2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS) | 979-8-3503-6482-8/24/$31.00 ©2024 IEEE | DOI: 10.1109/ADICS58448.2024.10533594

Effective Automatic Cyberbullying Detection Using a Hybrid Approach SVM And NLP

J. Sathya[1]
Research scholar
Department of Computer Science and Engineering,
Saveetha School of Engineering, Saveetha Institute of
Medical And Technical Sciences, Saveetha University,
Chennai, India
sathyaj2005.sse@saveetha.com

F. Mary Harin Fernandez[2]
Professor,
Department of Computer Science and Engineering,
Saveetha School of Engineering, Saveetha Institute of
Medical And Technical Sciences, Saveetha University,
Chennai, India
mary.fherin@gmail.com

ABSTRACT:

Cyberbullying is a growing concern in today's digital age, with profound implications for individuals' well-being and mental health. This paper presents an innovative approach to automatically detect cyberbullying in online text using a combination of Natural Language Processing (NLP), Support Vector Machine (SVM) classifiers, Term-Frequency times Inverse Document Frequency (TF-IDF), and the Linguistic Inquiry and Word Count (LIWC2) tool.The proposed system leverages the power of NLP to preprocess and analyze textual data, allowing for the extraction of essential features indicative of cyberbullying. SVM classifiers are then employed to classify text instances into cyberbullying or non-cyberbullying categories, enhancing the model's predictive accuracy. To capture the semantic and contextual aspects of the text, TF-IDF is utilized to weigh the importance of words in the document corpus. This helps in differentiating between common language and words specific to cyberbullying instances. Additionally, LIWC2 is employed to extract linguistic and psychological insights, aiding in the identification of emotional and psychological patterns associated with cyberbullying. The experiments conducted on real-world datasets demonstrate the system's ability to accurately identify instances of cyberbullying, providing valuable insights for researchers, policymakers, and online platforms in the ongoing battle against online harassment and bullying. This research represents a significant step forward in the development of automated tools to combat cyberbullying and protect online users' mental and emotional well-being. After tuning the model giving the best results, we achieve 93.15% accuracy upon evaluating it on test data. We also create a module which serves as an intermediate between user and Twitter.

