Sustainable Positive Communication in Cyberspace by Detecting Toxic Text

Pranchis Narzaree[1, *]
Research Scholar, Department of Computer Science and Technology, Bodoland University

Prof. (Dr.) Manoj Kr. Deka[2]
Department of Information Technology, School of Technology, Assam Skill University

ABSTRACT

With the emergence of social media technologies, it has been a great opportunity for the people to virtually come closer and communicate with each other irrespective of the location far distance apart. But, with the myriad growth of internet users, it is sometime disgraceful of using toxic or bad words in the messages while conveying dislike or like for something or someone, which is not done in any good society. This paper aims to study various toxic text studied in different papers, contrastive machine learning and deep neural network models, methods used in detection and classification of toxic text. As this is a review paper, Descriptive, survey method were used in the study where literature survey and observation techniques were tools utilized. The analysis was done using descriptive analysis theory based on the findings in the research paper where identifying, analyzing and summarizing were the contexts in the form of tables and figures. Results shows toxic text used in the papers are found to be toxic, insult, obscene, offensive, racist, sexist, hateful, attack, threat, abusive, non-abusive, entity directed criticism, hate speech, cyber-bullying where toxic, insult, threat, abusive were in common. The adaptive machine learning models used are Naïve Bayes, Random Forest, and Support Vector Machine and deep learning models used are CNN, Bi-LSTM, BERT and its variants MBERT, XLM-R. The studied ensemble techniques are utilized to predict and acquire a robust model that outperforms other baseline models. In the future, further review to be made using low research languages.

Keywords: Cyberspace, messages, toxic text, positive communication, social media

Title:-“Sustainable Positive Communication in Cyberspace by Detecting Toxic Text”

1) Introduction

Cyberspace offers a great deal of space to socially come closer and have conversation across the world round the clock. Communication and sharing has became an easy access for common people. It forms a great medium of electronic communication, even surpassing the traditional methods of communication. Cyberspace refers to the digital sphere that has become the consequence of development and use of computer internet. It encompasses the virtual world of digital data, communication, and interactions that occur through interconnected computers, smartphones, and other digital devices. It serves as a platform for a vast array of activities, including social media phenomenon, education, information transaction, entertainment, and various forms of digital services and dealings.

As cyberspace has become more pervasive, it has also introduced new challenges and opportunities, such as cyber security threats, data privacy concerns, and the potential for both positive and negative societal impacts.

1.1 Internet Users in India

As per 'Digital 2024' report [1], India was having 751.5 million internet users at the beginning of 2024 and became the location for 462.0 million social media users. It stands at 52.4 percent in spreading or penetration of internet while social media users were 32.2 percent of the total population. Cellular mobile connections were found to be 1.12 billion which positions at 78.0 percent of the total population in India.

1.2. About Toxic Text

Toxic texts in messaging is one such threat through internet social media (like Facebook, tweeter etc.) is found to be unacceptable activity that refer to online content that is harmful, offensive, or abusive in nature. This can include harassment, hate speech, bullying and content that promotes misinformation or conspiracies. Toxic texts can have negative psychological impacts on readers, especially assailable individuals like children and adolescents. Nevertheless, any text that is likely to cause harm or offense to its readers could be considered as a toxic text.

Therefore, an effort has been made to study toxic text messages that may be used in social media contents or internet sites with perceptive of internet users or various social networking sites. Addressing and eliminating toxicity is now-a-days a challenge for online platforms, content moderation, and digital accomplishment. As with the growth of social media, instant messaging, and video conferencing, it is now become easier than ever to stay connected with friends and family, even if they live far distance apart.

1.3. Aims and Objectives

The study aims to identify various toxic text so that it can be eradicated using various techniques and have good and positive communication environment in social networking sites and internet media as a whole. The research study has the following objectives.

