A Review of Different Approaches to Detect Online Cyberbullying and Hate Speech

K.M. Chaman Kumar[1*], Parth Mahajan[2], Tejas Naik[3], Navjyot Naik[4] and Aryan Gaonkar[5]

[1, 2,3,4,5] Dept. of Computer Engineering, SRIEIT, Goa University, Goa, India
*chaman_k2007@yahoo.co.in

Abstract. With the rapid rise in social media usage, platforms such as X(Twitter), Instagram and Facebook have become prevalent spaces for online interaction. These platforms allow users to share different forms of media such as text, image, video, etc. However, these platforms have also become a hotspot for harmful activities such as cyberbullying and hate speech. While these platforms employ solutions like machine learning models for text-based detection and deep learning, their ability to handle multi-modal content like images and videos remains limited. Our proposed solution introduces a web-based application that integrates multi-modal data analysis across the four major platforms: X(Twitter), Instagram and Facebook.

Keywords: Cyberbullying, Inception V3, Support Vector Machine, Convolutional Neural Network.

1 Introduction

Online bullying and hateful messages have become more prevalent on social media sites in our current era. These platforms often let users stay anonymous and don't have many rules, which has allowed harmful behaviour to spread without checks. Social media platforms, especially Twitter, have become breeding grounds for cyberbullying, particularly among vulnerable groups such as adolescents [9]. This leads to mental harm, emotional pain, and sometimes even serious real-world consequences. As more people use social media, we need better ways to spot and stop online bullying and hate speech that work and can handle lots of content.

The mental and emotional costs of cyberbullying and hate speech run deep. Victim teens often feel anxious, depressed, and in bad cases, think about suicide [18]. A Pew Research Centre study found that 59% of U.S. teens faced online harassment in 2018. More social media use during the COVID-19 lockdowns made things worse, with 56% of U.S. cyberbullying cases happening during this time [10]. Hate speech has also grown, having an impact on political talk and making social divides bigger.

S. Bhalerao et al. (eds.), *Proceedings of the International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2025),* Advances in Intelligent Systems Research 192,
https://doi.org/10.2991/978-94-6463-716-8_76

Creating automated detection systems comes with its share of hurdles. Online bullying and hateful comments often depend on context. They involve sarcasm, cultural references, slang, and ever-changing language trends. This makes it tough for models to grasp these complex behaviours [29]. On top of that, each platform has its own way of creating, organizing, and sharing content. This adds extra layers of difficulty, as models need to adapt to different settings and data types [32]. The rise of content that mixes different forms, like memes and videos also calls for the ability to handle tricky interactions between words and images. All these factors combined mean that automated systems must be flexible and advanced enough to deal with various inputs across different media types [19]. Additionally, annotating large datasets for training is a difficult task, leading some papers to propose transfer learning approaches in handling this task [22].

Recent breakthroughs in transfer learning, ensemble learning, and multimodal analysis have led to more advanced detection systems. Transfer learning models allow experts to fine-tune pre-trained models for specific tasks, which cuts down the need for big labelled datasets. Ensemble models merge the strong points of several classifiers to boost accuracy. Multimodal approaches bring together text and image data to get a fuller picture of online content [11]. This literature survey aims to look into and review the main models and methods used to spot cyberbullying and hate speech. It shines a light on what these models do well where they fall short, and how they perform with different sets of data. By taking a close look at studies that use models like InceptionV3, Random Forest, Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), Contrastive Language-Image Pre-Training (CLIP), and stack ensemble methods, this survey wants to give a full picture of where the field stands right now. The results of these studies will guide us to create a new system that can spot problems in real time. This system will work on many social media sites and adjust to different types of content. It will help us keep up with how online talks change and step in when needed.

Next, we'll look at the different machine learning and deep learning models used to spot cyberbullying and hate speech. We'll talk about how they work, what data they use, what they found, what's good about them, and what needs work. This review will help us see what's missing in current research and suggest ways to make automated detection systems stronger and able to grow.

2 Literature Review

2.1 Support Vector Machines (SVMs)

SVMs are effective supervised machine learning algorithms primarily used for classification but can also perform regression. In SVM’s approach as shown in table 1, a hypersurface which best divides a set of objects in a high dimensional space is determined, while increasing the distance between this surface and the closest objects of any category (support vectors). This method is particularly efficient in cases of linear separability, but can also be extended to non-linear cases using kernel functions to project data into another input space [1]. In [2], the performance of SVM was evaluated against other models including logistic regression and K-nearest neighbors. SVM also showed accuracy in languages inclusive of English, Tamil and Malayalam where [3] combined SVM and logistic regression with TF-IDF to attain 74.50% accuracy in identifying instances of cyberbullying. Thus, SVM is commonly used in cyberbullying detection because of its adeptness that wades through high dimensional imbalanced text data. For instance, the extent of the cyberbully was estimated in [1] through the use of SVM – which would include features such as TF-IDF. In their latest research, Bozyigit et al. [5] also stayed with SVM as one of the approaches but included other machine learning methods and noted that the performance in identifying cases of cyberbully was enhanced by adding text descriptions with social network characteristics (such as user activities). On the other hand, Elsafoury et al. [21] observed that SVMs, when enhanced with N-grams and other techniques of feature extraction, are very effective in classifying textual data, including the classification of cyberbullying.

