2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) | 979-8-3503-8944-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ACCAI61061.2024.10601796

Promoting Positive Discourse: Advancing AI-Powered Content Moderation with Explainability and User Rephrasing

Ananthajothi K[1], Meenakshi R[2] and Monica S[2]

[1]Associate Professor Computer Science and Engineering, Rajalakshmi Engineering College Chennai, India
[2]Student, Computer Science and Engineering Rajalakshmi Engineering College Chennai, India

E-mail : ananthajothi.k@rajalakshmi.edu.in, 200701143@rajalakshmi.edu.in, 200701151@rajalakshmi.edu.in

Abstract- Nothing is good or bad only thinking makes it so"- is a well-known phrase we might have heard while growing up. Social media and discussion platforms provide unsupervised and an unbound stage to share and discuss ideas, opinions and thoughts. Apart from these they knowingly or unknowingly pave way to a space that generates and harbors toxicity. The contents of these are sometimes targeted on a certain group of people based on gender, caste, religion and countries not keeping in mind the civic sense. This paper introduces an innovative AI-powered system for content moderation, designed to combat online toxicity. The system utilizes a rule-table-based filter for efficient initial screening, blocking content with clear violations. For nuanced toxicity detection, we employ an ensemble of RoBERTa and BiLSTM models. Promoting collaboration, we integrate GPT-3.5 to generate rephrasing suggestions, empow- ering users to modify their language constructively. The system incorporates explainable AI to clarify model decisions, enhancing transparency and understanding. Finally, a user feedback loop ensures the system's ongoing evolution, maintaining its relevance to community standards.

Index Terms—content moderation, toxicity detection, explainable AI, GPT-3.5, RoBERTa, BiLSTM, rule-table architecture, user feedback.

I. INTRODUCTION

Nowadays there is an increase in toxic comments and hostility in the digital space [1]. This provides the people a haven to express all of their thoughts and opinions undefended and unbounded through the networks and many layers, leading to an abundance of debates and toxic comments [2] which sometimes include hateful content with tons of offensive language. This often leads to bullying and confusion in the social platforms [3]. Many at times it is intended to create disruption and confusion among people to jeopardize the harmony by targeting specific people of a class of people, inducing community-based hatred, thus leading to a negative, intolerant and abusive online environment.

The development of a system that can find toxicity in the online social grounds and flagging them, to filter and flag such offensive content is essential for a more positive online social life. This will provide a more comfortable and respectable space where the dignity and good will of an individual person or groups are preserved and respected. Most of the available system for toxicity detection, as the one used by some big platforms like Facebook and Twitter, involves different neural layers like the CNN for determining local patterns, GRU for forming long sequential patterns and additional pooling layers to rule out overfitting. Along with varied word embedders and regularization units like SoftMax, Elastic Net, Dropout [4]. Yet, we see these toxic comments that leads to cyberbullying and dissent. In accordance to our work, RNNs and BiLSTMs [5]. Many models gave better performance in coordination to our work LSTM, RNN+CNN, Deep Neural Networks with word embedders like Word2Vec / GloVe were also used to detect and classify the toxic contents.

The proposed project integrates specialized models for content moderation bias assessment and mitigation. Emojis are analyzed for improved sentiment understanding. The primary objectives encompass real- time contextual toxicity analysis and automated rephrasing

II.LITERATURE REVIEW

The paper [1] aimed to initially find and detected toxicity in online conversation by trying out various commonly used and simple algorithms and natural language processing (NLP) technique. The Deep learning methods that used the above NLP word embedding mechanisms for numerous classifications was implemented that finally produced two datasets. The experiment was divided into primarily two phases: The first one dealt with the studying and classification of the commonly religious toxic comments, and the other one produced a classification based upon race and ethnically targeted toxic comments with numerous embeddings. The results showed that the CNN model produced the best results in both scenarios. The model was not able to comprehend with imbalanced data.