1. To study the different types of toxic text present
2. To explore assorted methods adopted in detecting toxic text
3. To highlight the results conferred in classifying toxicity

1.4 Related Work

As the focus of the study is on positive communication in Cyberspace as well as identification of toxic text, Huda, Miftachul (2023) indicated improvement of virtual interaction and information management for safe cyberspace communication[2] with organizational sustainability. Hence with the strategic support, digital social connections in the cyberspace can be utilized. Stanley D. Brunn (2014) focused on cyberspace boundaries which exist between and within fields and disciplines studying sustainability[3]. Marshan et. al., (2023) made a study on to fight against harassment in online platforms by detecting the severity of abusive comments[4]. They made a study on various machine learning models such as Naïve Bayes, Random Forest, Support Vector Machine, and deep earning models as Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM) for detecting and classifying toxic text where they found Random Forest with bi-grams could perform better with accuracy of (0.94), a precision of (0.91), a recall of (0.94). Towards such context Kiritchenko et. al. (2024), made a study on abusive language with the aspect of ethical and human rights in online messaging[5]. They highlighted the need for examining the broad social impacts such as task formulation and dataset design to model training and evaluation and to deployment. Desai et. al. (2021) studied about cyber bullying on social media using machine mearning. Such type of bullying can be thought of threatening, calumny, and chastising the individual[6]. It has also led to suicide attempts. Fan et. al. (2021) [7], highlighted the classification of toxicity using deep learning approaches. They used BERT models with two dataset where their proposed model performs well. However, Sheth et. al. (2022) focused on defining and detecting toxicity in social media. They found such toxic words are related to online hate speech, internet trolling, and sometimes outrage culture[8]. Even in the study of Teng and Varathan (2023), they proposed two approaches for cyberbullying[9] detection such as Conventional Machine Learning and Transfer Learning. Shrestha et. al. (2023), has studied on detection of toxic language and threats in Swedish which he expressed as harmful communication[10]. They used BERT model for detection of abusive words in communication.

Song et. al., 2021 made a study on multilingual toxic text detection in imbalance sample distribution[11]. Mohammed and Rania[12] reviews on many deep learning methods to detect toxic text that set for a challange. Wand et. al., 2021, after survey of toxic comments, they classified toxic words which may results in insults, bulgar words, and threats [13]as benial words while classifying. While in [14] Nitya et. al., (2024) have recently developed an AI methods for detection of social media comments using CNN, Naive Bayes model, as well as LSTM. It was a aim to build models of higher accuracy than the previous result by others. Madhyastha et. al., (2023) made a study on universe of discourse for toxicity in online conversations [15]. Smith et. al., (2021), made a study on consequences of using social media in adolescents. [16] try to explore the kinship between social media use and loneliness and belonging among adolescents and young adults. Ognibene et. al. (2023) proposed a hypothesis on accommodative Social Media Virtual Companion to interact in social media environments in order to achieve desirable status[17]. Kindle et. al., (2024), presents a original method on network toxicity analysis [18] for the inductive analysis of the dynamics of discursive toxicity within social media. Matamoros-Fernández and Farka (2021), made a collective distinction with discussion on Racism, Hate Speech, and Social Media. As far as the communication positivity is concerned in workplace, Bhat et. al. (2021), studied in three stages such as taxonomy of toxic language, creating dataset based on taxonomy of toxic language in workplace and use of offensive language and hate-speech datasets not effective in identifying toxicity in workplace communication. With respect to the voluntary of researchers [19] found deficient of geographical and platform diversity as tending conflict increased race orientation to unpack bias on social media. [20] also highlighted the impact of toxic communications at workplace in gratification of job. In the content of security in cyberspace Khraisat et. al., (2019), were concerned about the rise of malicious software that is causing a threat in cyberspace thereby its

now a challenge to critically design intrusion detection systems (IDS)[21]. Ptaszynski et. al., (2016) studied on cyberbullying and formulated a refreshing technique for automatic detection of cyberbullying entries on the cybermedia. They tried on seed words where three categories of semantic orientation score and then increased relevance of categories were calculated. They [22] found the proposed model outperformed baseline framework in both training stage and real world constraints. Lashkarashvili and Magda (2022), made a study on identifying toxicity in Georgian discussions. The dataset was prepared from the online platform "Tbilisi forum" that contains 10,000 comments which were labeled as toxic and non-toxic. They[23] used models using deep learning architectures such as NCP, biRNN, CNN, biGRU-CNN, biLSTM, biGRU, transformer, and a baseline NB-SVM and found satisfactory results with NCP and best results performance with CNN achieving 0.888 ACC and 0.942 AUC. Though many methods were applied to detect toxicity in discussions, messaging through text, Kumar et. al. (2020), made classifications of toxicity uisng CNN and Gru network. They detect toxicity like obscenity, threats, insults and identity-based hatred. They used LSTM type of RNN model, Long Short-Term Memory (LSTM) for detection as well as grouping of toxic text. Their task for text classification involved word representations and study the performance on text mining methodologies. Their propitious results motivated them for further development of CNN based methodologies for text mining and classifications by utilizing adaptive learning and comparisons with n-gram based techniques[24].

