International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 2065- 2073
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0282

Purging the Poison: A Machine Learning Approach to Filtering Toxic Comments

Vijaya R. Pawar[1], Simantini. D. Garud[2], Anuja. A. Kadam[3], Aishwarya. G. Khairnar[4]
[1-4]Department of Electronics and Telecommunication Engineering, Bharati Vidyapeeth's College of Engineering for Women, Pune, Maharashtra, India.
Email ID: *vijaya.kashid@bharatividyapeeth.edu[1], simantinigarud403@gmail.com[2], anujakadam1707@gmail.com[3], aishwaryakhairnar444@gmail.com[4]*

Abstract
The rapid growth of online communication platforms has provided unprecedented opportunities for global dialogue. Yet, it has also introduced challenges such as the proliferation of toxic comments, which can have severe consequences for individuals and communities. This research paper proposes a machine learning-based approach to mitigate the impact of toxic comments by automatically identifying and filtering them from online discussions. Our study begins by curating a comprehensive dataset of labeled comments, encompassing a spectrum of toxicity levels. Leveraging state-of-the-art natural language processing techniques, we extract relevant features from the textual content, including sentiment, context, and linguistic patterns. These features serve as inputs to a machine learning model, trained on a diverse range of toxic and non-toxic comments. In conclusion, this research contributes to the development of intelligent content moderation systems that foster healthier online discourse. By implementing machine learning algorithms, we aim to provide a scalable and effective solution for identifying and filtering toxic comments, ultimately promoting a more inclusive and respectful online environment.

Keywords: *Machine Learning, Feature Extraction, Negative Comments, Training Data, Toxic comment, Nontoxic comments*

1. Introduction
Social media serves as a platform bustling with diverse discussions, where anonymity empowers individuals to voice their opinions without restraint freely. In the nascent stages of the internet, email was the primary mode of communication, yet it was inundated with spam, making it challenging to differentiate between genuine and unsolicited emails. As the internet landscape has evolved, particularly with the emergence of social networking platforms such as Facebook and Reddit, the need to classify posts as either positive or negative has become increasingly vital. This is essential to prevent societal harm and shield individuals from participating in detrimental or antisocial conduct. Such toxic comments, whether they are threatening, obscene, insulting, or rooted in identity-based hatred, present a significant risk of online abuse and harassment. Consequently, individuals may refrain from expressing their opinions or seeking alternative viewpoints, leading to unhealthy and biased discussions. This, in turn, makes it challenging for various platforms and communities to foster equitable conversations, often prompting them to either restrict user comments or cease them entirely, ultimately undermining their viability. [1] In summary, this research contributes to the ongoing efforts to create safer digital spaces by leveraging the capabilities of machine learning. By exploring novel approaches to filter toxic comments, we aim to enhance content moderation strategies and promote a more positive and respectful online discourse. This study aims to explore and implement advanced machine learning models to accurately identify and categorize toxic comments, contributing to the creation of safer and more inclusive online environments. [5]

2. Literature Review
A review of the work carried out by the researchers

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 2065- 2073
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0282

