Toxic Comment Classification

M.Shangavelan
Vellore Institute of Technology, Chennai,
India.
20BEC1230
shangavelan.m2020@vitstudent.ac.in

R.H. Sanjay
Vellore Institute of Technology, Vellore,
India.
20BIT0348
sanjay.rh2020@vitstudent.ac.in

Sanskar Sinha
Vellore Institute of Technology, Vellore,
India.
20BIT0273
sanskar.sinha2020@vitstudent.ac.in

***Abstract*— Twitter or comment sections of online news platforms are an essential space to express opinions and discuss political topics. However, the misuse by spammers, haters, and trolls makes costly content moderation necessary. There is growing concern that a toxic culture in online platforms such as Twitter presents a barrier to diverse participation in the digital world. Conversational toxicity is an issue that can lead people both to stop genuinely expressing themselves and to stop seeking other's opinions out of fear of abuse/harassment. The proposed methodology includes Random Forest classification algorithm to solve the problem of toxic comment classification. The dataset is provided by Jigsaw toxic comment classification challenge on Kaggle.**

I. Introduction

Social media platforms like Facebook, Instagram, Twitter, and more are giving people a chance to connect across distances. In other words, the whole world is at our fingertips all thanks to social media. The youth is especially one of the most dominant users of social media. However, these popular social media platforms have also caused the emergence of a toxic comments culture. The comment sections of social media platforms have become a new playground for online bullying. Hence, it has become a need of the hour to filter out these toxic comments. The goal of this project is to use various deep learning and machine learning techniques to identify toxic tweets and comments, based on the words they contain, and other features, such as sentiment they convey, which can be used to help deter users from posting potentially hurtful messages.

A. Implications of the study and benefits

Reducing Harmful Content: A toxic tweet comment detection system can help identify and remove harmful content, such as hate speech, cyberbullying, and harassment. This can make online communities safer and more welcoming for everyone.

Improving User Experience: Toxic comments can discourage users from engaging with content and interacting with other users. By filtering out toxic comments, a detection system can help create a more positive user experience, encouraging users to stay engaged and participate in discussions.

Protecting Brand Reputation: Toxic comments can damage the reputation of brands and organizations. A detection system can help identify and remove harmful content, protecting a brand's reputation and ensuring that their online presence remains positive.

Saving Time and Resources: Manually moderating comments can be a time-consuming and resource-intensive task. A toxic tweet comment detection system can automate the process, saving time and resources while ensuring that harmful content is identified and removed in a timely manner.

B. Applications

Social Media Moderation: A toxic tweet detection system using machine learning can be used to automatically moderate comments on social media platforms, identifying and removing harmful content such as hate speech, harassment, and cyberbullying.

News and Media Comment Moderation: A machine learning-based toxic tweet detection system can also be used by news and media websites to automatically moderate comments, ensuring that their platform remains a safe space for discussion.

Online Community Moderation: Online communities such as forums, blogs, and discussion boards can benefit from a toxic tweet detection system using machine learning to identify and remove harmful content, creating a safe and positive environment for members.

Online Advertising: Brands and marketers can use a toxic tweet detection system using deep learning to monitor and analyze user comments and sentiment towards their products or services on social media platforms, helping them to better understand and respond to their audience.

Public Safety: A machine learning-based toxic tweet detection system can be used by law enforcement agencies to monitor social media activity and identify potential threats or instances of hate speech, aiding in their efforts to ensure public safety.

Mental Health Support: A machine learning-based toxic tweet detection system can also be used in mental health support applications, such as online therapy or counseling platforms,

to monitor user comments and identify potential triggers or harmful content that could impact a user's mental health.

Political Campaign Monitoring: Political campaigns can use a toxic tweet detection system using deep learning to monitor social media platforms and identify instances of misinformation or toxic comments directed at their campaign or candidate. This can help them to better understand and respond to their audience and to counteract harmful content.

Brand Reputation Management: A deep learning-based toxic tweet detection system can also be used by brands to monitor and manage their online reputation, by identifying and removing harmful comments that could negatively impact their brand image. This can help to build a positive online presence and enhance customer trust and loyalty.Language understanding for virtual assistants: Test cases can be generated from text conversations between users and virtual assistants to evaluate the assistant's ability to understand natural language and perform tasks.