2) Materials and Methods

Collection of materials have been done from 22 reputed journals that include ACL (Association of Computational Linguistics), MDPI (Multi disciplinary Digital Publishing Institute), Cambridge University Press, Elsevier-ScienceDirect, Springer Open, EDP Sciences, IEEE Access, ACM(Association for Computing Machinery), SAGE, Taylor and Francis. Further, accumulation of concepts from journals were being considered from 2019 till 2024(August). Descriptive study Method was used in the study where literature survey and observation techniques were the tools used. The analysis was done using descriptive analysis theory where identifying, analyzing and summarizing were the contexts.

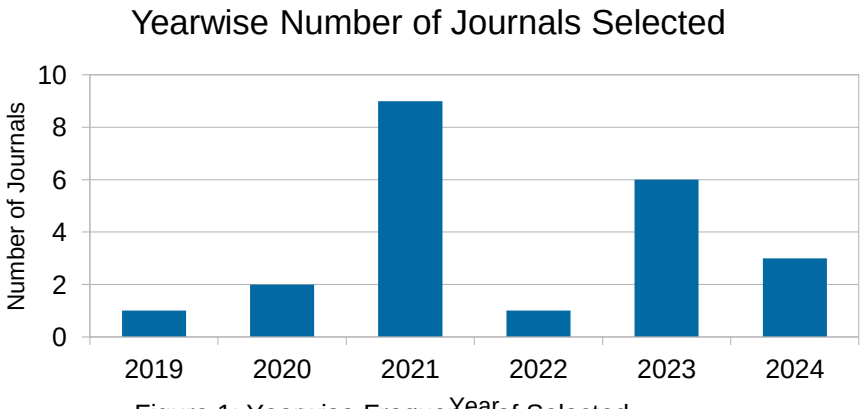

Figure 1: Yearwise Frequency of Selected Journals

3) Results and Discussion

With the identification of objectives, preprocessing and studying the materials accumulated, it was found to be promising and could provide an insight of the data.

3.1 Types of toxic messages

The following Table-1 illustrates the various types of toxic text present in the messages used in cyber media especially social networking sites.

Table-1: Different Types of Toxic Text

Author	Types of toxic text addressed	Dataset Source	Year
[4]	toxic, severe toxic, obscene, threat, insult, identity Hate	Kaggle's Toxic Comment Classification Challenge	2023
[5]	obscene, offensive, racist, insult, toxic, sexist, hateful, attack, threat, abusive, non-abusive, entity directed criticism, hate speech, cyber-bullying	Lexicons and annotated corpora are critical resources	2021
[6]	Bullying over social media, threatening, calumny, and chastising individual	tweeter	2021
[7]	hate speech, internet trolling, and sometimes outrage culture.	Twitter posts, Kaggle	2021
[8]	threats, obscenity, insults, and identity-based hate	facebook, tweeter	2021
[9]	cyberbully	Empath’s lexicon	2023
[10]	toxic, threats, Not toxic, Not Threats	Swedish social media and are annotated by at least three annotators. Hammer et al., Vidgen et al.	2023
[11]	non-toxic vs. Toxic.	Jigsaw Multilingual Toxic Comment dataset	2021
[13]	Toxic, Severe Toxic, Obscene, Threat, Insult, Identity Hate	Google Jigsaw	2021
[14]	cyberbullying,	Instagram-collected dataset	2024
[20]	Toxic, Impolite, Gossip, Offensive, Non-toxic	Microaggression dataset	2021
[22]	cyberbullying	Social media comments	2016
[23]	cyber-bullying, verbal harassment, or humiliation.	Tbilisi forum, an online platform for public discussions, Georgian dataset both for toxic comment classification and sentiment analysis.	2022
[24]	Toxic, Severe toxic, Obscene, Threatening, Insult, Identity hate	tweets	2020