in the area of depression detection and its analysis is done in detail. Cognizance of that work is presented here. Rahul and H. Khajla et al. [7] Observed that a pivotal strategy for augmenting the accuracy of a trained random forest classifier resides in the meticulous management of class imbalances within the training dataset. Various scholarly works and academic discourses underscore the paramount importance of conscientiously addressing these imbalances to foster a discernible amelioration in the predictive prowess of the model. By intricately tending to the equilibrium of class representation during the training phase, the model not only augments its discriminative capacities but also fortifies its ability to generalize across diverse instances. [3] These findings, elucidated by esteemed researchers, underscore the imperative nature of harmonizing class distribution for the overarching enhancement of the random forest classifier's performance. N. Chetty, S. Alathur et al. [4] In simpler terms, this paper investigates how artificial intelligence (AI) systems can help detect and analyse online hate content, which often contributes to communal violence. [5] By searching for relevant articles using specific keywords, the study gathers information from various sources. The literature review shows that social media platforms can utilize AI systems to identify and understand hate speech online. [6] Furthermore, the paper explores how cognitive processes influence both the perpetrators and victims of such content. It also discusses the challenges in managing online hate speech. Ultimately, the paper suggests that building effective AI systems and fostering healthier cognitive processes among individuals can help reduce hate content online. M. Husnain, A. Khalid et al. [12] This study employs two methods to identify different types of toxicity in comments. The first method trains separate classifiers for each type of toxicity, while the second method treats the problem as a multi-label classification task. [8] Various machine learning algorithms, such as logistic regression, Naïve Bayes, and decision trees, are used for analysis. The dataset is sourced from Kaggle, and 10-fold cross-validation is used to assess the model's robustness. A unique pre-processing technique is applied to transform the multi-label classification problem into a multi-class classification one, resulting in improved accuracy. Experimental findings reveal that logistic regression performs well in both binary and multi-class classification, suggesting the potential effectiveness of the pre-processing approach for neural classification models. Ashish, A. Rani et al. [2] This study aims to overcome these challenges by creating a more reliable toxic comment classification system. It plans to enhance the model's ability to recognize subtle toxic language by incorporating additional context and using techniques like adversarial training and data augmentation to introduce more diversity in the training data. Additionally, the study intends to test the model on various real-world datasets to ensure its effectiveness outside of controlled environments. [9-11] S. Smetanin et al. [15] This study examined the prevalence of toxic comments across various topics in Russian-language comments on the social network Pikabu. Firstly, we manually labeled a training dataset and fine-tuned multiple language models to classify toxic comments. We then made our pre-trained models publicly available to aid future toxic comment research. Secondly, we developed a method for labelling topics based on six key dimensions used by governmental and intergovernmental organizations to measure objective wellbeing. Finally, we analysed Pikabu data and discovered that the highest proportion of toxic comments occurred in discussions about politics, followed by security and socioeconomic topics. Other topics showed similar levels of toxic comments. [13,14] T. A. Belal et al. [20] This paper introduces a deep learning approach to categorize Bengali toxic comments. Initially, a binary classification model is employed to determine whether a comment is toxic or not. Subsequently, a multi-label classifier is used to identify the specific type of toxicity present in the comment. The dataset used in this study consists of 16,073 instances, with 8,488 labeled as toxic. Toxic comments may belong to one or more of six toxic categories: vulgar, hate, religious, threat, troll, and insult. [17-19] The binary classification task achieved 89.42% accuracy using Long Short Term Memory (LSTM) with BERT Embedding, while the multi-label classification task

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 2065- 2073
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0282

reached 78.92% accuracy and a weighted F1-score of 0.86 using a combination of Convolutional Neural Network and Bi-directional Long Short Term Memory (CNN-BiLSTM) with attention mechanism. To interpret the models' predictions and understand the importance of words in classification, the Local Interpretable Model-Agnostic Explanations (LIME) framework was utilized. H. R. Sifat et al. [26] The paper's results showed that a language model which was bidirectional trained can have a deeper sense of language context and flow than single-direction language models. In the paper, the researchers introduced a novel technique called Masked Language Model (MLM), enabling bidirectional training in previously unfeasible models. They utilized BERT embedding and incorporated them into a Transformer layer, followed by a Dense layer with 500 neurons. To prevent overfitting, a Dropout of 0.1 was applied. Keeping parameters consistent with previous models, they achieved an accuracy of 93.24%. However, one test case was misclassified, possibly due to training the model for 50 epochs, similar to other models. A. S. Kapse; A. Dubey et al. [28] The main objective of this research study is to detect and classify toxic comments on social media platforms, including those containing hate speech, abusive language, obscenities, threats, and insults. However, a significant challenge arises when dealing with datasets containing comments in multiple languages. [21,22] In such cases, the initial step in developing deep learning algorithms involves identifying the language of each comment before proceeding with toxicity detection. This language detection step is crucial for accurately analysing and categorizing comments as toxic or non-toxic. K. Machova, T. Tomcik et al. [24] The paper concentrates on identifying different types of toxic comments on social media platforms, with a specific focus on offensive language, hate speech, and cyberbullying. The dissemination of toxic content via social networks poses a significant challenge, potentially disrupting the functioning of democratic societies. The study conducts experiments using various machine learning techniques to determine the most effective approach for building recognition models. Specifically, it compares deep learning methods with ensemble learning to assess their suitability for this task. K. A. Kumar et al. [16] The primary objective of this paper is to determine whether a given comment can be classified as toxic or non-toxic using various machine learning techniques. [25,27] The study employs six different traits to analyse comments, and a dictionary is created using vectorization of known vocabulary (dataset) to train the machine learning model. Since multiple traits are considered, the model undergoes training multiple times against each trait to assess its performance. The research reveals that the Random Forest algorithm demonstrates strong performance across all traits, achieving an accuracy of 85% with a precision of 91%. Unlike previous studies focused on demographic or local languages, this research focuses on developing a classifier specifically for the English language. J. Roy et al. [23] This paper examines the impact of machine translation on automated toxic comment classification using the Google Perspective API. It tests comments from non-English Wikipedia talk pages in five languages, translating them into English. Results show high consistency in classification for French, Italian, and Spanish comments, but lower accuracy for Portuguese and Russian comments. The study underscores the influence of language on translation accuracy. [29,30]