II. LITERATURE SURVEY

[1] summarizes that, the internet has accelerated the generation of writer-generated sentiment content about various topics, which can be beneficial for businesses, governments, and individuals. However, analysing this content accurately using sentiment analysis techniques faces several challenges. These challenges hinder the identification of the precise meaning of sentiments and the detection of appropriate sentiment polarity. This paper provides a survey of these challenges and relevant approaches and techniques used in sentiment analysis.

[2] summarizes that, the rise of toxic content online has led to research in automatic detection and classification. However, there are still two limitations: the lack of support for multi-label classification and the impact of unbalanced datasets. This paper builds three state-of-the-art methods for multi-label classification and compares their performance based on the size of the training data. The methods include Support Vector Machine, Convolutional Neural Networks, and Long-Short-Term Memory Networks. Results show that CNN is the most robust method, challenging the belief that Neural Networks need significant data to train accurately. The paper also provides indicative thresholds for training data size to determine reliable classifier performance.

[3] summarizes that, the analysis of sentiment on social networks is valuable in understanding public opinion, but faces challenges in natural language processing. Deep learning models have shown promise in solving these challenges, and this paper reviews recent studies using deep learning for sentiment analysis, specifically sentiment polarity. The studies employed models using TF-IDF and word embedding on various datasets. A comparative study was conducted on the experimental results, highlighting the effectiveness of deep learning models in improving sentiment analysis accuracy.

[4] summarizes that, social media is a platform for people to express their feelings and opinions through short text messages. Sentiment analysis is crucial in identifying anxiety, depression, well-being, and mood of individuals or communities. It involves natural language processing and machine learning techniques to classify sentiment expressed in text documents. This paper discusses sentiment recognition using textual data and techniques used in sentiment analysis, emphasizing the importance of content-based classification in the field.

[5] summarizes that, social media platforms like Twitter are popular for sharing emotions and opinions, making sentiment analysis valuable for various fields. However, earlier studies focused on single-label classification and did not consider co-occurring emotions. This study proposes a multi-labelled emotion classification method using convolutional neural networks called CNN-EISC on the SEMEVAL2018 Task-1 dataset. The model classifies Anger, Sad, Fear, and Joy into four intensities under diverse classes and achieves a high average precision, recall, F-measure, and accuracy, demonstrating its effectiveness.

[6] introduces a Natural Language Processing technique to classify text into toxic and non-toxic categories, with the aim of creating a model for anti-bullying efforts. The study compared the performance of LSTM and Naive Bayes methods and found that LSTM had a 20% higher true positive rate, indicating its potential as a game changer in comment classification. The use of data science and smart technologies can lead to a healthier environment for virtual societies, as demonstrated by this study's promising results.

[7] proposes a contextual sentiment neural network (CSNN) model to address the black-box nature of deep neural networks in text sentiment analysis. It has interpretable layers that can explain the process of sentiment analysis prediction in a human-like way. A novel learning strategy called initialization propagation (IP) learning is used to realize the interpretability of each layer in the CSNN. Experimental results show that the IP learning is effective in improving interpretability, and the CSNN has high predictability and high explanation ability in sentiment analysis.

[8] proposes using deep learning models with LSTM neural networks for text mining to identify and filter out hate-speech in social media. With the increasing prevalence of hateful and abusive comments online, such a model could help create a safer and cleaner online environment. The model developed achieved a high level of accuracy, with precision of 94.49%, recall of 92.79%, and an accuracy score of 94.94% in classifying comments as toxic or non-toxic.

[9] states that, Online communication brings quality to human life, but it also has dangers like personal attacks and harassment. Machine learning tools are needed to manage the vast amount of information generated, and deep learning approaches using Convolutional Neural Networks (CNN) show promising performance in text classification. This paper uses CNNs to detect toxic comments in a large dataset of Wikipedia talk page edits and compares it with the traditional bag-of-words approach. Results show that CNNs improve toxic comment classification, supporting further research in this area.

The study done in [10] investigates the effectiveness of the dropout layer, a widely used technique to prevent overfitting in deep neural networks, on classical regression problems. A 3-layer deep learning net with a single dropout layer was tested on 8 real regression datasets. Results showed that the dropout layer did not improve overfitting in these datasets. This study suggests that the dropout layer may not be effective for classical regression problems and its usage in such tasks needs further exploration.

