WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

A Review on Automated Approaches to Detection of Social Media Toxicity

[1]Rajeshwari K, [2]Manish B S

[1]Assistant Professor, Information Science and Engineering, BMS College of Engineering, Bengaluru, India, rajeshwarik.ise@bmsce.ac.in
[2]Information Science and Engineering, BMS College of Engineering, Bengaluru, India, manishbs.is19@bmsce.ac.in

ABSTRACT—

The surge in cyberbullying and online abuse is a growing concern, particularly for the well-being of young individuals, contributing significantly to mental health issues like depression. The manual assessment of the overwhelming volume of daily data for toxic comments becomes an insurmountable task. A potential solution lies in automating the identification and censorship of such harmful comments by social media platforms, offering a promising avenue to address this escalating problem. However, the intricacies of detecting toxic comments add complexity, considering factors like context, perception, and vocabulary

The core objective of our project revolves around evaluating the toxicity embedded in individuals' social media profiles based on their activities. This encompasses an in-depth analysis of captions, bios, and other expressions employed on platforms such as Twitter. The comprehensive examination of an individual's entire profile involves a meticulous comparison with predefined data, culminating in the generation of a detailed report quantifying the level of toxicity within the social media presence. The derived report serves as a valuable resource for the cyber-security department, enabling them to ascertain whether an individual is exhibiting toxic or bullying behavior online. This approach fosters a more judicious management of social media content, fostering a safer and more positive online environment.

Keywords—Cyberbullying, Online abuse,Mental health, Toxic comments, Social media automation, Cyber-security

1. Introduction

The escalating use of online platforms has granted individuals unprecedented freedom to express their opinions and ideas. Popular social media sites like Twitter and Facebook, integral to daily life, especially among teenagers, serve as influential channels for free expression. However, the surge in social media usage comes with adverse consequences, particularly for adolescents who are frequently exposed to various behavioral and psychological threats. Cyberbullying, manifested through influential social attacks, emerges as a significant source of these threats. The anonymity afforded by social media platforms allows individuals to hide their identities, facilitating the misuse of technical features for unkind deeds. As a result, instances of cyberbullying have become more frequent over time, contributing to the broader context of Social Media Toxicity.

One of the most perilous activities on social networking platforms is the proliferation of offensive language containing abusive behavior aimed at causing harm to others. Cyberbullying encompasses various forms, including aggressive content, harassment, toxic comments, hate speech, sexism, and racism. The consequences of these hateful texts can be severe, leading to detrimental mental health effects such as anxiety, depression, self-harm, emotional perplexity, and even suicidal thoughts or attempts. Studies, like the one conducted by the Pew Research Center, reveal alarming statistics, indicating that over 60% of US citizens on social media platforms have experienced cyberbullying, with teenage girls enduring the worst forms. Consequently, the global prevalence of cyberbullying has reached epidemic proportions, firmly establishing it as a critical component of Social Media Toxicity.

In response to this epidemic, various global preventive and intervention approaches have been introduced to enhance the safety of internet users. Detecting offensive text, a pivotal aspect of combating cyberbullying, proves challenging due to the diverse language used by individuals. The task gains significance as people often employ sarcasm, intimidation, coarse language, and colloquialisms in a friendly manner, inadvertently causing harm. Researchers worldwide are actively engaged in identifying whether a post or a tweet is offensive, contributing to ongoing efforts to address the pervasive issue of Social Media Toxicity.

The figure [1] illustrates a poignant case of Social Media Toxicity, depicting an individual receiving corrupting remarks on their post. The United States has witnessed a surge in awareness regarding cyberbullying in the 2010s, driven in part by prominent cases that have captured public attention. Despite the enactment of laws specifically addressing cyberbullying in a few US states and various countries, the prevalence of cyberbullying cases continues to escalate. This persistent rise in cases raises questions about the efficacy of existing laws in mitigating the impact of Social Media Toxicity. Prior efforts

have primarily focused on the detection of cyberbullying after it has already transpired, indicating a reactive approach rather than proactive measures to combat the growing issue.

Fig-1 Depiction of Social Media Toxicity

Our review identified the following methodologies to detect Social Media Toxicity, combining techniques from Natural Language Processing (NLP) with other approaches. To assess and mitigate toxicity, the introduction of social media toxicity involves a detector generating reports on the toxicity level associated with a given account or user ID. In addition, prominent machine learning algorithms, including Multinomial Naive Bayes (MNB), Logistic Regression (LR), and Decision Tree Model (DT), contribute to the efforts to proactively address and reduce toxicity in online interactions.