3. Methodology

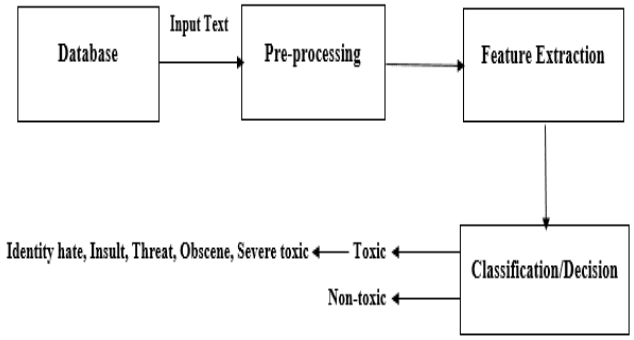

Figure 1 Schematic for a Deep Learning Approach to Filtering Toxic Comments

3.1. Step 1 Database

This work uses the 159572 comments samples from the Kaggle database. There are 8 columns and rows are 159572. The number of samples for each toxicity

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 2065- 2073
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0282

level class is Toxic: 15294, severe toxic: 1595, obscene: 8449, threat: 478, insult: 7877, and identity hate: 1405. The database typically consists of labeled comments where each comment is labeled as either toxic or non-toxic. This database is used for training and evaluating machine learning models to accurately classify new comments as toxic or non-toxic (refer figure 1)

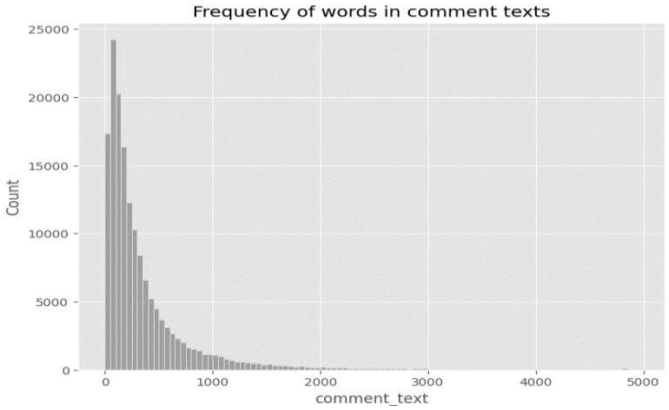

Figure 2 Frequency of Words in Comments Texts

3.2. Step 2 Preprocessing

We'll start by fetching raw text data from Kaggle. Then, we'll clean it up by removing commas, full stops, and other punctuations. Next, we eliminated common words known as stop words. Then, we applied stemming and lemmatization techniques to reduce words to their root form. Finally, we used count vectorization to convert the cleaned text data into a numerical format. Upon analyzing the cleaned data, we discovered a dataset comprising 159572 samples of comments along with corresponding labels, which can be imported from the train.csv file. In toxic comments classification using machine learning, preprocessing involves several steps to clean and standardize the text data before training the model. This typically includes tokenization, lowercasing, removal of punctuation, stop words, and handling special characters, emoji's, URLs, and numerical digits.