[11] presents a method for automatic detection of toxic South African tweets using an English corpus. The study evaluates the performance of Support Vector Machines with various n-gram features for classification, and finds that combining the classifiers improves accuracy and F-measure scores. The best performing classifier used a combination of unigram and bigram word n-grams and character n-grams of length 3 to 7. The results compare favourably to previous work on English corpus, indicating the model's reliability for detecting toxic tweets in South African context.

[12] proposes an aspect-based opinion mining system that uses sentiment lexicons and deep learning techniques such as Bi-LSTM for improved sentiment analysis. The proposed framework was evaluated using online product reviews from various repositories. Results indicate that Bi-LSTM outperforms other algorithms in terms of accuracy, and the proposed system can be used to determine customer satisfaction with products, prices, and customer service. The framework includes standard NLP processes such as data preparation and keyword extraction. The paper also provides future research directions for sentiment analysis.

[13] states about artificial Neural Networks are modelled after the human brain's network of neurons. They consist of different layers of interconnected cells that process and pass on information to the outermost layer which produces the output. Nonlinear activation functions are crucial in the learning process of these networks as they make sense of complicated mappings between inputs and outputs. In essence, Artificial Neural Networks mimic the functionality of the human brain in processing information.

[14] states that the MeTwo dataset is the first Spanish Twitter dataset created to understand and analyse the expression of sexism in online conversations. This work proposes a new task to detect different types of sexist behaviours, from explicit hate to subtle expressions, and investigates the feasibility of using traditional and deep learning techniques for automatic detection. Results show that sexism is prevalent in many forms in social networks and can be detected using deep learning approaches. The generalizability of the task to other subdomains, such as misogyny, is discussed.

[15] addresses the issue of toxic comment detection in Roman Urdu, a widely used language on social media in South Asia. The authors developed a labelled corpus of toxic and non-toxic comments called RUT, which contains over 72 thousand comments collected from popular social media platforms. They trained several classification models, including classical machine learning and deep models, and proposed an ensemble approach that achieved an F1-score of 86.35%, setting the first-ever benchmark for toxic comment classification in Roman Urdu.

[16] explores the harmful side effects of social media and develops an efficient model using Bidirectional Encoder Representations from Transformers (BERT) to detect and classify toxicity in social media content. The model is fine-tuned on a labeled dataset and tested with two datasets collected from Twitter related to the UK Brexit. The results demonstrate the proposed model's ability to classify and analyze toxic tweets effectively.

[17] states about TrollPacifier, which is a holistic system for troll detection on Twitter. The system analyses different features of trolls and legitimate users, such as writing style, sentiment, behaviour, social interactions, linked media, and publication time. The system combines existing and new techniques to achieve a very high accuracy of 95.5%. This work provides an up-to-date analysis of troll detection, a systematic grouping of features, a description of a working system, and a comparison among different features with machine learning.

[18] states about a model for temporal tracking of comments toxicity is proposed using tweets related to a specific hashtag. A Convolutional Neural Network is trained for toxic comment prediction and used to categorize tweets based on toxicity. An adapted change detection approach is applied to monitor toxicity trend changes over time. Experimental results showed that toxic comment classification on Twitter conversations can reveal significant knowledge, and changes in toxicity are accurately identified over time.

[19] focuses on a comparison study between the traditional deep learning techniques like LSTMs and GRUs with the latest state-of-the-art transformers for the task of classifying a comment based on its toxicity. Specifically, four different deep learning models are built, implemented and trained on a standard dataset for comparing the performance of the models. The models that are explored and compared are Bi-directional LSTMs, GRUs, Bidirectional LSTMs with CNNs and Transformers (with pre-trained RoBERTa weights).

The study in [20] proposes a method to classify online comments as toxic or non-toxic using machine learning. The dataset is pre-processed to remove noise and errors, and the TF-IDF technique is used to transform raw comments before training with logistic regression. The model can differentiate between different types of toxicity, including severe-toxic, obscene, threat, insult, and identity-hate. Confusion metrics are used to evaluate the performance of the model. The study shows promise in accurately identifying toxic comments in online domains.