II. Literature Survey

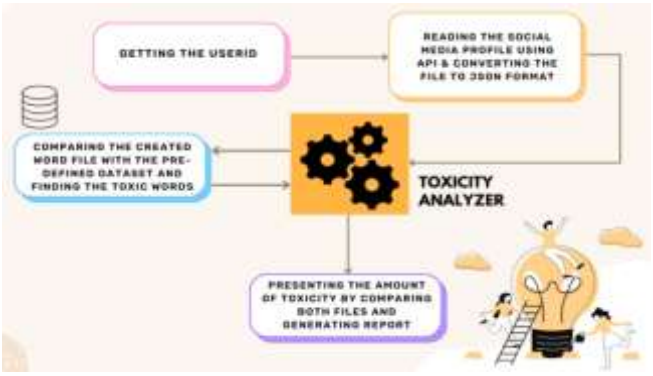

Fig 1. System Implementation in [1]

Narain C A, Tirugnanam, Shrii Sudhan, Mr. S. Rajesh Kumar, et al. [1] have proposed an approach leveraging Python as the platform, incorporating the snscrape library module, JSON library, and Flask library module for social media toxicity detection. Their project involves an HTML page designed to collect the user's Twitter account name and the total number of last tweets to retrieve posts and comments by the user. These details are then sent to the Python Flask login page function through an HTTP POST/GET request. Subsequently, the information is provided to the snscrape's TwitterUserScraper module, where tweet contents up to the specified number are fetched and appended to a Python array named 'tweets.' The collected tweets are saved in a newly created result.json file at the current code location. The data undergoes parsing through for loops, checking for swear words from the dataset.json. If a match is found, the matched word is added to the swear category strings, and counts are tallied accordingly. The count of swear words based on specified categories in the dataset file and categorized swear words are then sent to the success function of the Python module. This function incorporates an embedded HTML file, connecting the sentence with the embedded code and ultimately displaying the contents in a new HTML success page for the user.

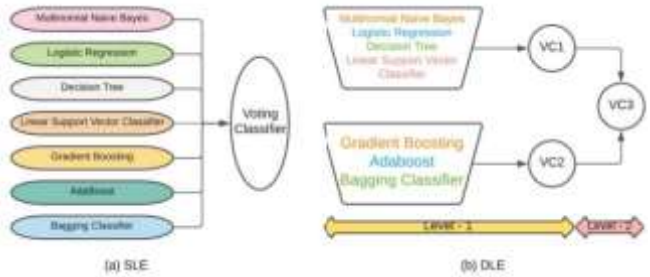

Fig 2. Proposed Models in [2]

Kazi Saeed Alam, Shovan Bhowmik, et al. [2] have undertaken a research endeavor focusing on the construction of two ensemble-based voting models for the identification of offensive or non-offensive texts. Their proposed model has exhibited superior performance compared to independently applied machine learning (ML) algorithms and ensemble techniques. Notably, they have achieved an impressive accuracy rate of 96% for the Twitter-extracted dataset. Looking ahead, the researchers express their intention to expand their dataset collection efforts by incorporating multiple diversified datasets, including private ones, to further evaluate and enhance their model's performance. Currently, their categorization efforts have been limited to two groups, prompting consideration for an extension to explore the model's efficacy in addressing multi-class classification challenges. The suggested models, as per the researchers, hold the potential for application in other text classification endeavors, contributing to more meaningful analyses in related fields.

Classifier	F1-score		Accuracy	
	Dat1	Dat0	Dat1	Dat0
Naives'Bayes [Wo-Tf-Idf LIWC]	0.86	0.81	0.89	0.82
Logistic regression [Wo Tf-Idf LIWC]	0.87	0.82	0.90	0.81
CNN [Embeddings]	0.89	0.83	0.94	0.86
LSTM [Embeddings]	0.88	0.82	0.94	0.85

Table 1. Results in [3]

Cheniki Abderrouaf, Mourad Oussalah, et al. [3] address hate-speech online identification in their paper. Employing original feature engineering and a transfer learning scheme utilizing a negated dataset for balanced training, they test their methodology on the Wikipedia comment Corpus, covering hate speech categories such as Toxicity, Personal Attack, and Aggression. The study contrasts a convolutional neural network (CNN) with long short-term memory (LSTM) architectures using FastText word embeddings against baseline algorithms like Logistic Regression and Naive Bayes classifiers. Feature sets, including Word-level Tf-Idf, Character-level Tf-Idf, and LIWC, are compared, highlighting the feasibility of the developed transfer learning scheme to outperform standard random sampling approaches. The study emphasizes the superiority of the constructed CNN and LSTM-based classifiers in the overall classification of hate speech categories, paving the way for future hybridization schemes incorporating more sophisticated sentence constructs in testing datasets and feature selection processes.