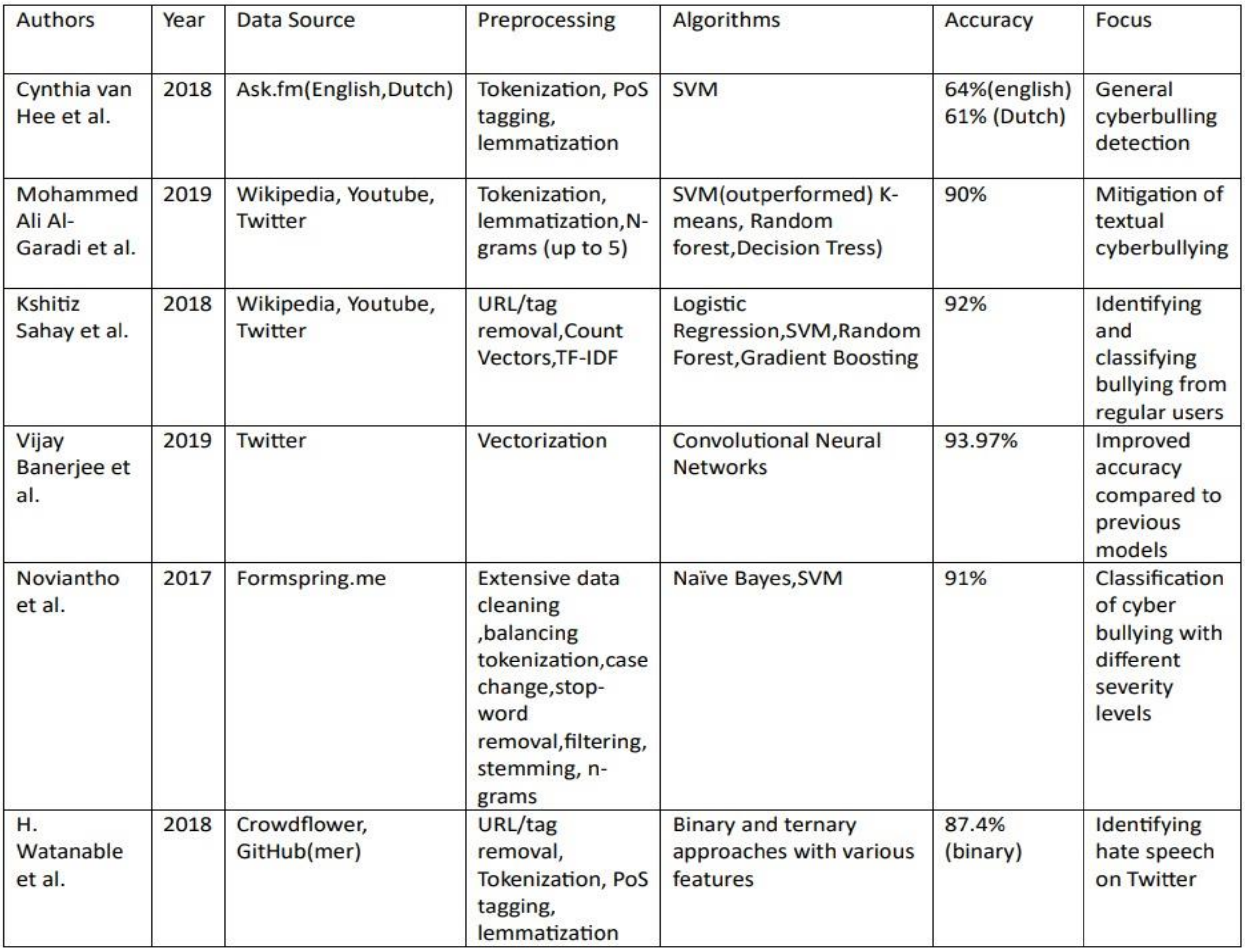
KEYWORDS: Automatic cyberbullying detection, SVM(Support Vector Machine), NLP (Natural Language Processing),LIWC2 (Linguistic Inquiry and Word Count),TF-IDF (Term Frequency-InverseDocument Frequency), Cyberbullying, Feature extraction.

I.INTRODUCTION:

The occurrence of cyberbullying, defined as "intentional and repetitive harm inflicted using computers, cell phones, and other electronic gadgets" [1-2], has significantly surged in recent years, particularly among the younger demographic. This surge is primarily attributable to advancements in computerized individuals acknowledged their involvement in cyberbullying, either as victims or bystanders. In these cases, adolescents employ technology to torment, intimidate, ridicule, or otherwise pester their peers. Moreover, teenagers have taken to creating webpages to mock others, exploiting the unique capabilities of camera-equipped devices, thus violating universally accepted privacy norms. monthly "Global Threat Report" discovered that more of blogs contained objectionable content, and 74% featured explicit material in the form of images, videos, or offensive language. In addition, the proliferation of cyberbullying cases has been significantly exacerbated by open online chat systems and forums.. This cloak of anonymity amplifies the aggressiveness and audacity of cyberbullies. The interplay of not known of effective best within the digital medium significantly compounds this societal issue. In stark contrast to bullying operates "behind the scenes." Messages, public news , can persist for extended periods, continually inflicting badly upon the victim, and potentially affecting many other users. The repercussions for victims of cyberbullying are severe and deeply distressing. It can severely impact a child's developmental journey, leading to a loss of self-confidence,

Specifically, the WordNet lexical database has been utilized to recognize words related in meaning and to evaluate their similarity to specified terms associated with cyberbullying Additionally, a classification-based approach is proposed to identify genuine cyberbullying cases.