[21] paper presents a proposed detection scheme for identifying hateful content on social media. The scheme is an ensemble of Recurrent Neural Network classifiers that use various features related to user information, such as their tendency towards racism or sexism, and word frequency vectors derived from the text. The scheme was evaluated on a corpus of 16k tweets and demonstrated its effectiveness in distinguishing hateful messages from normal text, achieving higher classification quality than current state-of-the-art algorithms.

Aspect-based sentiment analysis (ABSA) is a task that aims to identify the sentiment polarity of a specific aspect in text. In [22], a model called Synthetic Attention in Bidirectional Encoder Representations from Transformers (SA-BERT) was proposed for sentiment detection in review datasets. The model uses attention networks to effectively represent the target and context of the reviews, and a transformer to input the word vectors in parallel. The synthetic attention mechanism is then used to learn essential parts of the context and aspects in the reviews. The proposed SA-BERT and SA-BERT-XGBoost models achieved high accuracy and F1 scores on restaurant16 and restaurant14 datasets, outperforming baseline models such as DLCF-DCA-CDM, R-GAT+BERT, ASGCN-DG, AEN-BERT, and BERT-PT. The average accuracy and F1 scores were approximately 2 and 3.04% higher than the baseline models, respectively. In conclusion, the proposed models provide a more effective

approach for aspect-based sentiment analysis in comparison to the baseline models.

[23] proposes a novel recurrent neural network (RNN) architecture, called Markovian RNN, for nonlinear regression of nonstationary sequential data. In real-life applications, time series data often exhibit non stationarity due to the temporally varying dynamics of the underlying system. The proposed model adaptively switches between internal regimes in a Markovian way to capture the nonstationary nature of the data. A hidden Markov model (HMM) is used for regime transitions, where each regime controls hidden state transitions of the recurrent cell independently. The whole network is jointly optimized in an end-to-end fashion. The proposed method outperforms conventional methods such as Markov Switching ARIMA, RNN variants, and recent statistical and deep learning-based methods through extensive experiments with synthetic and real-life datasets. The inferred parameters and regime belief values are also interpreted to analyze the underlying dynamics of the given sequences.

[24] presents a model, called BG-GCNN, for English toxic comment classification. The model combines bidirectional gated recurrent unit (Bi-GRU) and global pooling optimized convolution neural network (CNN) to treat each type of toxic comment as a binary classification. First, Bi-GRU is used to extract time-series features of the comment, then the dimensionality is reduced through global pooling optimized convolution neural network, and finally, the classification result is output by a Sigmoid function. Comparative experiments show that the BG-GCNN model has better classification performance than Text-CNN, LSTM, Bi-GRU, and other models. The Macro-F1 value of the toxic comment dataset on the Kaggle competition platform is 0.62, and the F1 values of the three toxic label classification results (toxic, obscene, and insult label) are 0.81, 0.84, and 0.74, respectively, which are the highest values in the comparativeexperiment.

Automated content moderation is widely used on social media platforms to promote healthy discussions, and toxic span prediction is a crucial step towards building such systems. In [25], the authors propose a multi-task learning (MTL) model for joint toxic comment classification and toxic span prediction. The model uses ToxicXLMR for bidirectional contextual embeddings and a Bi-LSTM CRF layer for toxic span identification. To enable MTL, the authors curated a dataset from Jigsaw and toxic span prediction datasets. The proposed model outperformed single-task models on both classification and toxic span prediction datasets, with improvements of 4% and 2%, respectively. The model was also evaluated on out-of-domain text from Twitter datasets and showed a 3% improvement in F1 score over single-task models. The proposed MTL model has the potential to enhance the effectiveness of automated content moderation systems.

[26] proposes a deep-learning-based approach for automated scoring of the severity of toxic comments on the Internet using the ELECTRA model. The model uses a downstream regression task accomplished by multi-layer perceptron, convolutional neural network, and attention head layers. The dataset used for training the model is from the Kaggle competition Toxic Comment Classification Challenge, and the model performance is evaluated through another Kaggle competition Jigsaw Rate Severity of Toxic Comments. The proposed approach achieves a competition score of 0.80343, ranking 71/2301 (top 3.1%) in the leaderboard, and can get a silver medal in the competition. The results demonstrate that the proposed method can effectively filter toxic comments and harmful text information on the Internet and significantly reduce the cost of manual review, thus building a healthier online social environment.