Table 1. SVM

Ref No.	Methodologies	Datasets Used	Tools and Technology	Merits	Demerits	Results
[1]	Linear SVM	Twitter	Not Specified	Auto annotation	Annotation focus	Baseline: 91%, Retrain: 95%
[2]	SVM, CNN	YouTube	Comment scraper, sklearn, Langdetect	Multilingual	Class imbalance	CNN: Best results
[3]	SVM	Twitter	NLP Tools, HTML, CSS, Bootstrap	Language flexible	Text-only focus	F1-score: 74%
[5]	SVM, AdaBoost	Twitter	Scikit-learn Spyder, Twitter API	High accuracy	Limited generalization	AdaBoost accuracy: ~90.1%

2.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT, in table 2 is a framework for understanding languages which looks at the context of words both in left and right directions, hence making it more effective for natural language processing (nlp) tasks such as question answering, text classification and sentiment analysis. Designed by Google, BERT allows understanding of words based on the meaning of words surrounding them owing to its bi-directional structure. The use of BERT in cyberbullying detection has proven to be very effective. For example, BERT in [26] fine-tuned for the Upwork dataset concerning inter-user abuse in Twitter, Formspring, Wikipedia and several works drew the best findings possible. This research resolved the issue of class imbalance by large datasets and oversampling techniques and provided evidence on BERT's ability to effectively identify bullying content in many sources. In another study, the linear neural network layer with BERT's pre-trained word vectors was used for the classification of the comments in cyberbullying, the efficiency of which was higher in comparison with the previous machine learning methods proposed for the datasets of Wikipedia and Formspring. BERT's higher performance was observed compared to GloVe embedding based models in dealing with complex abusive language phenomena [27]. Furthermore, an adapted version of BERT known as COVID-HateBERT was also introduced to help identify hate speech during the pandemic and was found to be better than previous models in identifying contextual hate speech, hence demonstrating BERT's flexibility to the changing social conditions [15].

Table 2. BERT

Ref No.	Methodologies	Datasets Used	Tools and Technology	Merits	Demerits	Results
[26]	BERT	Twitter, Wikipedia, Formspring	PyTorch, SMOTE, Scikit-learn	High accuracy	Data issues	F1: 0.94, 0.91, 0.92
[27]	Pretrained & Fine-tuned BERT	Formspring, Wikipedia	Hugging Face, PyTorch, Python	Context accuracy	Class imbalance	F1: 0.94, 0.81
[15]	BERT	HatEval 2019, COVID-HATE Dataset	Hugging Face, Adam Optimizer	Outperforms models	Bias, multilingual	F1: 69.59%

2.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks – or CNNs – are a category of deep learning architecture that is predominantly used to analyze visual data shown in table 3. The architecture of CNNs features convolutional layers designed to process input images by applying filters to learn about spatial characteristics like edges and textures, which is why CNNs are widely used in applications such as object detection and nowadays even speech and social content analysis. For instance, in detecting hate speech in memes which involves integrating images and text, CNNs are known to process the images [31]. CNNs have also been more than capable of handling text by regarding the text being a series of ‘pixels’ or features in which this enables the model to detect the changes in patterns over periods of time. Results in [28] demonstrated that with the help of other processes in Natural Language Processing (NLP), detection of hate speech and its classification was possible due to the localized features the CNNs possess for the illustrations, which is very important when dealing with obscured and indirect expressions that are often found in hate speech. Such methods have improved the detection of hate speech in networks such as Twitter. In addition, the applicability of CNNs in cyberbullying detection has been successful. As raised in [17], when combining CNNs with other machine learning algorithms, CNNs were able to recognize the very delicate elements present in the text, which made them very useful in deep meaning comprehension tasks. In addition, the study investigated importation of model changes relative to the datasets used, which increased accuracy in categorization. Finally, these results consolidate the encouragement to use CNNs in processing either still images or written text, particularly, in social media text and images with malicious intent.