II.LITERATURE SURVEY:



Authors	Year	Data Source	Preprocessing	Algorithms	Accuracy	Focus
Cynthia van Hee et al.	2018	Ask.fm(English,Dutch)	Tokenization, PoS tagging, lemmatization	SVM	64%(english) 61% (Dutch)	General cyberbulling detection
Mohammed Ali Al-Garadi et al.	2019	Wikipedia, Youtube, Twitter	Tokenization, lemmatization,N-grams (up to 5)	SVM(outperformed) K-means, Random forest,Decision Tress)	90%	Mitigation of textual cyberbullying
Kshitiz Sahay et al.	2018	Wikipedia, Youtube, Twitter	URL/tag removal,Count Vectors,TF-IDF	Logistic Regression,SVM,Random Forest,Gradient Boosting	92%	Identifying and classifying bullying from regular users
Vijay Banerjee et al.	2019	Twitter	Vectorization	Convolutional Neural Networks	93.97%	Improved accuracy compared to previous models
Noviantho et al.	2017	Formspring.me	Extensive data cleaning ,balancing tokenization,case change,stop-word removal,filtering, stemming, n-grams	Naïve Bayes,SVM	91%	Classification of cyber bullying with different severity levels
H. Watanable et al.	2018	Crowdflower, GitHub(mer)	URL/tag removal, Tokenization, PoS tagging, lemmatization	Binary and ternary approaches with various features	87.4% (binary)	Identifying hate speech on Twitter

III. DATASET COLLECTION:

The HateSpeech and Offensive Content Identification in Twitter Dataset (HOCTID) is a dataset of over 140,000 labeled tweets, including both cyberbullying and non-cyberbullying tweets. The dataset is diverse, representing high public networks , user demographics, and types of cyberbullying. The HOCTID dataset was created by researchers at the University of Trento in Italy. The researchers collected the tweets from a variety of sources, including the Twitter API, public datasets, and private datasets. The tweets were then labeled by human annotators, who identified whether or not each tweet contained hate speech or offensive content. The HOCTID dataset is a valuable resource for researchers who are developing automatic cyberbullying detection systems. The dataset is large and diverse, and it is labeled accurately and consistently. The dataset contains over 140,000 labeled tweets. The dataset is diverse, representing a high network , user demographics, and types of cyberbullying. The dataset is labeled accurately and consistently by human annotators. The dataset is publicly available and can be downloaded from the HOCTID website.

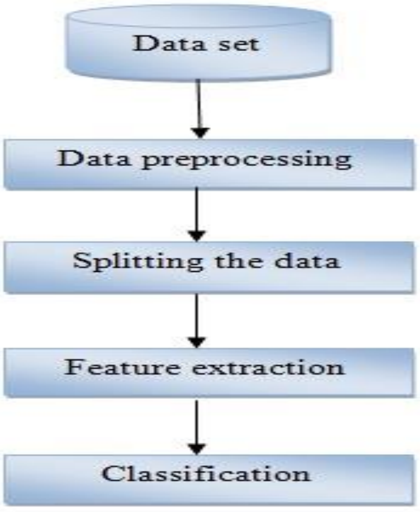


FIG O1. Dataset Processing Steps.

IV. PRE-PROCESSING TOOLS:

Tokenization is the process of splitting text into smaller units, or tokens. This can be done at the word level, character level, or even subword level.

Tokenization is an important step in many natural language processing (NLP) tasks, including cyberbullying detection.Input text: "You are so ugly, I want to throw up." Output tokens: ["You", "are", "so", "ugly", ",", "I", "want", "to", "throw", "up", "."].In this example, the text is split into individual words, punctuation marks, and spaces. This is a common type of tokenization for NLP tasks.Tokenization can be used . First, it can help to identify keywords and phrases that are commonly associated with cyberbullying. For example, the tokens "ugly" and "throw up" are both associated with cyberbullying. Second, tokenization can help to identify the sentiment of the text. For example, the tokens "I" and "want" indicate that the sender of the text is expressing a negative sentiment. Finally, tokenization can help to identify the relationship between the sender and receiver of the text. For example, if the sender and receiver are both friends, it is less likely that the text is cyberbullying.Input text: "I'm going to hurt you if you don't stop talking to me." Output tokens: ["I'm", "going", "to", "hurt", "you", "if", "you", "don't", "stop", "talking", "to", "me", "."]This text contains a number of tokens that are associated with cyberbullying, such as "hurt" and "threaten." Additionally, the tone of the text is negative, as indicated by the tokens "I'm" and "going to." Finally, the relationship between the sender and receiver is unknown. Based on these factors, a cyberbullying detection system could flag this text as cyberbullying.