Kanwal Yousaf Tabassam, Nawaz, et al. [4] present a novel deep learning-based framework for detecting and classifying inappropriate child video content. Utilizing transfer learning with EfficientNet-B7 architecture, the model extracts video features processed through a BiLSTM network, enabling effective video representations and multiclass video classification. The evaluation, conducted on a manually annotated cartoon video dataset of 111,156 clips from YouTube, reveals the proposed EfficientNet-BiLSTM framework (with hidden units = 128) outperforming other models, achieving an impressive accuracy of 95.66%. Comparative analysis with state-of-the-art models underscores the BiLSTM-based framework's superiority, showcasing the highest recall score of 92.22%.

Hong Fan, Wu Du, Abdelghani Dahou, Ahmed A. Ewees, Dalia Yousri, Mohamed Abd Elaziz, Ammar H. Elsheikh, Laith Abualigah, Mohammed A. A. Al-qaness, et al. [5] investigate toxicity detection in social media using deep learning techniques. Employing Bidirectional Encoder Representations from Transformers (BERT), the study classifies toxic comments from user-generated data, focusing on tweets. The BERT-base pre-trained model is fine-tuned on a labeled toxic comment dataset from Kaggle public datasets and tested on real-world data, specifically two different tweet datasets collected in distinct periods related to the UK Brexit case study. Evaluation results underscore BERT's ability to classify and predict toxic comments with a high accuracy rate. Comparative analysis with Multilingual BERT, RoBERTa, and DistilBERT reveals that the BERT-base model outperforms all counterparts, securing the best results.

Classifier	2-Gram	3-Gram	4-Gram	Average
SVM	89.42%	89.5%	**90%**	89.6%
Neural Network	**93%**	92.5%	91.7%	92.4%

Table 2. Precision Scores of the models

John Hani, Mohamed Nashaat, Mostafa Ahmed, Ammar Mohammed, et al.[6] propose a machine learning approach for cyberbullying detection. Using SVM and Neural Network classifiers with TFIDF and sentiment analysis, the Neural Network achieves 92.8% accuracy with 3-grams, outperforming SVM (90.3% accuracy with 4-grams). Comparison with related work shows the Neural Network's superior accuracy and f-score (91.9% vs. 89.8%). The study aims to improve cyberbullying detection for safer social media use but notes limitations related to the training data size. Larger cyberbullying datasets and the adoption of deep learning techniques are suggested for enhanced performance.

Sinchana C., Sinchana K., Pradyumna C. S., Deepika S., et al.[7], conducted a study reviewing and applying various machine-learning techniques, including SVM, Neural Network (NN), C4.5 decision tree algorithm, instance-based learning, and J48 algorithm, to efficiently detect cyberbullying.

Amit Sheth, Valerie L. Shalin, Ugur Kursuncu, et al. [8], in their paper, explore the influences on toxic exchange detection beyond conventional content analysis. Their goal is to establish a framework that incorporates multiple dimensions of toxicity, leveraging explicit knowledge in a statistical learning algorithm to address ambiguity. The study emphasizes the significance of multi-level analysis—content, individual, and community—for toxicity detection, underlining the numerous features essential for determining toxicity. The infusion of knowledge representation in a learning algorithm emerges as a solution for toxicity detection and related challenges.