III. Proposed Methodology

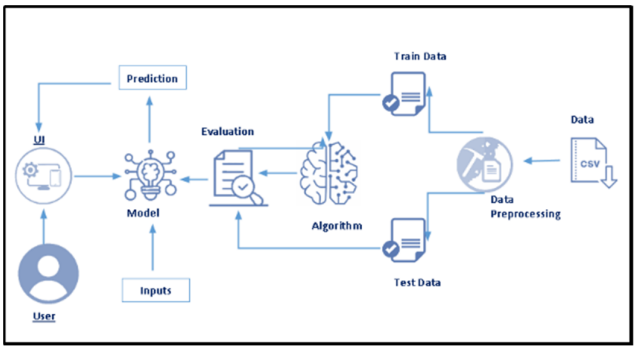
This paper uses the Random Forest classification algorithm to detect toxic comments.

Random Forest is a machine learning algorithm that is used for both classification and regression tasks. It is an ensemble method that combines the predictions of multiple decision trees to make more accurate predictions. Each decision tree is trained on a random subset of the data and features, and the final prediction is determined by taking a majority vote (in case of classification) or averaging (in case of regression) of the predictions from individual trees.
The Random Forest algorithm is particularly useful in handling complex and high-dimensional datasets. It has several advantages, including accuracy, robustness, feature importance.
When it comes to toxic comment classification, Random Forest can be applied to identify and classify toxic comments or text that contains offensive, abusive, or harmful language. The algorithm can be trained on a labeled dataset where each comment is multi labelled across several labels such as toxic, severely toxic, obscene, threat, insult, identity_hate.

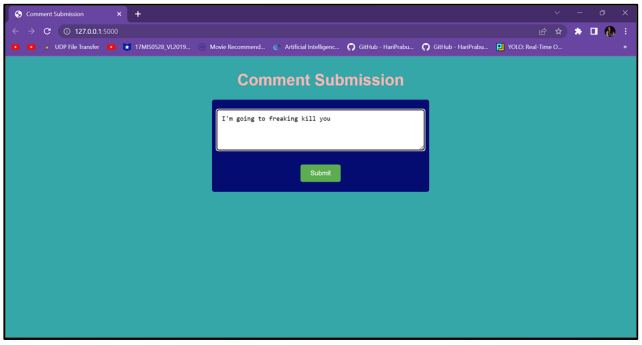
IV. Dataset

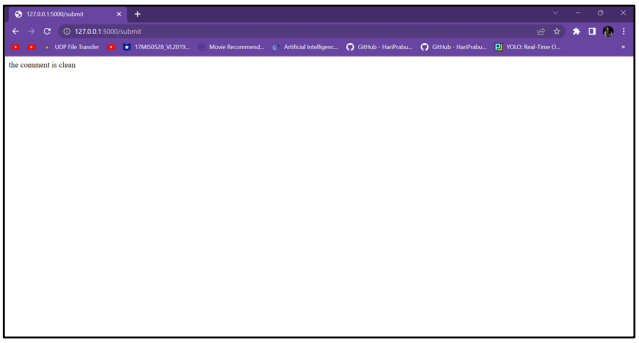
The Kaggle toxic comment classification challenge dataset is a well-known benchmark dataset in natural language processing for detecting different types of toxic comments. The dataset consists of a large collection of comments from Wikipedia's talk page edits, where each comment can be labelled into one or more of the following categories: toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset contains a total of 159,571 comments, out of which 15,594 comments are labelled as toxic, 1,580 comments as severe toxic, 9,572 comments as obscene, 4,001 comments as threat, 8,991 comments as insult, and 1,814 comments as identity hate. The remaining comments are labelled as non-toxic. The dataset poses some challenges for classification models due to the imbalanced class distribution, where the number of comments in some categories is significantly lower than the others. In addition, the dataset contains comments with various levels of toxicity, and it is not always clear what type of toxicity a comment contains, making the classification task more challenging.