3.3. Step 3 Feature Extraction

The Word Cloud & TF-IDF is used for feature extraction of Toxic & Non-Toxic comment. It consists of toxic or non-toxic word in one single comment. Word clouds are generated to visualize the most frequently occurring words in comments for different toxicity levels. These word clouds provide a visual representation of the most. Although not explicitly mentioned in the provided code snippet, TF-IDF vectorization is a common feature extraction technique used in text classification tasks. It transforms text data into numerical vectors based on the frequency of words in documents, weighted by their inverse document frequency. common words used in toxic, severe toxic, obscene, and insult comments. Word Clouds for top occurring words: (table 3-6)

Figure 3 Severe Toxic Comment

Figure 4 Toxic Comments

Figure 5 Obscene Comments

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 2065- 2073
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0282

Figure 6 Insult Comments

3.4. Step 4 Classification / Decision

The classification/decision module serves as the final stage in the process, where the model evaluates the extracted features to determine whether a given comment is toxic or non-toxic. This module utilizes machine learning algorithms such as support vector machines, Logistic Regression, Naive Bayes (Multinomial), Random Forest Classifier. By leveraging the extracted features and the learned patterns from the training data, the classification/decision module makes informed predictions, assigning a label of toxic or non- toxic to each input comment, thereby enabling effective filtering of negative or harmful content. In toxic comments classification, the work of classification or decision-making involves training a machine learning model to analyze text inputs and predict whether they contain toxic language or not. This typically involves preprocessing the text, extracting relevant features, selecting an appropriate model, and then training it on labeled data to make accurate predictions.

4. Algorithm

4.1. Logistic Regression

Logical regression analyzes the relationship between one or more independent variables and classifies data into discrete classes. Equation of Logistic Regression:

$$f(x) = \frac{1}{1 + e^{-x}}$$

4.2. Naive Bayes (Multinomial)

Naive Bayes classifiers are probabilistic classifiers based on Bayes' theorem. The Multinomial Naive Bayes variant is suitable for classification with discrete features, such as word counts in text classification tasks. The formula for Bayes' theorem is given in Figure 7.

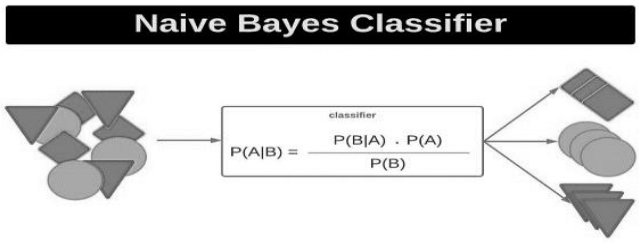

Figure 7 Naïve Bayes Classifier

where,

- PA= the prior probability of occurring A
- PBA= the condition probability of B given that A occurs
- PAB= the condition probability of A given that B occurs
- PB= the probability of occurring B

4.3. Random Forest Classifier

Random Forest is a supervised machine learning algorithm made up of decision trees as shown in figure 8. Random Forest is used for both classification and regression. for example, classifying whether an comment is “Toxic” or “Nontoxic”. It's known for its high accuracy and robustness.