After thorough observing the Table-1, it was found to be interesting in identifying various types of toxic text. The study made by [4][5][8][10][13][24] reflects having common properties in studying toxicity like toxic, insult and threat. [5] unwrap critical massive resources of lexicons and annotated corpora for acquisition of racist perceptive, criticism, hate speech, cyber-bullying also. The proposed BERT model of [6] achieved 91.90% accuracy using Twitter dataset which was a very good result as compared to other traditional machine learning models when used related datasets to study bullying over social media, calumny, chastising. Now-a-days internet trolling[7] is also being seen in common in facebook or tweeter. It has both positive and

negative effect. Paper shows it has used datasets like Twitter and Kaggle. Most of the study made, have taken materials from tweeter as well as created their own dataset from the existing corpus such as [4][11][13] used Google Jigsaw. [5][6][9][1][4][22][23] studied cyber-bullying in common. It was also seen in [10] and [23] that researchers are also using dataset other than English language such as Swedish, Georgian dataset etc. As social media platforms provides online discussions, exchange of thoughts and ideas, a critical examining of the contents becomes necessary. It also facilitates varied ways for users to explicitly upload their expressions and willingness to involve themselves in that environment. The textual matter nevertheless, pertain high level of toxicity like using indecent words to insult, demotivate, harass which may cause potential risk to the users.

3.2 Heterogeneous methods and models

With the advancement in technologies, different methods and models were also built for the research study or experiments. Following Table-2 shows the various methods or techniques used in solving research problems for the benefit of the society and research community as a whole.

Table-2: Methods & Models Adopted in Detecting Toxic Text

Paper	Purpose of Study	Methods/Models
[4]	To detect and classify toxic comments and to investigate the effect of text pre-processing on the performance of the machine and deep learning models. Comparing Machine Learning and Deep Learning Techniques for Text Analytics: Detecting the Severity of Hate Comments Online	Naïve Bayes, Random Forest, and Support Vector Machine, with deep learning models such as Convolutional Neural Network (CNN) and Bi-directional Long Short-Term Memory (Bi-LSTM)
[5]	To examine the broad social impacts of this technology, and to bring ethical and human rights considerations to every stage of the application life-cycle, from task formulation and dataset design, to model training and evaluation, to application deployment.	Survey
[6]	To detect cyberbullying and implement features with the help of a bidirectional deep learning model called BERT.	BERT model, SVM and Naive Bayes
[7]	To classify Social Media Toxicity using Using Deep Learning	BERT
[8]	To provide a framework that identifies and utilizes the multiple dimensions of toxicity and incorporates explicit knowledge in a statistical learning algorithm to resolve ambiguity.	Psychological survey
[10]	To Examining the intersection between toxic language and threats in Swedish language	BERT
[11]	To Study Multilingual Toxic Text Detection Approaches under Imbalanced Sample Distribution	MBERT, XLM-R

[12]	To provide comprehensive reviews of the various strategies for ensemble learning, especially in the case of deep learning.	Survey
[13]	To build a toxicity detector using machine learning methods including CNN, Naive Bayes model, as well as LSTM.	CNN, Naive Bayes model, as well as LSTM.
[14]	To study on cyberbullying detection system in the Moroccan dialect on an Instagram-collected dataset.	LSTM(GloVe)
[20]	To study and detect toxicity in Workplace Communications	Bert+ MLP
[22]	To study on sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization	Survey
[23]	To detect toxicity in online Georgian discussions	NCP, biRNN, CNN, biGRU-CNN, biLSTM, biGRU, transformer, and a baseline NB-SVM
[24]	It aims to classify toxic message from tweets collected hate speech on the subjects of religion and refugees	Convolution And Gru, Naive Bayes- multinomial Bernoulli event model with the n-gram bag-of- Words

Many researchers have used classical machine learning methods and models for the research study such as [4][6] where they have investigated the effects of preprocessing on the machine learning models and compared with the performance of deep learning models. To the highest degree [4][6][7][10][11][13][14][20][23][24] have used deep learning methods to get most accuracy out of their models. [6][7][10][11][20] used BERT model in common and other used survey methods to study the effect of different models in toxic text in research.

3.3 Performance of models

From the Table-3, it is very apparent that the performance of the models designed for the research are found to be greater than 90%. [4][6][14][23] used various learning models and compared with baseline models and represented their efficiency. Like as in [4] RF with bi-grams able to perform well with an accuracy of (0.94) whereas [6] achieved 91.90% accuracy and LSTM was capable to detect toxic text and classify 91.24% of its features.