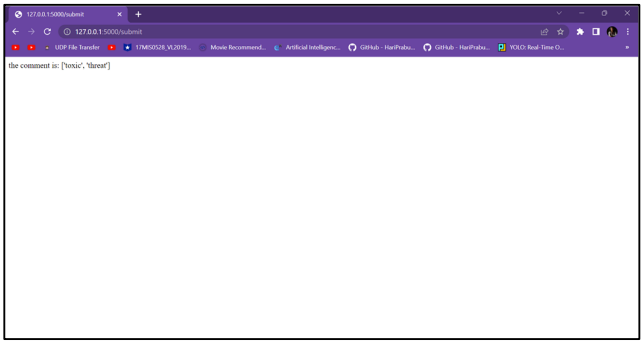
V. Block Diagram

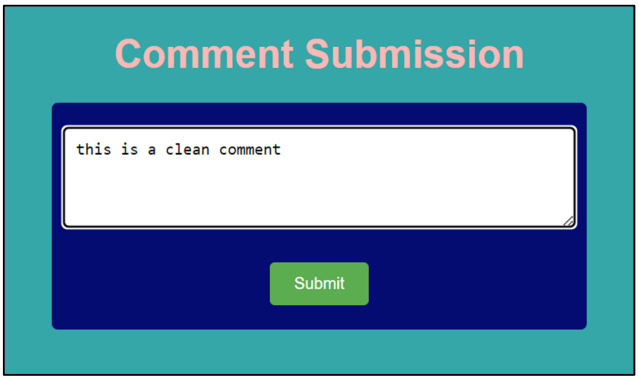


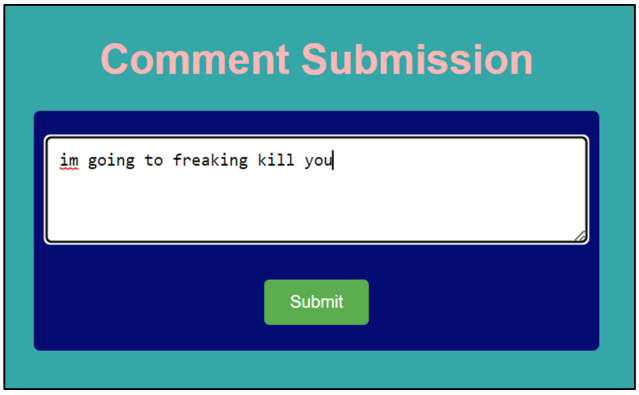
VI. Results

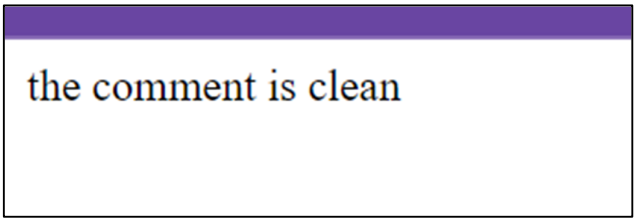

Comment Submission
I'm going to freaking kill you
Submit

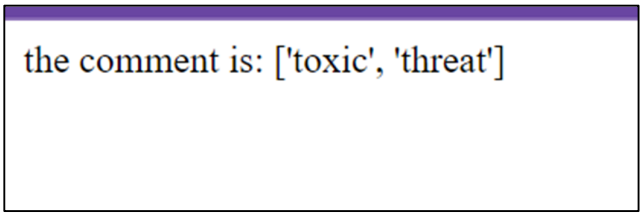

the comment is clean


the comment is: ['toxic', 'threat']

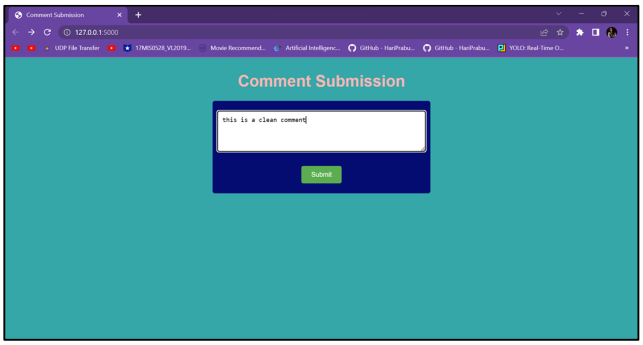

Comment Submission
this is a clean comment
Submit

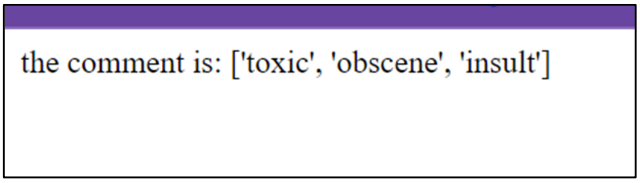

Comment Submission
im going to freaking kill you
Submit


the comment is clean


the comment is: ['toxic', 'threat']


Comment Submission
you are such an annoying idiot
Submit


Comment Submission
this is a clean comment
Submit


the comment is: ['toxic', 'obscene', 'insult']

VII. Advantages and Disadvantages

Advantages:

Promotes healthier online conversations: A toxic comment classifier can help create a more positive and respectful environment in online communities by identifying and filtering out toxic comments. This can encourage constructive dialogue and reduce the spread of harmful content.

