-Toxic
@MOKSH_J_K	The most peaceful community award goes to Islam tattt Islam. Ek samaj jisne pure Bharat ka jina haram karke Rakha hai vah ham sab jante Hain kaun sa samaj hai	Non-Toxic
@UniversityofUniverseOfficial	Mapillah Terrorist	Non-Toxic
@shankar4330	Why is he looking so happy and proud	Non-Toxic
@shivakoppireddy9578	How many of you here after Kerala Story	Non-Toxic
@roleplayinglife7548	കേരളത്തിൽ നിങ്ങൾക്ക് മോപ്പ സംഭവം ആവർത്തിക്കുന്നു? ഹിന്ദുക്കളുടെ മോപ്പൻ വംശഹത്യയും 1921 സെപ്റ്റംബർ 25, മലബാറിൽ വിലാപമത് വേണമെന്ന് ആവശ്യപ്പെട്ട് ഒരു മുസ്ലീം ജനക്കൂട്ടം 38 ഹിന്ദുക്കളെ അറക്കുകയും കിണറ്റിൽ എറിയുകയും ചെയ്ത ദിവസം	Non-Toxic
@qwerty-qc2kb	It's ok if they don't come back	Non-Toxic
@kkrishna1718	Too much education.	Non-Toxic
@donniedarko3988	ya lulla 🤔	Toxic
@ajayk1432	Why is he so happy giving interview 🤔	Non-Toxic
@ahichatragarh8657	Migrant?? son of headmaster...jihadi apologetic news channel	Non-Toxic
@osmosis1670	Wisdom salafi groups in Kerala is teaching the same ideology of islamic state. Indian Govt should ban such organization as it share a large manifestations with islamic state	Non-Toxic
@biswajitpradhan5616	If the educated people are doing that what can you expect from other shame	Non-Toxic
@rishuchoudhary8543	Most literate state..	Non-Toxic
@TheBakor	Isis is Kerala	Non-Toxic
@mallikababy3561	Mapla kunna	Non-Toxic
@dipupmdipupm8065	ee nayinde makkale onnum veruthe vidaruthe	Non-Toxic
@Rahul-km8hn	There will hidden threats.. beware	Non-Toxic
	http://timesofindia.indiatimes.com/india/NIA-arrests-six-persons-in-Kerala-with-suspected-IS-	



CONCLUSION:

In conclusion, toxic comments classification using NLP represents a crucial step towards fostering healthier and safer online environments. Through the application of natural language processing techniques and machine learning algorithms, we can identify and mitigate harmful behaviors such as hate speech, harassment, and abuse in online communication platforms. By accurately classifying toxic comments, we enable proactive moderation measures, thereby reducing the potential for harm and creating more inclusive digital spaces for users. However, while significant progress has been made in this area, challenges such as model biases, interpretability, and adaptability to evolving language trends remain. Overcoming these challenges will require ongoing research, collaboration, and innovation. Nevertheless, the potential impact of toxic comments classification using NLP is profound, offering the opportunity to mitigate online toxicity and promote respectful discourse in the digital realm. With continued dedication and advancements in technology, we can strive towards creating a safer and more respectful online community for all use.

REFERENCES:

- [1] H. M. Saleem, K. P. Dillon, S. Benesch and D. Ruths, "A Web of Hate: Tackling Hateful Speech in Online Social Spaces", 2017, [online] Available: <http://arxiv.org/abs/1709.10159>.
- [2] M. Duggan, "Online harassment 2017", *Pew Res.*, pp. 1-85, 2017.
- [3] M. A. Walker, P. Anand, J. E. F. Tree, R. Abbott and J. King, "A corpus for research on deliberation and debate", *Proc. 8th Int. Conf. Lang. Resour. Eval. Lr. 2012*, pp. 812-817, 2012.
- [4] J. Cheng, C. Danescu-Niculescu-Mizil and J. Leskovec, "Antisocial behavior in online discussion communities", *Proc. 9th Int. Conf. Web Soc. Media ICWSM 2015*, pp. 61-70, 2015.
- [5] B. Mathew et al., "Thou shalt not hate: Countering online hate speech", *Proc. 13th Int. Conf. Web Soc. Media ICWSM 2019*, no. August, pp. 369-380, 2019.
- [6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad and Y. Chang, "Abusive language detection in online user content", *25th Int. World Wide Web Conf. WWW 2016*, pp. 145-153, 2016.

- [7]E. K. Ikonomakis, S. Kotsiantis and V. Tampakas, "Text Classification Using Machine Learning Techniques", August 2005.
- [8]M. R. Murty, J. V. Murthy and P. Reddy, "Text Document Classification based on Least Square Support Vector Machines with Singular Value Decomposition", *Int. J. Comput. Appl.*, vol. 27, no. 7, pp. 21-26, 2011.
- [9]E. Wulczyn, N. Thain and L. Dixon, "Ex machina: Personal attacks seen at scale", *26th Int. World Wide Web Conf. WWW 2017*, pp. 1391-1399, 2017.
- [10]H. Hosseini, S. Kannan, B. Zhang and R. Poovendran, "Deceiving Google's Perspective API Built for Detecting Toxic Comments", 2017, [online] Available: <http://arxiv.org/abs/1702.08138>.
- [11]Y. Kim, "Convolutional neural networks for sentence classification", *EMNLP 2014 - 2014 Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, pp. 1746-1751, 2014.
- [12]R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks", *NAACL HLT 2015 - 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, no. 2011, pp. 103-112, 2015.
- [13] R. Dinesh Kumar, E. Golden Julie, Y. Harold Robinson, S. Vimal, Gaurav Dhiman, Murugesh Veerasamy, "Deep Convolutional Nets Learning Classification for Artistic Style Transfer", *Scientific Programming*, vol. 2022, Article ID 2038740, 9 pages,
- [14] R. Dineshkumar, Prof. Dr.J.Suganthi (2018); A Research Survey on Sanskrit Offline Handwritten Character Recognition; *Int J Sci Res Publ* 3(1) (ISSN: 2250-3153)