Table-3: Summary of the Papers

Year	Findings	Paper
2023	After studying the various models and compared its performances demonstrated that the Random Forest with bi-grams achieved the best overall performance with an accuracy of (0.94), a precision of (0.91), a recall of (0.94), and an F1 score of (0.92). The study also takes into account to detect severity of abusive language in online platforms by building an efficient model that, contributes essential entailment in theory as well as live activity.	[4]

2021	It is learnt that the paper rigorously gave an effort to study and collect various sub-fields of abusive language detection and examine the fields through the views of ethics and human rights.	[5]
2021	It was considered after study of various features, with one such feature, BERT model achieved 91.90% accuracy when trained over dual cycles ofcourse, that was sentimental feature. Therefore, the proposed model outperformed the traditional machine learning Models which was a little further achievement.	[6]
2021	The outcome showed that the proposed model can efficiently classify and analyze toxic tweets.	[7]
2021	The author identified multiple influences on the exchange of toxic contents and thereby detecting offensive text beyond conventional content analysis.	[8]
2023	The effort highlights the assorted difficulties with harmful language and the paper focused on the need of using contrasting methods to detect such different form of harmful textual content.	[10]
2021	The experimental results demonstrates the introduction of fusion method based on different loss functions. It is found that it could effectively solve the problem of imbalance in exactitude and recall due to imbalanced samples. The paper suggested to use their approach to build an automated content moderation system and moreover, it can be native to any globalized social media platform such as Twitter, Facebook, Instagram, Snapchat, and Discord.	[11]
2021	The author found ensemble learning as advantageous that can is combine different individual models to improve performance prediction and obtain a stronger model that outperforms them.	[12]
2021	LSTM with GloVe embedding layer achieved the best accuracy and Kaggle score, and GloVe embedding layers have an overall better performance.	[13]
2024	The experimental results gave accuracies of around 77% to 91% from both the machine learning and deep learning algorithms. The LSTM model gave the best performance by 91.24% where outcome was far more better than other machine learning models.	[14]
2021	The author studies the significance of toxicity and presented a novel dataset that is annotated for toxic online content. The paper focused on understanding annotator perceptions of toxicity as well as confounding factors much as demographic factors along three dimensions such as gender, age and political orientation.	[15]
2021	The paper symbolized the various forms of online communication and related social outcomes such as belonging and loneliness. In this paper, consideration of individual, societal, and cultural factors that may explain preferable descriptors are sought.	[16]

2023	The paper proposes a theoretical framework based on an adaptive "Social Media Virtual Companion". It was for educating and supporting an entire community, teenage students, to interact in social media environments. The author studied such theory in order to achieve desirable conditions, defined in terms of a community specific and participatory designed measure of Collective Well-Being (CWB).	[17]
2020	The author provided a review and literary criticism on racism, hate speech, and social media contents. It focused on methodological, theoretical, and ethical challenges of the scholarly research field. The paper also presented their involvement in future Research.	[19]
2021	The paper presented low performance result on their dataset the rare of its kind using Microaggression dataset that they mention the only resource applicable to this domain. The authors also presented a taxonomy and annotating road map to study toxic language in workplace emails. They used ToxiScope and further demonstrated the necessity of new dataset to detect workplace toxicity since the models trained on existing excessively toxic datasets did not detect delicate toxic text.	[20]
2019	This survey paper presents a taxonomy of contemporary Intrusion Detection Systems of notable recent works. It provided an overview of the datasets commonly used for evaluation purposes. It also presents evasion techniques used by attackers to avoid detection as cyber criminals have shown their capability to obscure their identities, hide their communication, distance their identities from illegal profits, and use infrastructure that is resistant to compromise.	[21]
2022	The author presented a novel approach in toxic comment classification through a brain-inspired Neural Circuit Policy (NCP) model. They compared many baseline models, including NCP, showed satisfactory results. CNN performed well with 0.888 ACC and 0.942 AUC.	[23]
2023	The outcome provided by proposed Concurrent GRU model can outperform other models with well established methodologies. The paper provided enough evidence that models was appropriate for toxic comment classification.	[24]

[12] studied ensemble techniques to predict and acquire a robust model that outperforms other models. [16] symbolizes the assorted forms of online communication that become easy to connect and message across other parts of the world. The authors in [20] also presented a taxonomy and annotating programme to study toxic language in emails of workplace. They have also used Microaggression dataset for their study. [23] made a study on brain-inspired Neural Circuit Policy (NCP) model and inferred satisfactory accuracy but could not outperform baseline models.