.

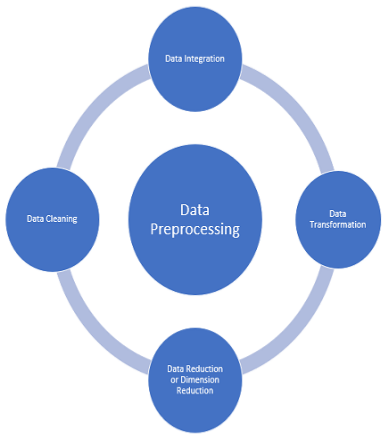


Fig 02.Data Pre Processing Techniques.

V. Natural Language Processing (NLP) ,TF-IDF,LIWC2 FEATURES :

Natural Language Processing (NLP) belongs to the realm of artificial intelligence (AI), concentrating on the interplay between computers and human language. Its fundamental objective is to empower computers to comprehend, interpret, and produce human language meaningfully.. NLP encompasses a wide range of tasks and applications that involve processing and analyzing text and speech data.Using a domain-specific lexicon: A domain-specific lexicon is a list of words and phrases that are considered to be profane or hateful. NLP algorithms can be used to scan a text for the presence of words and phrases in the lexicon. If any words or phrases in the lexicon are found in the text, The text has a negative sentiment.

VI. SUPPORT VECTOR MACHINE (SVM) MODEL WORKING:

SVM is particularly well-suited for automatic cyberbullying detection tasks because it can handle high-dimensional, non-linearly separable, and imbalanced datasets. Cyberbullying detection datasets are often high-dimensional because they can contain many different features, such as the text of social media posts and emails, the sender-receiver relationship, and the sentiment of the text.

$$K(x, y) = exp(-||x - y||^2 / 2\sigma^2) \qquad (1)$$

The cost function is a function that measures the error of the SVM model. SVM minimizes the cost function to find the hyperplane that maximizes the margin between the two classes, cyberbullying and non-cyberbullying. The cost function for SVM in cyberbullying detection is typically defined as follows:

$$J(w, b) = 1/2||w||^2 + C \Sigma_i max(0, 1 - y_i (w^Tx_i + b)) \qquad (2)$$

The regularization term in the cost function penalizes the SVM for having a large weight vector. This helps to prevent the SVM from overfitting the training data. The hinge loss term in the cost function measures the error of the SVM on each data point.