The paper [2], dealt with NLP along with DNNs to identify the solution for the given task of finding the toxic and harmful comments. models such as Word embeddings were implemented in concurrence with RNNs, in addition to the commonly utilized Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN),to determine which model will provide with the desired results. Text classification was done using techniques such as Tokenizing, Stemming, and Embedding. The model used less algorithms, that were actually employed for categorizing online comments based on their toxicity levels. The comments were then compared to determine the accuracy of the model withalternative models such as Long Short-Term Memory (LSTM)and Convolutional Neural Networks (CNN). Also the paper focused more on Commonly used Machine learning using LSTM-CNN and oversampling, Bag of Word (BaG), random undress sampling technique that were used to classify texts. However, the paper was unable to handle over-sampled datasets, resulting in poor results for imbalanced data and the resultant computational complexity was higher.

The paper [3] conducted experiments on social media comments belonging to a particular region, in this case Indonesia. The study aimed to conduct a study to identify toxic sentence-containing comments on social media platforms in Indonesia. There was a pre-trained model that was already implemented for the language spoken there-Indonesian was used. This experiment initially, consisted of multilabel classification was conducted, and the outcomes produced by classification models like Multilingual BERT (MBERT), IndoBERT, and Indo RoBERTa Small were assessed. Other architectures in BERT were used to make it handier to handle problems related to text and NLP. At the data pre-processing stage even the methods to convert informal or colloquial language into standard terms and derive the meaning behind emoticons was used so that it was initialized to find the essence of those words worked upon by the model and convey the correct meaning. The study was concluded by stating that the Indo RoBERTa model displayed a higher accuracy and F1score. However, the research did not consider a dataset that consisted toxic comments including defamation, hate speech, radicalism, and pornography. The study had not considered the validation data and the final value was similar to the value derived from the trained data, this indicated that the proposed model showed signs of overfitting.

In [4] a model was presented, that consisted of a three-tier Convolution Network Model also known as CNN model galvanized by the model AlexNet that was propo studied and acted upon datasets derived from popular search engines. The proposed system initially dealt with cleaning/extracting the user comments by deleting things such as stop words, punctuations and others. One of the models of the Fast text- crawl was implemented to devise the pre-trained word embeddings matrix. The deep learning activation function (ELU) was used in the CNN blocks for a much faster confluence. The technique dropout was done on different layers of the network to prevent overfitting. The determination score ROC-AUC was considered and used as an evaluation metric. Upon analysis, it was seen that it gave a decent score of the accuracy and a pretty average F1 score. This model provided an optimal result with accuracy varying in multi-dimensional datasets.

The paper [5] dealt with finding how generative AI models will perform the task of mitigating social media comments and determine /classify comments based on their nature. Here AI model ChatGPT was used and compared its performance with MTurker annotations for the three frequently discussed topics /contents related to toxic content: Hateful, Offensive, and Toxic also abbreviated as (HOT). The choice of prompts that were used to interact with ChatGPT impacted the performance. The results also suggested that classifications from ChatGPT aligned well with the provided HOT contents. The disagreement of the model aroused due to MTurkers that was not able to give high-quality and useful annotations.

The paper [6], discusses the need for the use of DNN. DNNs are considered the best approach or method the for harmful comments detection using text data. The DNN compared the toxic comments detection with many classical features present like BoWV-known as Bag of word vectors and (GloVe) and Term Frequency-Inverse Document Frequency (TF-IDF) values. It was found that the DNN methods significantly outperformed the commonly existing methods like CNNs and RNNs. The classifiers that were used had the recognition of being top-performing models. Apart from that the BERT-Fine tuning approach was used along with DNN classifiers. The Binary Classification Task and Regression Task were implemented. Word embedding models such as Mikolovs word embedding, fast Text sub word embedding was used. These were compared against transformers based on BERT finetuning. Additionally, it was noticed that the CNN architecture was used compare against Bi-LSTM and Bi-GRU classifiers. The final conclusion was that BERT model is efficient at the crux of hate speech detection. Among the various DNN based classifiers, it was seen that bi- LSTM performs better than any of the models. The model however though did not support multi-class classification and most of the models used were susceptible to appending attacks.