Protects users from harassment and abuse: By identifying toxic comments, the classifier can help protect users from harassment, bullying, and online abuse. This is especially important for vulnerable individuals who may be targeted by malicious individuals or groups.

Saves time and resources: Moderating user-generated content can be a time-consuming and resource-intensive task. By automating the process with a toxic comment classifier, moderators can focus on other important aspects of community management, saving time and resources.

Scalability: Online platforms often have a large user base, and manually moderating every comment can be impractical. A toxic comment classifier can be easily scaled to handle a high volume of comments, ensuring efficient moderation even in large communities.

Disadvantages:

False positives and negatives: Toxic comment classifiers may sometimes misclassify comments, leading to false positives (classifying benign comments as toxic) or false negatives (failing to identify toxic comments). This can result in the unintended censorship of legitimate content or the presence of harmful comments that go undetected.

Contextual understanding challenges: Understanding the nuances of language and context can be difficult for a classifier. Sarcasm, irony, cultural references, and other factors can impact the interpretation of a comment. A toxic comment classifier may struggle to accurately identify toxicity in such cases, potentially leading to misclassifications.

Bias and fairness concerns: Machine learning models, including toxic comment classifiers, can inherit biases from the training data or the biases of the individuals labeling the data. This can result in unfair treatment of certain groups or perspectives, amplifying existing biases or marginalizing specific communities.

Adversarial behavior: People who are determined to spread toxic or harmful content may intentionally find ways to bypass the classifier. They may modify their language or use alternative tactics to evade detection, making it challenging for the classifier to stay ahead of malicious actors.

Applications

Social media platforms: Toxic comment classifiers can be used by social media platforms to automatically detect and filter out offensive, harassing, or harmful comments. This helps create a safer and more positive environment for users and discourages abusive behavior.

Online forums and communities: Toxic comment classifiers can be employed in online forums and communities to moderate user-generated content. By identifying and flagging toxic comments, moderators can more effectively enforce community guidelines and maintain a respectful and inclusive space for discussion.

News websites and comment sections: Comment sections on news websites often attract toxic and inflammatory comments. A toxic comment classifier can automatically identify and moderate such comments, ensuring that the conversation remains civil and constructive.

E-commerce platforms: Toxic comment classifiers can be used in product reviews and customer feedback sections on e-commerce platforms. This helps filter out comments that contain offensive language, spam, or malicious content, providing more reliable and helpful information to potential buyers.

Chat platforms and messaging apps: Toxic comment classifiers can be integrated into chat platforms and messaging apps to identify and flag toxic or abusive messages. This can assist in preventing cyberbullying, harassment, and the spread of harmful content in private or group conversations.

Online gaming communities: Gaming platforms often experience toxic behavior in chat channels and multiplayer interactions. Toxic comment classifiers can help identify and address offensive or abusive language, fostering a more positive and enjoyable gaming experience for all players.

Content moderation services: Companies that provide content moderation services to other platforms can utilize toxic comment classifiers as a part of their workflow. These classifiers can assist human moderators in efficiently reviewing and filtering large volumes of user-generated content, ensuring compliance with community guidelines.

Conclusion

A toxic comment classifier can be a valuable tool in promoting healthier online conversations, protecting users from harassment, and improving the overall quality of user-generated content. By automating the identification and filtering of toxic comments, platforms can create a safer and more inclusive environment for their users. However, it's important to acknowledge the limitations of toxic comment classifiers.

The contextual understanding of language and the challenges of bias and fairness are also areas of concern. Additionally, determined individuals may find ways to bypass the classifier, necessitating continuous updates and improvements.

REFERENCES

[1] Hussein, D.M.E.D.M., 2018. A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), pp.330-338.