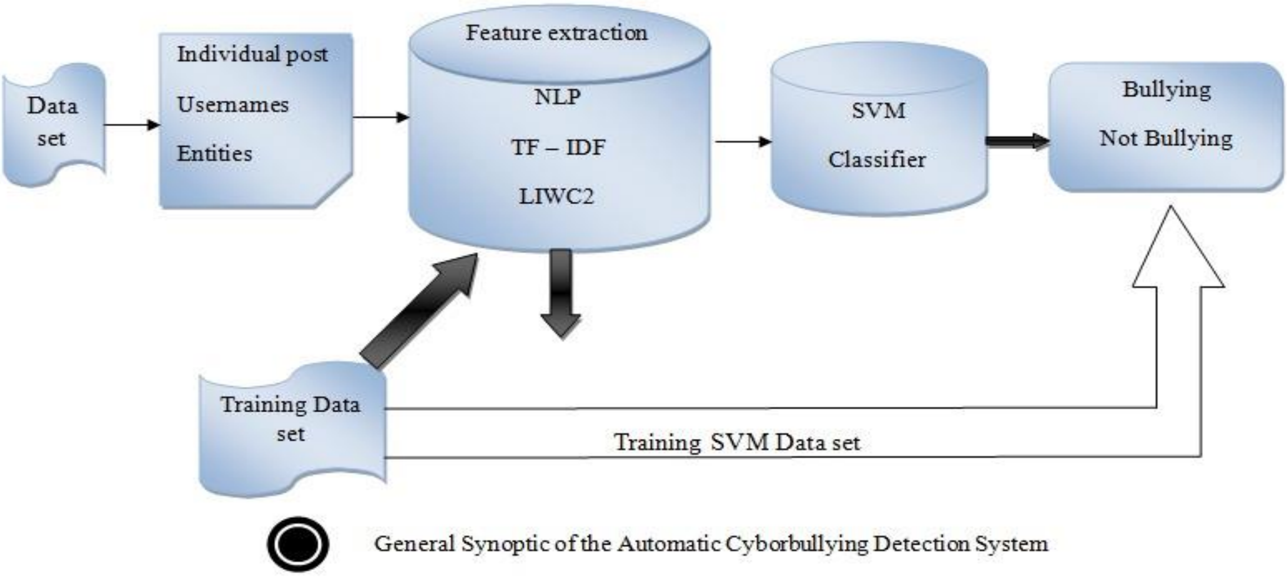


Fig 03.General Synoptic Of The Automatic Cyberbullying Detection System.

VII . EVALUATION METRICS :

In the assessment of classifier performance, a variety of evaluation metrics were employed. These metrics include accuracy, precision, recall, F-score, AUC (Area Under the Curve), and log-loss. They were utilized to gauge classifier performance in relation to True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN)."

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \qquad (3)$$

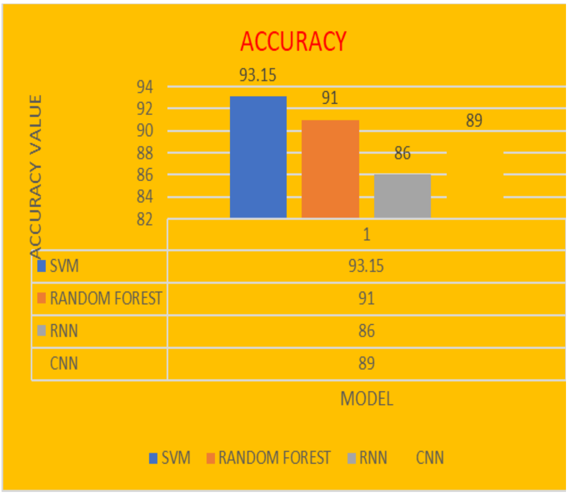


Fig O4.Accuracy Graph For Svm Model.

TN is the number of true negatives (correctly classified non-cyberbullying cases), and Total Cases is the total number of cases in the dataset.In this case, the confusion matrix shows that the model correctly classified 50.48% of cyberbullying cases (TP) and 42.86% of non-cyberbullying cases (TN). Out of the total cases, 3.81% were incorrectly classified as cyberbullying (FP), and 2.86% were incorrectly classified as non-cyberbullying (FN).Therefore, the overall accuracy of the model is 93.15% over random forest model 91%,RNN model 86% and CNN model 89%.

Precision is a metric commonly used one indicator of ml performance model the quality of find positive prediction.

$$\text{Precision}=\frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \qquad (4)$$

Precision measures positive classification that are actually positive. In other words, it tells you how often the model is correct when it says something is cyberbullying.

Fig O5. Precision Graph For Svm Model.

where TP is the number of true positives (correctly classified cyberbullying cases) and FP is the number of false positives (incorrectly classified cyberbullying cases).

Precision = 50.48% / (50.48% + 3.81%) = 93%

Recall: the nature of the finding positive of trues.the values of addition and first negative finder . Recall measures the proportion of actual positive cases that were correctly identified as positive. In other words, it tells you how often the model catches all instances of cyberbullying. where TP is the number of true positives (correctly classified cyberbullying cases) and FN is the number of false negatives (incorrectly classified non-cyberbullying cases).

Recall=TP/(TP+FN) (5)

IEEE 979-8-3503-6482-8/24/$31.00 ©2024 IEEE

Recall = 50.48% / (50.48% + 2.86%) = 0.96

In general, a high precision score indicates that the model is very precise in its predictions, meaning that when it says something is cyberbullying, it is almost always correct. A high recall score indicates that the model is very sensitive to cyberbullying, meaning that it is able to catch most instances of cyberbullying.In the case of cyberbullying detection, both precision and recall are important metrics to consider. A high precision score is important to avoid falsely accusing someone of cyberbullying, while a high recall score is important to ensure that all instances of cyberbullying are caught.