4. Conclusion

As merging towards zenith of technology, Cyber media and others have noticeable impact on our life and how we connect with others. If positive side is considered, technology has made it easier to stay connected with family and friends far distance away. It has also provided us with news and events updates all across the globe at our finger tips. People can

share their knowledge and experiences with others. But, if other side of the aspect is considered, cyber-media can also have a negative impact on the relationships. It could create difficulties in building meaningful relationships as that may lead to feelings of isolation and loneliness, as well as a lack of face-to-face interaction.
Different researchers have profound feelings of having a congenial relationship in communication in social media or other internet sites. As with the myriad rise of messaging in social networking sites such as tweeter, Facebook, YouTube and the like, toxic contents are found to be common now-a-days. Therefore, researchers have contributed a lot to identify toxic text and classify toxicity so that it can be eradicated and have positive communication in cyberspace.

Acknowledgment
The scholar authors would like to thank Professor (Dr.) M. K. Deka, for encouraging to write an article in positive communication.

Funding
This research did not receive any specific grant from any funding agencies.

Conflict of Interest
The authors declare no conflict of interest.

Author Contributions
Mr. Pranchis Narzaree: Conceptualization, Data curation, Methodology, Analysis, Writing – original draft.
Mr. Amkar Brahma: Methodology.
Prof. (Dr.) Manoj Kr. Deka: Supervising, Writing – review and editing.

Ethics approval
Not Applicable.

Data availability
Not Applicable

5. References

[1] https://datareportal.com/reports/digital-2024-india
[2] Huda, Miftachul. 2023, Trust as a key element for quality communication and information management: insights into developing safe cyber-organisational sustainability, International Journal of Organizational Analysis, 10.1108/IJOA-12-2022-3532. DOI:10.1108/IJOA-12-2022-3532
[3] Brunn, Stanley D. 2014. "Cyberspace Knowledge Gaps and Boundaries in Sustainability Science: Topics, Regions, Editorial Teams and Journals" *Sustainability* 6, no. 10: 6576-6603. https://doi.org/10.3390/su6106576
[4] Marshan, Alaa, Farah Nasreen Mohamed Nizar, Athina Ioannou, and Konstantina Spanaki. "Comparing Machine Learning and Deep Learning Techniques for Text Analytics: Detecting the Severity of Hate Comments Online." *Information Systems Frontiers*, November 24, 2023. https://doi.org/10.1007/s10796-023-10446-x.
[5] Kiritchenko, Svetlana, Isar Nejadgholi, and Kathleen C. Fraser. "Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective." *Journal*

of Artificial Intelligence Research 71 (July 15, 2021): 431–78. https://doi.org/10.1613/jair.1.12590.

[6] Desai, Aditya, Shashank Kalaskar, Omkar Kumbhar, and Rashmi Dhumal. "Cyber Bullying Detection on Social Media Using Machine Learning." Edited by M.D. Patil and V.A. Vyawahare. *ITM Web of Conferences* 40 (2021): 03038. https://doi.org/10.1051/itmconf/20214003038.

[7] Fan, Hong, Wu Du, Abdelghani Dahou, Ahmed A. Ewees, Dalia Yousri, Mohamed Abd Elaziz, Ammar H. Elsheikh, Laith Abualigah, and Mohammed A. A. Al-qaness."Social Media Toxicity Classification Using Deep Learning: Real- World Application UK Brexit." *Electronics* 10, no. 11 (June 1, 2021): 1332. https://doi.org/10.3390/electronics10111332.

[8] Sheth, Amit, Valerie L. Shalin, and Ugur Kursuncu. "Defining and Detecting Toxicity on Social Media: Context and Knowledge Are Key." *Neurocomputing* 490 (June 2022): 312– 18. https://doi.org/10.1016/j.neucom.2021.11.095.

[9] Teng, Teoh Hwai, and Kasturi Dewi Varathan. "Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches." *IEEE Access* 11 (2023): 55533–60. https://doi.org/10.1109/ACCESS.2023.3275130.

[10] Shrestha, Amendra, Lisa Kaati, Nazar Akrami, Kevin Linden, and Arvin Moshfegh. "Harmful Communication: Detection of Toxic Language and Threats on Swedish." In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, 624–30. Kusadasi Turkiye: ACM, 2023. https://doi.org/10.1145/3625007.3627597.