Machine learning models were used in [7] to revamp the accuracy of the comment toxicity detection systems. It was done using commonly used unsupervised methods that showed dependence on the state-of-the-art models and other matters such as external and internal embeddings to improvise the accuracy while keeping in mind the bias. Logistic Regression Model and Neural Network Model were used to classify the comments. A much higher accuracy and a seemingly F1 score was achieved using an ensemble BERT and LSTM model. The accuracy and FI score was relatively less compared to other machine learning and neural network models. Also, the models were not able to predict and highlight toxic comments and did not support multi- dimensional data.

The model in [8] extensively explored various propositions for finding morbidity in texts which were then classified, assessed and displayed, with the purpose of enhancing the present total quality of texts classifications. It relied on a hybrid model of (LSTM) with Glove word embeddings and another one of LSTM with word embeddings that were generated and also generated (BERT). These were then used to train and test on a large number of datasets that was already classified as toxic and non-toxic. Upon execution it was noted that the combination of LSTM and BERT together performed better than the previous models mentioned. As many Machine Learning models and NLP models were used, errors were found in toxicity identification because of inconsistent data. It was noted that Sub word embeddings were not included, which eventually reduced F1 score and accuracy.

The paper [9], initially started off with comparing a diverse array of models applied to a complex multi-labeled hate speech dataset. The comments were compared using models in natural language processing that included the likes of BERT, ResNet and Vision Transformers. The trained models for three epochs were done using (pwBCE) loss function that stands for positive weighted Binary Cross Entropy. Apart from that, the model implemented various transformer-based models, - BERT, DistilBERT, AlBERT etc, solely for the purpose of classification of the online chats into toxic and non-toxic. The same methods and data were used throughout the analysis to focus on performance, bias detection and inference. The research showed that the models BERT, RNN, and XLNet had a very similar performance. RNNs were seen to be much faster at inference compared to many of the BERT tested. It was noted that, the model DistilBERT combined a good classification performance with a low inference time per dataset batch. The proposed model exhibited bias in associating identities with toxicity, potentially leading to incorrect classifications and harm, and the use of Convolution-based models like CNN and CCT exhibited more sensitivity to association bias between identities and insults.

In [10], the proposed system aimed to address the semi-supervised toxicity detection by addressing the gap by creating a fair semi-supervised hybrid model designed to traverse social bias in the classification of harmful text. At the outset, the proposed model primarily comprised two components: firstly, a trained semi-supervised text classification conducted on benchmark toxicity datasets. Then, in the successor step, the model traversed the various social bias present in the trained classifier during the initial stage. to improve fairness. After the initial steps two different generative-based semi- supervised text classification models, were used namely NDA- GAN and GANBERT, following this, fair semi-supervised models known as FairNDAGAN and FairGANBERT were incorporated. These models were subsequently compared with baselines in terms of accuracy and fairness to elucidate the challenges related to social fairness in semi-supervised clas- sification of harmful and toxic text. The objective was to demonstrate the vulnerability of both supervised and semi- supervised models with regards to both accuracy and fair- ness. The models undergo two-step training: pre-training on benchmark toxicity datasets and post-training with adversarial debiasing to mitigate gender and race demographic biases. Expanding the amount of unlabelled data did not show a direct correlation with enhanced model performance or fairness. Through the experiments, FairNDAGAN and FairGANBERT models presented a significantly fairer semi-supervised frame- work that consistently surpassed their non-fair semi-supervised counterparts in terms of fairness. This highlighted the potential weaknesses of semi-supervised models when confronted with imbalanced datasets, emphasizing the importance of integrat- ing data augmentation methods and implementing balancing techniques in scenarios where similar results were observed. The model however failed to mention that these models were sensitive to highly imbalanced datasets, limiting their use in practical scenarios with such data. Fairness was assessed only for gender and race demographics, leaving out other social groups. Apart from that the models performed differently for various demographic groups, indicating the necessity of tailored strategies or bias-mitigating methods.