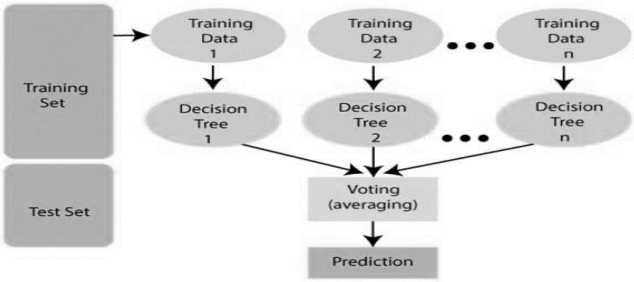

Figure 8 Random Forest Classifier

4.4. Support Vector Machine (SVM)

Support Vector Machine is a powerful supervised learning algorithm used for classification tasks. It's commonly used in text classification tasks due to its ability to handle high-dimensional feature spaces. Ex: SVM algorithm can be used for Face detection, image classification, text categorization, etc. (Fig. 9)

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 2065- 2073
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0282

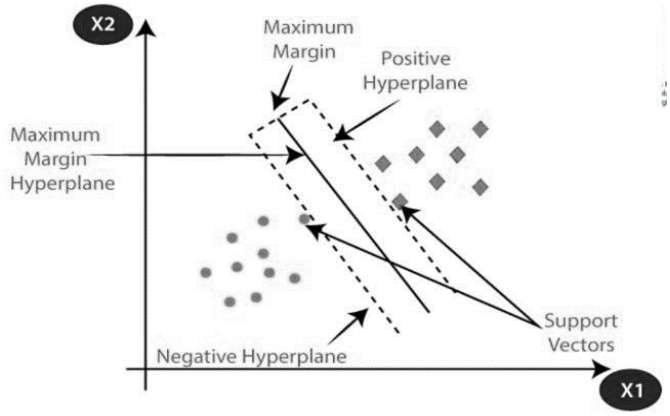

Figure 9 Support Vector Machine (SVM)

Random Forest Classifier and Support Vector Machine (SVM) are types of algorithms used in machine learning. They're often used to solve problems like classifying text, which means figuring out if a piece of text is toxic or not, for example. The code you provided seems to be preparing the data for this kind of task. It's cleaning up the text, figuring out which words are important, and getting everything ready to teach a computer to recognize toxic comments. However, the actual teaching part—the use of Random Forest or SVM to train a computer to recognize toxic comments—doesn't seem to be in the code you shared. That part would involve taking the cleaned-up text and toxicity labels, teaching the computer what toxic comments look like, and then testing how well it learned. So, in short, the code you shared gets the data ready, but it doesn't actually train a computer to recognize toxic comments using Random Forest or SVM.