Model	Accuracy	Precision	Recall	Macro-F1
Majority Class	58.89	19.63	33.33	24.71
LR-BoW	84.15	83.45	81.11	82.19
BERT-cahya	86.53	85.90	**84.22**	84.95
BERT-indolem	83.95	81.87	83.12	82.47
BERT-indobenchmark	**86.99**	**87.69**	83.88	**85.59**

Table 3. F1-scores of the Bert model in [6]

Muhammad Amien Ibrahim, Noviyanti Tri Maretta Sagala, Samsul Arifin, Rinda Nariswari, Nerru Pranuta Murnaka, Puguh Wahyu Prasetyo, et al. [9], focus on classifying hate speech, abusive messages, and normal tweets on Indonesian Twitter. Utilizing a dataset from a previous study with 5860 tweets categorized as hate speech, abusive, or normal, the research develops a benchmark model, logistic regression model, and transformer models of BERT for multiclass classification. Overall, all machine learning models achieve high accuracy, with BERT models slightly outperforming logistic regression. The best-performing model, BERT-indo benchmark, achieves a Macro-F1-score of 85.59, demonstrating its ability to distinguish between hate speech and abusive language on Indonesian Twitter.

Yau-Shian Wang, Yingshan Chang, et al. [10], focus on automatic toxicity detection for online moderation, emphasizing its significance for positive societal impact in NLP research. They incorporate the SOCIAL BIAS FRAME (Sap et al., 2020) into prompt design for detecting implicit biases. The approaches are evaluated on three hate speech/toxicity detection datasets, showcasing the advantages of generative classification over discriminative classification in both quantitative and qualitative results. The study concludes with an analysis of prompt wording's interaction with model behavior, followed by discussions on the ethical implications of self-diagnosis.

Paula Reyero Lobo, Enrico Daga, Harith Alani, et al. [11], although limited to toxic speech in a single platform, language, and time, identifies drawbacks in widely used annotation protocols for toxicity detection and related phenomena (e.g., hate speech, abusive and offensive language) (Fortuna, Soler-Company, and Wanner 2021). The study suggests promising directions, such as using ontologies to validate texts verified by human annotators or assist them in the annotation process, for building more consistent and reliable benchmarks. Additionally, leveraging knowledge of specific demographic groups can enrich and preprocess datasets where information remains unidentified, mitigating risks associated with using "one-size-fits-all" models.

Sanoussi, Mahamat Saleh, Xiaohua Chen, et al. [12], propose an automated mechanism for detecting hate, offensive, and insult content on the Facebook platform. Combining Fasttext features with an SVM classifier yields the best results, achieving high f1-scores for Insult and Hate class labels at 95.4% and 93.9%, respectively. The study marks a crucial step in hate speech detection, with future plans including deep annotation for various types of offensive text in mixed languages and the exploration of robust features through pre-trained models and deep learning approaches.

Wijesiriwardene, Thilini et al. [13], introduced the ALONE dataset for adolescents, categorizing offensive language into appearance, intellectual, political, race, religion, and sexual preference types. The dataset, with unique features and interaction-based design, provides ground truth for understanding online toxic behavior and training machine learning models. It aids in identifying toxicity patterns, allowing researchers to develop guidelines for various toxic behaviors. The dataset is available upon request for research purposes, subject to a non-public dissemination agreement.

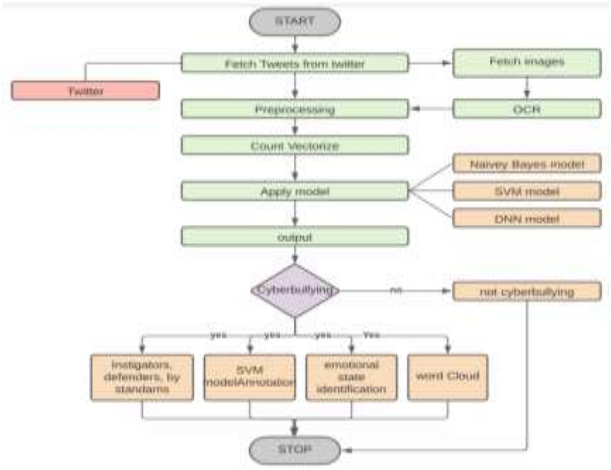

Fig 3. Flow Diagram in [14]

Nirmal, Niraj et al. [14] aimed to automatically detect cyberbullying-related posts on social media, addressing the challenges of manual monitoring in the era of information overload. The project introduces a system for prompt detection of cyberbullying signals, covering various types and involving posts from bullies, victims, and bystanders. The study employs Naive Bayes, SVM, and DNN algorithms for effective detection.

Leite, Joao A. et al. [15] introduced ToLD-Br, a dataset for classifying toxic comments on Brazilian Portuguese Twitter. Their analysis emphasized the dataset's necessity for studying automatic classification and underscored the challenges, including significant class imbalance in multi-label classification. The study showcased the ongoing relevance of monolingual approaches over multilingual experiments, emphasizing the need for large-scale datasets in building reliable models.