The paper [11], handles the issue of social media comments and texts targeted towards a particular community such as Blacks, Muslims, LGBTQ+ etc which consists of only toxic statements taken with the appropriate context. To tackle the above issue the paper focuses on deep learning methods and natural language processing embeddings with the dataset initially being classified into two sets – one focusing on the harmful social media comments and the other one being ethnicity or race-based comments using jigsaw unintended bias in toxicity classification. Standard machine learning algorithms such as Support Vector Machines (SVM), Random Forests (RF), Naive Bayes (NB), and Logistic Regression (LR) rely on manually crafted features and cannot inherently extract contextual information from toxic text. These algorithms typically operate on predefined numerical or categorical features derived from the text, such as word frequencies, TF-IDF scores, n-grams, or other linguistic attributes. Widely used deep learning models like RNN along with LiSTM and GRU. Apart form this the model fusion technique was also used to represent two different approaches and develop two different classifier techniques. The conclusion was that the pretrained data models with DL techniques showed higher accuracy.

The paper [12], deals with multifaceted components pertaining to toxic and harmful comments such as cursing, harassment and extremism. Probabilistic models such as R esNet, Inception and BERT were used and existing models such as HOLE, TRANS-E, TRANS-H were also considered that can generate embeddings from a knowledge graph. The model emphasized the importance of conducting multi-level data analysis, which encompasses content, individual, and community perspectives, along with the diverse array of features required to assess toxicity but did not provide a definite conclusion on which models provided the desired output.

The model prescribed in [13], consisted of various deep learning models consisting of a hybrid model of LSTM and CNN along with word and character embeddings. The CNN layer further dealt in such a way that it incorporated words and characters differently to reach a conclusion such as word-based convolution neural network and character based convolutional neural network. The model also performed comparative analysis in terms of performance on the trained models using the Gradio app. It anticipates the toxicity level and categorizes the comment into different levels of toxicity, presenting this classification in the output section. The analysis of the model went as follows, Neural network models underwent training and testing using real-world communication comments and tweets. Balancing techniques were addressed in imbalanced text datasets. Prior to processing, comments were refined to eliminate stop words, retaining only the most meaningful words, while Lemmatization was applied to unify words into their base forms.All the methodologies used here showed an impressive score and clearly the hybrid model performed better than the rest.

The paper [14], proposed a machine learning solution aimed at detecting toxic images by leveraging embedded text content. The methodology outlines the utilization of Long Short-Term Memory (LSTM) Gated Recurrent Unit (GRU) models and their Bidirectional variants. The model commenced with a review of related works, encompassing data processing methodologies and machine learning techniques, to setup for identifying 'trolls' and toxic content. The framework used was adaptable for training models with varied datasets and labels. The proposed methodology consisted of the creation of two modules: an image text extraction module and a text classification module. To extract text from images, it utilized the Python-tesseract module, renowned for its OCR capabilities. Tesseract stands out as one of the most widely used OCR engines, ensuring precise character extraction from images for text classification, a deep learning model tailored for detecting toxicity within the text embedded in the image was selected. The chosen models for this task were recurrent neural networks: Bidirectional LSTM and Bidirectional GRU. The utilization of the word embeddings increased the accuracy, apart from that the models showed a slight dip in accuracy while extracting the images.

The paper [15], proposed a simple and commonly used deep learning models and techniques to counterattack toxic media comments along with the novel application of natural language processing technique. It targeted the social media comments into four categories namely – Pornography, hate speech, radicalism and hate using the SVM method. For the model training process LSTM was used along with various NLP libraries and tools. The proposed approach involved leveraging Bidirectional Encoder Representations from Transformers (BERT) to categorize toxic comments within user-generated content, such as tweets. The BERT-base pretrained model was fine-tuned using a reputable labeled dataset of toxic comments sourced from Kaggle's public datasets. Evaluation results demonstrated BERT's capability to accurately classify and predict toxic comments with a notably high accuracy rate. The BERT-base model exhibited superior performance across all compared models, achieving the most favorable outcomes.

III.METHODOLOGY

The system though separated as different environments, on the whole includes five major packages which in turn consists of several modules. These packages are highly isolated with the help of docker containers so as to support hassle free maintenance, updation and scalability. Hence the development phase doesn't take the environment separation into consideration, for each of the models is developed as different containers.