5. Result and Discussions

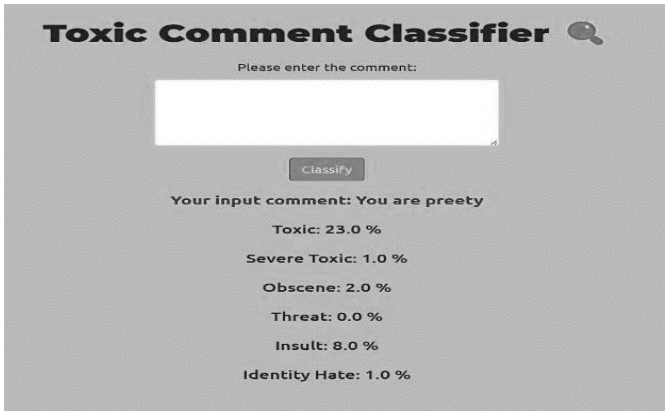

Figure 10 Non-toxic Comment

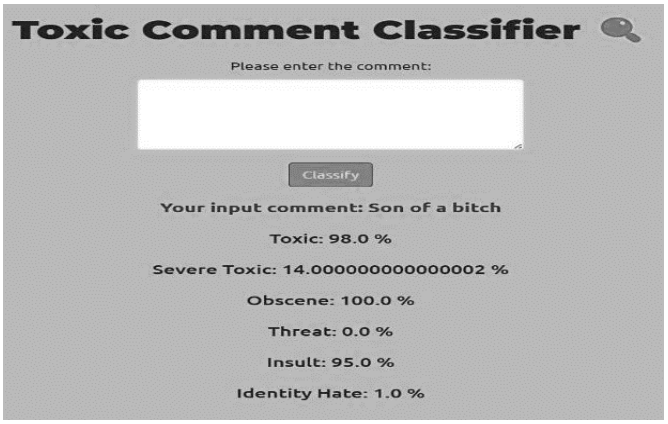

Figure 11 Toxic comment

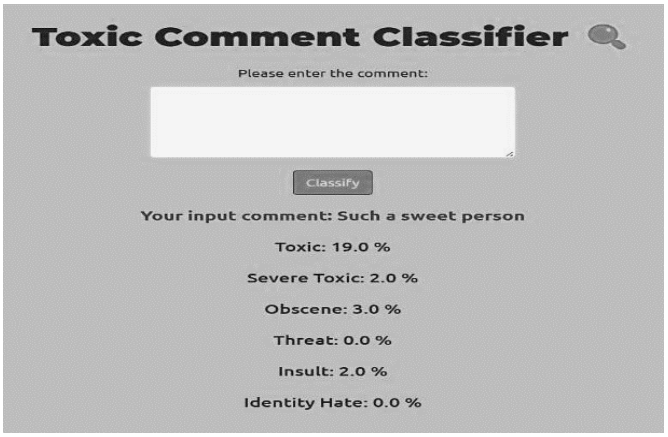

Figure 12 Non-toxic comment

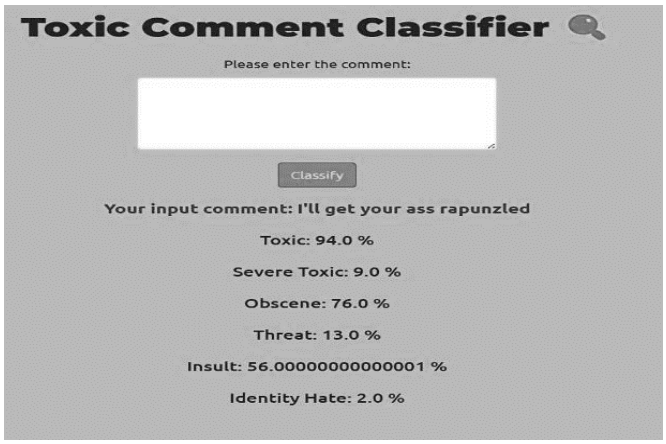

Figure 13 Toxic comment

Fig.10 depicts the non-toxic comment & Fig.11; Fig.12 & Fig.13 depicts toxic comment. In the present work deep learning classifier is employed to classify toxic comment & non-toxic comment sentences & words.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 02 Issue: 07 July 2024
Page No: 2065- 2073
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2024.0282

6. Performance Analysis

Table 1 Performance Parameter Details W.R.T Toxic & Non-Toxic Comments

SENTENCES	TOXICITY CATEGORY						NON-TOXIC	ACCURACY %
	Severe Toxic	Obscene	Threat	Insult	Identity Hate	Only Toxic		
1. You are pretty							90%	90%
2. I hate you	3%	3%	2%	17%	10%	97%		97%
3. You jerk	1%	77%	0%	93%	2%	99%		99%
4. Such a sweet person.							95%	95%
5. You're a doofus.	1%	11%	0%	37%	1%	63%		63%
6. Her positive aura fills the room with joy.							96%	96%
7. Son of a bitch	14%	100%	0%	95%	1%	98%		98%
8.I Love U.							97%	97%
9. I'll get your ass rapunzled.	7%	84%	2%	68%	2%	95%		95%
10.Your friendship means a lot to me.							98%	98%

Conclusion

This work not only advances the frontier of toxic comment detection but also lays the groundwork for future explorations in leveraging deep learning and NLP for enhanced content moderation. As the work move forward, the insights gained from this endeavor serve as a valuable contribution to the broader discourse on harnessing technology for fostering healthier online interactions. The present system attains 99% for toxic & non-toxic comment content. Table 1 shows Performance parameter details w.r.t Toxic & Non-Toxic comments

Acknowledgement

The authors truly appreciate college staff members for their invaluable help and guidance. Thank you for taking time out of your time during all stages of this work.