[11] Song, Guizhe, Degen Huang, and Zhifeng Xiao. "A Study of Multilingual Toxic Text Detection Approaches under Imbalanced Sample Distribution." *Information* 12, no. 5 (May 12, 2021): 205. https://doi.org/10.3390/info12050205.

[12] Mohammed, Ammar, and Rania Kora. "A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges." *Journal of King Saud University - Computer and Information Sciences* 35, no. 2 (February 2023): 757–74. https://doi.org/10.1016/j.jksuci.2023.01.014.

[13] Wang, Kehan, Jiaxi Yang, and Hongjun Wu. "A Survey of Toxic Comment Classification Methods." arXiv, 2021. https://doi.org/10.48550/ARXIV.2112.06412.

[14] Nithya, Dr D, Nanthine K S, Thenmozhi S, and Varshinipriya R. "Advanced Social Media Toxic Comments Detection System Using AI." *International Journal for Research in Applied Science and Engineering Technology* 12, no. 4 (April 30, 2024): 4719–24. https://doi.org/10.22214/ijraset.2024.61145.

[15] Madhyastha, Pranava, Antigoni Founta, and Lucia Specia. "A Study towards Contextual Understanding of Toxicity in Online Conversations." *Natural Language Engineering* 29, no. 6 (November 2023): 1538–60. https://doi.org/10.1017/S1351324923000414.

[16] Smith, Douglas, Trinity Leonis, and S. Anandavalli. "Belonging and Loneliness in Cyberspace: Impacts of Social Media on Adolescents' Well-Being." *Australian Journal of Psychology* 73, no. 1 (January 2, 2021): 12–23. https://doi.org/10.1080/00049530.2021.1898914.

[17] Ognibene, Dimitri, Rodrigo Wilkens, Davide Taibi, Davinia Hernández-Leo, Udo Kruschwitz, Gregor Donabauer, Emily Theophilou, et al. "Challenging Social Media Threats Using Collective Well-Being-Aware Recommendation Algorithms and an Educational Virtual Companion." *Frontiers in Artificial Intelligence* 5 (January 9, 2023): 654930. https://doi.org/10.3389/frai.2022.654930.

[18] Kiddle, Rupert, Petter Törnberg, and Damian Trilling. "Network Toxicity Analysis: An Information-Theoretic Approach to Studying the Social Dynamics of Online Toxicity." *Journal of Computational Social Science* 7, no. 1 (April 2024): 305–30. https://doi.org/10.1007/s42001-023-00239-2.

[19] Matamoros-Fernández, Ariadna, and Johan Farkas. "Racism, Hate Speech, and Social Media: A Systematic Review and Critique." *Television & New Media* 22, no. 2 (February 2021): 205–24. https://doi.org/10.1177/1527476420982230.

[20] Bhat, Meghana Moorthy, Saghar Hosseini, Ahmed Hassan Awadallah, Paul Bennett, and Weisheng Li. "Say 'YES' to Positivity: Detecting Toxic Language in Workplace Communications." In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2017–29. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.173.

[21] Khraisat, Ansam, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges." *Cybersecurity* 2, no. 1 (December 2019): 20. https://doi.org/10.1186/s42400-019-0038-7.

[22] Ptaszynski, Michal, Fumito Masui, Taisei Nitta, Suzuha Hatakeyama, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. "Sustainable Cyberbullying Detection with Category-Maximized Relevance of Harmful Phrases and Double-Filtered Automatic Optimization." *International Journal of Child-Computer Interaction* 8 (May 2016): 15–30. https://doi.org/10.1016/j.ijcci.2016.07.002.

[23] Lashkarashvili, Nineli, and Magda Tsintsadze. 2022, "Toxicity Detection in Online Georgian Discussions." *International Journal of Information Management Data Insights* 2, no. 1 (April 2022): 100062. https://doi.org/10.1016/j.jjimei.2022.100062., https://www.sciencedirect.com/science/article/pii/S2667096822000064

[24] D. Ramana kumar, P.Manideep Pasulad, Afreen Shaik, Nimish Reddy, 2020, Toxic Message Classification Using Convolution And Gru Network, International Journal of Research, e-ISSN: 2348-6848, p-ISSN: 2348-795X, Volume 07 Issue 05, May 2020