The first environment is the active environment that deals with the real-time fetching of data and using an already pre- trained core BERT model, which will assign toxicity scores for the input. The toxicity score is then used by the rephrasing and the reinforcement modules for further actions. The secondary environment is a passive environment that deals with the batch-wise learning and training of the core model and also includes the Bias modules to continually check for possible biases in the model demographic attributes.The above Figure.1, presents the architecture of the entire system in a detailed manner.

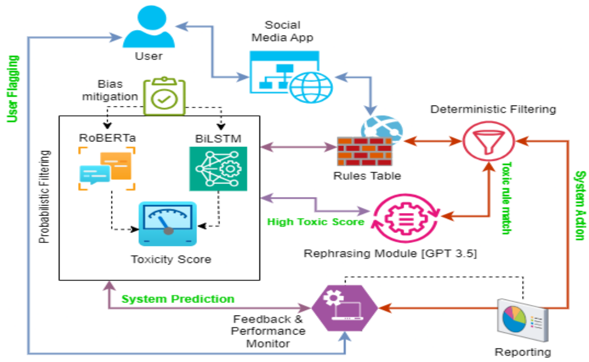


Fig. 1. Architecture Diagram of the system

A. *UI MODULE*

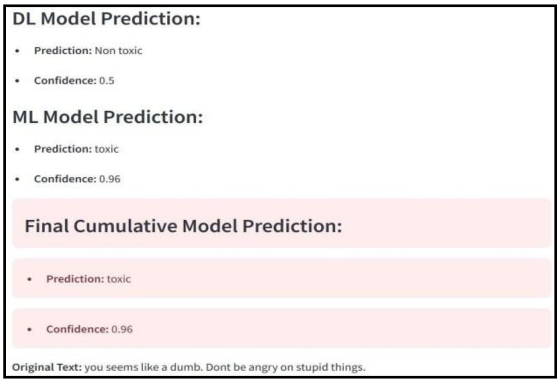
The UI module WILL have a similar structure to that of the social media platforms, with basic sections like profile, settings, etc. The home page should have the recent posts, and also the user will have an option to post something new or to comment on the existing posts. If the user is going to post something, he will be provided with ability to input texts, emojis, emoticons and images. The comment section is restricted for text and emojis or emoticons. Additionally, a menu containing the reactions like like, dislike, love, etc. can be added for users to react to the post without comments.

B. *DATA INGESTION MODULE*

This module is the most extensive part of the entire system as it includes 2 types of data pipeline, one dealing with online learning and the other for real time data streaming. Despite this both the pipeline modules consist of a standard step of processes. At first, the raw data input is processed and cleansed and then it is sent for tokenization. Since the core module is a transformer-based model, the tokenizer is chosen to be compatible with it. Hence BERT tokenizer is the primary choice along which customization is done to conserve non-ascii characters, slangs, masked and emojis which are usually ignored during tokenization. The data will be then embedded with the ensemble of GloVe, BERT and Emo2vec. The ensemble is chosen so as to increase the accuracy of the model. Though BERT works exceptional in handling textual information based on the context, emojis are ignored. Hence, this system first embeds the information with GloVe + Emo2vec layer that produces non contextual embeddings and then a secondary layer of BERT embedder for a contextual embedding. After that the data is sent to the core module. Data pipeline also has the other components like Cloud Stack, storage unit and DBT along with Apache Airflow and docker to automate and ease the data cleansing, transformation and storage. Unlike the traditional storage techniques, this module increases the ease of scalability.

C. *CORE MODULE*

The core prediction module will contain the Roberta+logistic regression and Bi-LSTM model that will take in the content and will check for the presence of toxicity in any forms like [hate, profanity, offensive, sarcasm, dark humor, etc], as shown in Figure.2 the user comment has been flagged as toxic in nature. For this package the accuracy should be greater than 92%. The core module should also check for any unintended biases and use ways to avoid it [Training or testing biases]. Techniques like special dropout can be adopted for the Bi-lstm layer to avoid overfitting. The weighted average method is used to get the final output, which is based on the accuracy of the two models BiLSTM module and the Roberta+logistic regression module (ensemble).