References

[1]. Risch, Julian & Krestel, Ralf. (2020). Toxic Comment Detection in Online Discussions. 10.1007/978-981-15-1216-2_4.

[2]. Rani, A., & Shyan, H. (2023, June). A Comparative Study and Analysis on Toxic Comment Classification. In 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS) (pp. 783-787). IEEE.

[3]. T. C. Nagavi and A. D. S., "Detection and Classification of Toxic Content for Social Media Platforms," 2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), Noida, India, 2021, pp. 368-373, doi: 10.1109/RDCAPE52977.2021.9633647.

[4]. N. Chetty, S. Alathur, D. Andrews and V. Kumar, "Computational Analysis of Online Hate Content using Cognitive -AI," 2021 6th International Conference on Computing, Communication and Security (ICCCS), Las Vegas, NV, USA, 2021, pp. 1-4, doi: 10.1109/ICCCS51487.2021.9776339.

[5]. A. Bodaghi, B. C. M. Fung and K. A. Schmitt, "Technological Solutions to Online Toxicity: Potential and Pitfalls," in IEEE Technology and Society Magazine, vol. 42, no. 4, pp. 57-65, Dec. 2023, doi: 10.1109/MTS.2023.3340235.

[6]. K. Lin, "Do Abstractions Have Politics? Toward a More Critical Algorithm Analysis," 2021 Conference on Research in Equitable and Sustained Participation in Engineering, Computing, and Technology (RESPECT),

Philadelphia, PA, USA, 2021, pp. 1-5, doi: 10.1109/RESPECT51740.2021.9620635.

[7]. Rahul, H. Kajla, J. Hooda and G. Saini, "Classification of Online Toxic Comments Using Machine Learning Algorithms," 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2020, pp. 1119-1123, doi: 10.1109/ICICCS48265.2020.9120939.

[8]. N. Haque, M. B. Alam, A. A. Towfiq and M. Hossain, "Bangla Toxic Comment Classification and Severity Measure Using Deep Learning," 2022 International Conference on Recent Progresses in Science, Engineering and Technology (ICRPSET), Rajshahi, Bangladesh, 2022, pp. 1-5, doi: 10.1109/ICRPSET57982.2022.10188551.

[9]. M. N. Fauzan, A. G. Putrada, N. Alamsyah and S. F. Pane, "PCA-AdaBoost Method for a Low Bias and Low Dimension Toxic Comment Classification.," 2022 International Conference on Advanced Creative Networks and Intelligent Systems (ICACNIS), Bandung, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICACNIS57039.2022.10055017.

[10]. R. Rivaldo, A. Amalia and D. Gunawan, "Multilabeling Indonesian Toxic Comments Classification Using The Bidirectional Encoder Representations of Transformers Model," 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), Medan, Indonesia, 2021, pp. 22-26, doi: 10.1109/DATABIA53375.2021.9650126.

[11]. N. K. Singh and S. Chand, "Machine Learning-based Multilabel Toxic Comment Classification," 2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), Greater Noida, India, 2022, pp. 435-439, doi: 10.1109/ICCCIS56430.2022.10037626.

[12]. M. Husnain, A. Khalid and N. Shafi, "A Novel Preprocessing Technique for Toxic Comment Classification," 2021 International Conference on Artificial Intelligence (ICAI), Islamabad, Pakistan, 2021, pp. 22-27, doi: 10.1109/ICAI52203.2021.9445252.

[13]. N. Boudjani, Y. Haralambous and I. Lyubareva, "Toxic Comment Classification For French Online Comments," 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Miami, FL, USA, 2020, pp. 1010-1014, doi: 10.1109/ICMLA51294.2020.00164.

[14]. M. Vichare, S. Thorat, C. S. Uberoi, S. Khedekar and S. Jaikar, "Toxic Comment Analysis for Online Learning," 2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Ernakulam, India, 2021, pp. 130-135, doi: 10.1109/ACCESS51619.2021.9563344.