[2] Zhao, Z., Zhang, Z. and Hopfgartner, F., 2019. Detecting toxic content online and the effect of training data on classification performance. EasyChair.

[3] Dang, N.C., Moreno-García, M.N. and De la Prieta, F., 2020. Sentiment analysis based on deep learning: A comparative study. Electronics, 9(3), p.483.

[4] Joshi, S. and Deshpande, D., 2018. Twitter sentiment analysis system. arXiv preprint arXiv:1807.07752.

[5] Mari, K. and Ganesh, V., 2020. Multi-Labelled Emotion with Intensity Based Sentiment Classification Model in Tweets using Convolution Neural Networks [J]. International Journal of Advanced Trends in Computer ence and Engineering, 9(2), pp.1650-1656.

[6] Zaheri, S., Leath, J. and Stroud, D., 2020. Toxic comment classification. SMU Data Science Review, 3(1), p.13.

[7] Ito, T., Tsubouchi, K., Sakaji, H., Yamashita, T. and Izumi, K., 2020. Contextual sentiment neural network for document sentiment analysis. Data Science and Engineering, 5(2), pp.180-192.

[8] 8. K. Dubey, R. Nair, M. U. Khan and P. S. Shaikh, "Toxic Comment Detection using LSTM," 2020 Third International Conference on Advances in Electronics, Computers and Communications (ICAECC), Bengaluru, India, 2020, pp. 1-8.

[9] Mishra, V. and Tripathi, M., 2022, April. A toxic content detection technique in sentimental analysis with convolution neural networks. In 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 398-402). IEEE.

[10] Özgür, A. and Nar, F., 2020, October. Effect of dropout layer on classical regression problems. In 2020 28th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE. [11] Oriola, O. and Kotzé, E., 2019, November. Automatic detection of toxic south african tweets using support vector machines with n-gram features. In 2019 6th international conference on soft computing & machine intelligence (ISCMI) (pp. 126-130). IEEE. [12] Sushmitha, M., Suresh, K. and Vandana, K., 2022, June. To Predict Customer Sentimental behavior by using Enhanced Bi-LSTM Technique. In 2022 7th International Conference on Communication and Electronics Systems (ICCES) (pp. 969-975). IEEE. [13] Sharma, S., Sharma, S. and Athaiya, A., 2017. Activation functions in neural networks. Towards Data Sci, 6(12), pp.310-316. [14] Rodríguez-Sánchez, F., Carrillo-de-Albornoz, J. and Plaza, L., 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. IEEE Access, 8, pp.219563-219576. [15] Saeed, H.H., Ashraf, M.H., Kamiran, F., Karim, A. and Calders, T., 2021. Roman Urdu toxic comment classification. Language Resources and Evaluation, pp.1-26. [16] Fan, H., Du, W., Dahou, A., Ewees, A.A., Yousri, D., Elaziz, M.A., Elsheikh, A.H., Abualigah, L. and Al-qaness, M.A., 2021. Social media toxicity classification using deep learning: real-world application UK Brexit. Electronics, 10(11), p.1332. [17] Fornacciari, P., Mordonini, M., Poggi, A., Sani, L. and Tomaiuolo, M., 2018. A holistic system for troll detection on Twitter. Computers in Human Behavior, 89, pp.258-268. [18] Georgakopoulos, S.V., Tasoulis, S.K., Vrahatis, A.G. and Plagianakos, V.P., 2020. Convolutional neural networks for twitter text toxicity analysis. In Recent Advances in Big Data and Deep Learning: Proceedings of the INNS Big Data and Deep Learning Conference INNSBDDL2019, held at Sestri Levante, Genova, Italy 16-18 April 2019 (pp. 370-379). Springer International Publishing. [19] Akash, G., Kumar, H. and Bharathi, D., 2021. Toxic comment classification using transformers. In Proceedings of the 11 th Annual International Conference on Industrial Engineering and Operations Management Singapore (pp. 1895-1905). [20] Ozoh, P.A., Adigun, A.A. and Olayiwola, M.O., 2019. Identification and classification of toxic comments on social media using machine learning techniques. International Journal of Research and Innovation in Applied Science (IJRIAS), 4, pp.142-147. [21] Pitsilis, G.K., Ramampiaro, H. and Langseth, H., 2018. Detecting offensive language in tweets using deep learning. arXiv preprint arXiv:1801.04433.