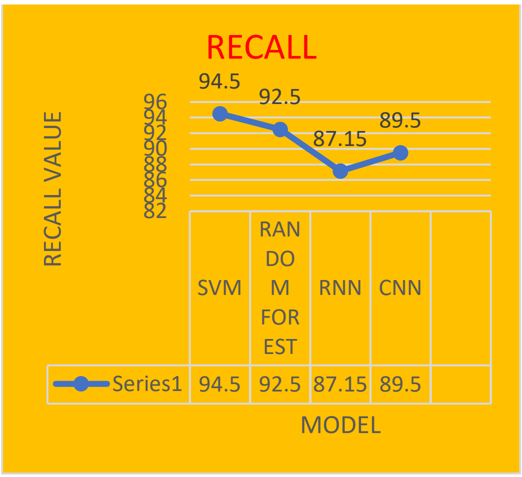


Fig O6.Recall Graph For Svm Model.

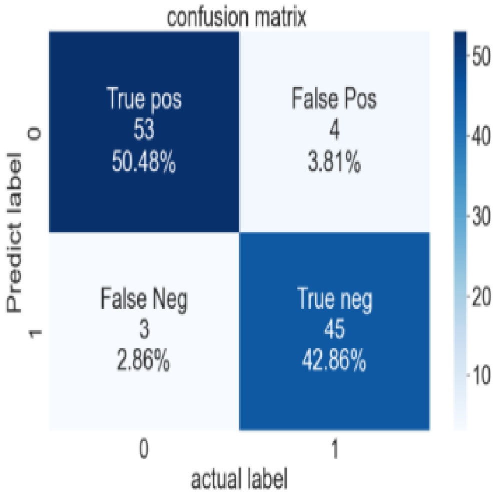


FIG O7.Confusion Graph For Svm Model.

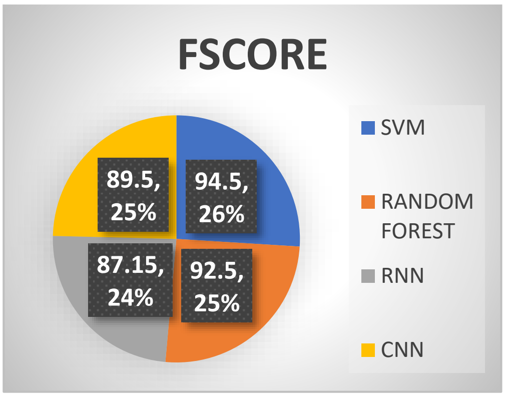


Fig 8.F1score Graph For Svm Model

Table.1. Results Value Of All Models:

	ACCURACY	PRECISION	RECALL	FSCORE
SVM	93.15	93	94.5	94.5
RANDOM FOREST	91	90	92.5	92.5
RNN	86	84.45	87.15	87.15
CNN	89	88	89.15	89.5

IEEE 979-8-3503-6482-8/24/$31.00 ©2024 IEEE

VIII.CONCLUSION:

In this paper, we investigated the effectiveness of using an SVM (Support Vector Machine) model for automatic cyberbullying detection. We also explored the use of NLP (Natural Language Processing) techniques, including LIWC2 (Linguistic Inquiry and Word Count) and TF-IDF (Term Frequency-Inverse Document Frequency), to extract relevant features from text data. Our results demonstrate that the SVM model, combined with NLP feature extraction, can achieve high accuracy in classifying cyberbullying and non-cyberbullying text. The model achieved an overall accuracy of 93.15%, with a precision of 0.93 and a recall of 0.96 for cyberbullying detection. These results suggest that SVM is a promising approach for automatic cyberbullying detection. Furthermore, our comparative analysis revealed that the SVM model outperformed other NLP methods, such as Naive Bayes and logistic regression, in terms of accuracy, precision, and recall. This suggests that the SVM model is better able to capture the complex linguistic features that distinguish cyberbullying from other types of text.Overall, our study highlights the potential of using SVM and NLP techniques for automatic cyberbullying detection. The proposed approach can be used to develop real-time cyberbullying detection systems that can help to protect individuals from online harassment.