[15]. S. Smetanin and M. Komarov, "Share of Toxic Comments among Different Topics: The Case of Russian Social Networks," 2021 IEEE 23rd Conference on Business Informatics (CBI), Bolzano, Italy, 2021, pp. 65-70, doi: 10.1109/CBI52690.2021.10056.

[16]. K. A. Kumar and B. Kanisha, "Analysis of Multiple Toxicities Using ML Algorithms to Detect Toxic Comments," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2022, pp. 1561-1566, doi: 10.1109/ICACITE53722.2022.9823822.

[17]. A. N. M. Jubaer, A. Sayem and M. A. Rahman, "Bangla Toxic Comment Classification (Machine Learning and Deep Learning Approach)," 2019 8th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 2019, pp. 62-66, doi: 10.1109/SMART46866.2019.9117286.

[18]. Shambharkar, P. G., Singh, H., Raghav, H. R., & Verma, H. (2023, May). Exploring the Efficacy of Deep Learning Models for Multiclass Toxic Comment Classification in Social Media Using Natural Language Processing. In 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-8).

IEEE.

[19]. Rani, A., & Shyan, H. (2023, June). A Comparative Study and Analysis on Toxic Comment Classification. In 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS) (pp. 783-787). IEEE.

[20]. T. A. Belal, G. M. Shahariar and M. H. Kabir, "Interpretable Multi Labeled Bengali Toxic Comments Classification using Deep Learning," 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE), Chittagong, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ECCE57851.2023.10101588.

[21]. M. Ibrahim, M. Torki and N. El-Makky, "Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 2018, pp. 875-878, doi: 10.1109/ICMLA.2018.00141.

[22]. A. A. Kumar, P. B. Pati, K. Deepa and S. T. Sangeetha, "Toxic Comment Classification Using S-BERT Vectorization and Random Forest Algorithm," 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/InC457730.2023.10263218.

[23]. J. Roy et al., "Investigating the Effect of Machine-Translation on Automated Classification of Toxic Comments," 2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS), Denver, CO, USA, 2022, pp. 764-769, doi: 10.1109/MASS56207.2022.00120.

[24]. K. Machova and T. Tomcik, "Comparison of Deep Learning and Ensemble Learning in Classification of Toxic Comments," 2023 21st International Conference on Emerging eLearning Technologies and Applications (ICETA), Stary Smokovec, Slovakia, 2023, pp. 353-358, doi: 10.1109/ICETA61311.2023.10343820.

[25]. S. Jain, G. Kaushik, P. Prabhu and A. Godbole, "Detox: NLP Based Classification And Euphemistic Text Substitution For Toxic Comments," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-5, doi: 10.1109/ICCCNT51525.2021.9579846.

[26]. H. Rahman Sifat, N. H. Nuri Sabab and T. Ahmed, "Evaluating the Effectiveness of Capsule Neural Network in Toxic Comment Classification Using Pre-Trained BERT Embeddings," TENCON 2023 - 2023 IEEE Region 10 Conference (TENCON), Chiang Mai, Thailand, 2023, pp. 42-46, doi: 10.1109/TENCON58879.2023.10322429.

[27]. V. Swetha, R. Anuhya, E. S. Sowmya and A. Geethanjali, "Building a Toxic Comments Classification Model," 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2021, pp. 1519-1523, doi: 10.1109/ICECA52323.2021.9675911.

[28]. A. S. Kapse, A. Dubey, H. Bisen, K. Kumar and M. Tamheed, "Multilingual Toxic Comment Classifier," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 1223-1228, doi: 10.1109/ICICCS56967.2023.10142540.

[29]. T. C. Nagavi and A. D. S., "Detection and Classification of Toxic Content for Social Media Platforms," 2021 4th International Conference on Recent Developments in Control, Automation & Power Engineering (RDCAPE), Noida, India, 2021, pp. 368-373, doi: 10.1109/RDCAPE52977.2021.9633647.

[30]. Shukla, A., & Arora, D. (2023, June). Deep Learning Model for Identification and Classification of Web Based Toxic Comments. In 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT) (pp. 274-279). IEEE.