DL Model Prediction:

- **Prediction:** Non toxic
- **Confidence:** 0.5

ML Model Prediction:

- **Prediction:** toxic
- **Confidence:** 0.96

Final Cumulative Model Prediction:

- **Prediction:** toxic
- **Confidence:** 0.96

Original Text: you seems like a dumb. Dont be angry on stupid things.

Fig. 2. Study of the social media comments.

D. *THE HUMAN IN THE LOOP MODULE*

This module will focus on voting mechanism. First the user whose content has been flagged toxic, if they feel the model's decision is wrong will downvote the outcome of the model. The downvotes will be moved to the other section of the application [web page] which is visible only to the admin page, will have the humans to review the ai actions and if a majority of the downvotes are acceptable, then the human mode can push the core predictor for a re- training phase.

E. *REPHRASING MODULE*

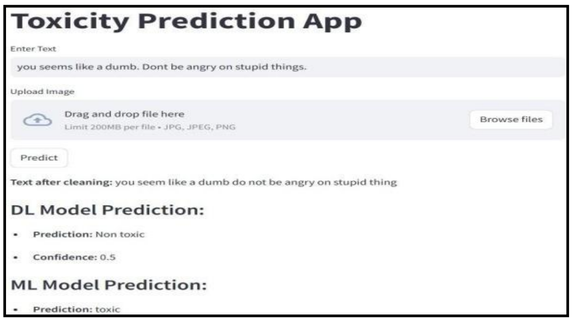
This module will be triggered when the output of the core module is toxic. The rephrasing module will be using the GPT based models, to rephrase the given content to remove toxicity in it, while preserving maximum information. There must be 2-3 options presented before user in the UI, and the user will select the suitable ones. When the user uses the ai provided options, it can be directly posted, without another cycle of toxicity analysis as shown in Figure.4 the final prediction of the model will determine the comment to be toxic or non-toxic and notify the user.

F. *REINFORCEMENT MODULE*

This module, will be provided with the decision of the model, along with the explanation for the decision. The human user, must vote if the model's reasoning is valid or not. Based on the majority voting mechanism, the feedback module will be triggered. This module ensures the adaptability of the model, as certain context will be “toxic” for one period and not in other timelines. Through reinforcement such changes in the model, is updated timely, for a more stable and reliable system. This visualization also reflects the effects of the data augmentation performed during the processing stage.

IV.Result And Analysis

The implementation of multiple models of the system along with the analysis is provided in this section. The visualization helped in providing valuable insights for developing models and fine tuning them. This section exhibits the nature of the data ingestion module and how it contributed to the model configuration is also highlighted. The Figure.3 shows the social media comment that the user had entered and the comment being toxic in nature. This then undergoes to the packages that will then provide the user with the rephrased text as shown in Figure.4, that will finally allow the user to enter their comment.



Toxicity Prediction App

Enter Text: you seems like a dumb. Dont be angry on stupid things.

Text after cleaning: you seem like a dumb do not be angry on stupid thing

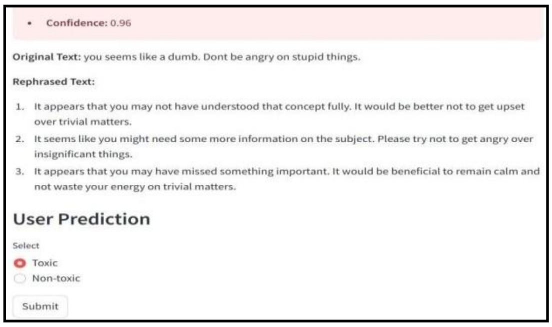
DL Model Prediction:

- **Prediction:** Non toxic
- **Confidence:** 0.5

ML Model Prediction:

- **Prediction:** toxic

Fig. 3. Study of the social media comments.



- **Confidence:** 0.96

Original Text: you seems like a dumb. Dont be angry on stupid things.

Rephrased Text:

1. It appears that you may not have understood that concept fully. It would be better not to get upset over trivial matters.
2. It seems like you might need some more information on the subject. Please try not to get angry over insignificant things.
3. It appears that you may have missed something important. It would be beneficial to remain calm and not waste your energy on trivial matters.