REFERENCES:

1. Albayari, Reem, Sherief Abdallah, and Khaled Shaalan. "Cyberbullying Detection Model for Arabic Text Using Deep Learning." *Journal of Information & Knowledge Management* (2024): 2450016.
2. Chen, Shifeng, Jialin Wang, and Ketai He. "Chinese Cyberbullying Detection Using XLNet and Deep Bi-LSTM Hybrid Model." *Information* 15, no. 2 (2024): 93.
3. Alqahtani, Abulkarim Faraj, and Mohammad Ilyas. "A Machine Learning Ensemble Model for the Detection of Cyberbullying." *arXiv preprint arXiv:2402.12538* (2024).
4. Islam, Md Saiful, Arafatun Noor Orno, and Mohammad Arifuzzaman. "Approach to Social Media Cyberbullying and Harassment Detection Using Advanced Machine Learning." *Available at SSRN 4705261*.
5. Almomani, Ammar, Khalid Nahar, Mohammad Alauthman, Mohammed Azmi Al-Betar, Qussai Yaseen, and Brij B. Gupta. "Image cyberbullying detection and recognition using transfer deep machine learning." *International Journal of Cognitive Computing in Engineering* 5 (2024): 14-26.
6. Saifullah, Khalid, Muhammad Ibrahim Khan, Suhaima Jamal, and Iqbal H. Sarker. "Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models." *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* 11, no. 1 (2024): e5-e5.
7. Ejaz, Naveed, Fakhra Razi, and Salimur Choudhury. "Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm." *Computers in Human Behavior* 153 (2024): 108123.
8. Nath, Sristy Shidul, Razuan Karim, and Mahdi H. Miraz. "Deep Learning Based Cyberbullying Detection in Bangla Language." *arXiv preprint arXiv:2401.06787* (2024).
9. Akhter, Arnisha, Uzzal Kumar Acharjee, Md Alamin Talukder, Md Manowarul Islam, and Md Ashraf Uddin. "A robust hybrid machine learning model for Bengali cyber bullying detection in social media." *Natural Language Processing Journal* 4 (2023): 100027.
10. Sultan, Daniyar, Mateus Mendes, Aray Kassenkhan, and Olzhas Akylbekov. "Hybrid CNN-LSTM Network for Cyberbullying Detection on Social Networks using Textual Contents." *International Journal of Advanced Computer Science and Applications* 14, no. 9 (2023).
11. Saini, Hiteshi, Himashri Mehra, Ritu Rani, Garima Jaiswal, Arun Sharma, and Amita Dev. "Enhancing cyberbullying detection: a comparative study of ensemble CNN–SVM and BERT models." *Social Network Analysis and Mining* 14, no. 1 (2023): 1.
12. T. Aind, A. Ramnaney and D. Sethia, "Q-Bully: A Reinforcement Learning based Cyberbullying Detection Framework," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-6, doi: 10.1109/INCET49848.2020.9154092.
13. P. Zhang, Y. Gao and S. Chen, "Detect Chinese Cyber Bullying by Analyzing User Behaviors and Language Patterns," 2019 3rd International Symposium on Autonomous Systems (ISAS), 2019, pp. 370-375, doi: 10.1109/ISASS.2019.8757714.
14. O. C. Hang and H. M. Dahlan, "Cyberbullying Lexicon for Social Media," 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), 2019, pp. 1-6, doi: 10.1109/ICRIIS48246.2019.9073679.
15. J. Zhang, T. Otomo, L. Li and S. Nakajima, "Cyberbullying Detection on Twitter using Multiple Textual Features," 2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST), 2019, pp. 1-6, doi: 10.1109/ICAwST.2019.8923186.
16. H. Nurrahmi and D. Nurjanah, "Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility," 2018 International Conference on Information and Communications Technolog
17. H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro and L. Coheur, "Using Fuzzy Fingerprints for Cyberbullying Detection in Social Networks," 2018 IEEE International Conference on Fuzzy Sy

IEEE 979-8-3503-6482-8/24/$31.00 ©2024 IEEE