III. Conclusion

The pervasive issue of cyberbullying and offensive content on social media is significantly impacting individuals, especially teenagers, and has severe consequences, including suicidal thoughts among the victims. This highlights the pressing concern of social media toxicity, emphasizing the need for effective measures to address and mitigate its harmful effects

The research delves into the exploration and assessment of diverse machine-learning techniques for detecting cyberbullying, showcasing their effectiveness in various datasets. Beyond traditional machine learning, the study also explores the utilization of Neural Networks and web scraping tools to enhance early detection capabilities in addressing the broader spectrum of social media toxicity.

The research findings pave the way for future advancements in combating social media toxicity. Potential applications include the development of more sophisticated machine-learning models and Neural Networks for enhanced cyberbullying detection. Additionally, the integration of innovative web scraping tools could further refine the accuracy and efficiency of toxicity detection in evolving online environments. The insights garnered from this research contribute to building robust frameworks, fostering safer online spaces, and potentially informing the creation of preventive measures and interventions against cyberbullying and social media toxicity.

References

[1] Narain C A[1] , Tirugnanam[1] ,Shrii Sudhan[1] , Mr. S. Rajesh Kumar[2]. Social Media Toxicity Analyser

[2] Alam, Kazi & Bhowmik, Shovan & Prosun, Priyo. (2021). Cyberbullying Detection: An Ensemble Based Machine Learning Approach. 710-715. 10.1109/ICICV50876.2021.9388499.

[3] Cheniki, Abderraouf & Oussalah, Mourad. (2019). On Online Hate Speech Detection. Effects of Negated Data Construction. 5595-5602. 10.1109/BigData47090.2019.9006336.

[4] Yousaf, Kanwal & Nawaz, Tabassam. (2022). A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3147519.

[5] Yousaf, Kanwal & Nawaz, Tabassam. (2022). A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos. IEEE Access. 10. 1-1. 10.1109/ACCESS.2022.3147519.

[6] John Hani , Mohamed Nashaat , Mostafa Ahmed , Zeyad Emad , Eslam Ame(2019). Social Media Cyberbullying Detection using Machine Learning

[7] C. Sinchana, et al.; International Journal of Advance Research, Ideas and Innovations in Technology(2020). Detection of Cyberbullying using Machine Learning

[8] Amit Sheth1 , Valerie L. Shalin2,1 , Ugur Kursuncu1,3 (2021).Defining and Detecting Toxicity on Social Media: Context and Knowledge are Key

[9] Ibrahim, Muhammad Amien & Sagala, Noviyanti & Arifin, Samsul & Nariswari, Rinda & Murnaka, Nerru & Prasetyo, Puguh. (2022). Separating Hate Speech from Abusive Language on Indonesian Twitter. 187-191. 10.1109/ICoDSA55874.2022.9862850.

[10] Wang, Yau-Shian & Chang, Yingshan. (2022). Toxicity Detection with Generative Prompt-based Inference. 10.48550/arXiv.2205.12390

[11] Reyero Lobo, Paula & Daga, Enrico & Alani, Harith. (2022). Supporting Online Toxicity Detection with Knowledge Graphs. Proceedings of the International AAAI Conference on Web and Social Media. 16. 1414-1418. 10.1609/icwsm.v16i1.19398.

[12] . Sanoussi, Mahamat Saleh & Xiaohua, Chen & Agordzo, George & Guindo, Mahamed & Omari, Abdullah & Mahamat, Boukhari. (2022). Detection of Hate Speech Texts Using Machine Learning Algorithm. 0266-0273. 10.1109/CCWC54503.2022.9720792.

[13] .Wijesiriwardene, Thilini & Inan, Hale & Kursuncu, Ugur & Gaur, Manas & Shalin, Valerie & Thirunarayan, Krishnaprasad & Sheth, Amit & Arpinar, Ismailcem. (2020). ALONE: A Dataset for Toxic Behavior among Adolescents on Twitter.

[14] INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET) E-ISSN: 2395-0056 VOLUME: 07 ISSUE: 12 | DEC 2020. Automated Detection of Cyberbullying Using Machine Learning

[15] Toxic Language Detection in Social Media for Brazilian Portuguese: New Dataset and Multilingual Analysis(2020)