User Prediction

● Toxic
○ Non-toxic

Fig. 4. Production of Rephrased comments.

V.Conclusion

In this paper, we have discussed about the problem statement and objectives for the proposed system, and have constructed a complete system that would providea multi-featured solution to the existing problems in the landscape of today’s online forums. There is a growing need for a system, that could not only determine and filter out the toxicity of the online social platform, but also preserve the ethical rights of the users. We have proposed an all-rounder application, that could perform the task in an unbiased and streamlined manner. It is been noted that the functioning of the system involves multiple technologies from containerization, cloud computing to Deep learning. Hence, the complexity of the system is much higher and demands more computational units.

References

[1] Abbasi, A. R. Javed, F. Iqbal, N. Kryvinska, and Z. Jalil, “Deep learning for religious and continent-based toxic content detection and classification,” Sci Rep, vol. 12, no. 1, p. 17478, Oct. 2022, doi: 10.1038/s41598-022-22523-3.

[2] K. Poojitha, A. S. Charish, M. A. K. Reddy, and S. Ayyasamy, “Classification of social media Toxic comments using Machine learning models,” 2023, doi: 10.48550/ARXIV.2304.06934.

[3] G. Z. Nabiilah, S. Y. Prasetyo, Z. N. Izdihar, and A. S. Girsang, “BERT base model for toxic comment analysis on Indonesian social media,” Procedia Computer Science, vol. 216, pp. 714–721, 2023, doi: 10.1016/j.procs.2022.12.188.

[4] I. Singh, G. Goyal, and A. Chandel, “AlexNet architecture based convolutional neural network for toxic comments classification,” Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 9, pp. 7547–7558, Oct. 2022, doi: 10.1016/j.jksuci.2022.06.007.

[5] L. Li, L. Fan, S. Atreja, and L. Hemphill, “‘HOT’ ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media”.

[6] A. G. D’Sa, I. Illina, and D. Fohr, “Towards Non-Toxic Landscapes: Automatic Toxic Comment Detection Using DNN”.

[7] M. A. Saif, A. N. Medvedev, M. A. Medvedev, and T. Atanasova, “Clas- sification of online toxic comments using the logistic regression and neu- ral networks models,” presented at the PROCEEDINGS OF THE 44TH INTERNATIONAL CONFERENCE ON APPLICATIONS OF MATH- EMATICS IN ENGINEERING AND ECONOMICS: (AMEE’18), Sozopol, Bulgaria, 2018, p. 060011. doi: 10.1063/1.5082126.

[8] C. I. A. Neuroscience, “Retracted: An Automated Toxicity Classification on Social Media Using LSTM and Word Embedding,” Computational Intelligence and Neuroscience, vol. 2023, pp. 1–1, Feb. 2023, doi: 10.1155/2023/9850820.

[9] C. Duchene, H. Jamet, P. Guillaume, and R. Dehak, “A benchmark for toxic comment classification on Civil Comments dataset.” arXiv, Jan. 26, 2023. Accessed: Nov. 13, 2023.

[10] S. Shayesteh, “Social Fairness in Semi-Supervised Toxicity Text Clas- sification”.

[11] Deep learning for religious and continent-based toxic content detection and classification” Ahmed Abbasi1, Abdul Rehman Javed, Farkhund Iqbal, Natalia Kryvinska & “Defining and Detecting Toxicity on Social Media: Context and Knowledge are Key” Amit Sheth, Valerie L. Shalin, Ugur Kursuncu

[12] Toxic Comment Classification using Deep Learning. B.Ramesh Naidu, Naresh Tangudu, Ch. Chandra Sekhar.

[13] Using deep learning to detect social media ‘trolls’Áine MacDermott a, Michal Motylinski , Farkhund Iqbal , Kellyann Stamp , Mohammed Hussain , Andrew Marrington / DFRWS 2022 APAC - Proceedings of the Second Annual DFRWS APAC

[14] CLASSIFICATION OF TOXIC COMMENTS USING DEEP LEARNING Chaitanya Sonawane*1, Preeti Kawade*2, Tejaswini Bagale*3, Swarada Ogale*4,