Table 3. CNN

Ref No.	Methodologies	Datasets Used	Tools and Technology	Merits	Demerits	Results
[31]	TRN, R-CNN	MSCOCO, Hateful Memes	PyTorch, MMF	Image-captioning	Weak alignment	AUROC: 78.86
[28]	CNN, LSTM, BERT	SemEval, AMI, COVID-HATE	PyTorch	Methods overview	Annotation issues	BERT > ML models
[17]	CNN, ANN	YouTube Comments	Scikit-learn, TensorFlow	Human-AI synergy	Small dataset	ANN: 96% Accuracy

2.4 Contrastive Language-Image Pre-training (CLIP)

CLIP is an innovative and new-age model OpenAI has developed for multimodal tasks (text and image actions) simultaneously. The model is trained to associate images with corresponding text thereby performing effectively in cases such as memes which utilizes interaction between text and picture. This model is very good in generalized learning; hence it is able to classify unseen images and texts. In [8], application of CLIP techniques on the Facebook Hateful Meme dataset was reported. The model achieved an accuracy of 87.42% indicating efficacy in hateful meme detection. GABRI, due to its capability in minimizing false positives, proved to be beneficial in the evaluation of various datasets but required modifications to become effective across different dataset types. One of the characteristics of CLIP as shown in table 4 that makes it popular for social media content analysis where hate speech detection is required is its comparative learning approach that aids in understanding word, picture connections even with complex relationships.

Table 4. CLIP

Ref No.	Methodologies	Datasets Used	Tools and Technology	Merits	Demerits	Results
[8]	CLIP	Facebook Hateful Meme	PyTorch, Torchvision	False-positive robust	Limited generalization	Acc: 87.42 %, AUROC: 88.35
[14]	DBSCAN, CLIP	4chan /pol/ dataset	Perspective API, Rewire	Tracks meme evolution	Some false positives	Found 3,321 meme variants

2.5 Other Models

The issue of cyberbullying has led researchers to investigate other machine learning solutions. Image-based cyberbullying detection yielded an accuracy of 89% using Inception V3 which outperformed simpler millimetric's constructed from laid CNNs [6]. Traditional approaches also utilize Naive Bayes (NB) which, although very easy to understand has been implemented to a great extent, many times acting as a reference point in related studies. Logistic Regression (LR), on its own, has not been very effective, but when combined with feature engineering as in the case of implied threat detection [23], where an F1 of 77.13% was achieved. AdaBoost also worked well with other miscellaneous classifiers however it was the best among such classifiers in terms of accuracy (90.1%) for classifying Turkish tweets with twitter attributes [5]. The Gradient Boosting method has been tried, though studies indicate that it may find difficulty in predicting accurately as compared to the Random Forest model [20]. These models shown in table 5 and table 6 provide different trade-offs between speed and accuracy of detection, with average performance of ensemble methods being robust across dataset and languages.

Table 5. Other Models

Ref No.	Methodologies	Datasets Used	Tools and Technology	Merits	Demerits	Results
[6]	VGG16, Inception V3, CNN	Image datasets	Python, Keras, TensorFlow	Better than CNN	Limited data, labeling issues	Acc: 89% (Inception V3), 67% (CNN)
[23]	SVM, LR, NB, AdaBoost, RF, DL	Machine-gen sentences	ChatGPT, Google Colab	Includes threats	Biased negative class	DNN F1: 91.49%, Bi-LSTM F1: 91.61%
[5]	SVM, LR, KNN, NBM, AdaBoost, RF	Twitter	Scikit-learn, Spyder, Twitter	High accuracy	Limited generalizability	AdaBoost Acc: 90.1%

Table 6. A summary table of different performance metrics

Ref No	Model Used	Performance Metrics
[1]	SVM	Accuracy 91%
[2]	CNN	F1 Score 75%
[3]	LSTM	F1 Score 93.6% and
[4]	Random Forest Classifier	Accuracy 96%
[5]	AdaBoost	Accuracy 90.1%
[6]	Inception V3	Accuracy 89%
[8]	CLIP	Accuracy 87.42%, AUROC 88.35%

[11]	SVM, LR	Accuracy between 83% and 97%
[12]	LSTM	Accuracy 98%
[15]	BERT	F1 Score 69.59%
[17]	ANN	Accuracy 96%
[21]	MLP	AUROC 0.95
[23]	BiLSTM	F1 Score 91%
[26]	BERT	F1 Score 94%
[27]	BERT	F1 Score 94%S
[31]	TRN	AUROC 0.7886

3 Dataset Description and Inference

The field of cyberbullying and hate speech detection now benefits from diverse datasets from social media platforms, each with unique attributes and challenges. MMHS150K, one of the largest sources with 150k tweets that contain both images and text, has enabled subsequent investigations employing transfer learning architectures, such as VGG16 and InceptionV3, for the identification of harmful content [6]. This dataset has also availed well the social media posts that are rich in both text and illustrations. In [4], a small YouTube dataset containing 300 videos that were considered either harmful or non-harmful in hate speech detection tasks was reported. The limited scope of this particular dataset shows the struggles associated with acquisition of labelled video content. In 2023, a dataset comprising 131,867 entries from Twitter, YouTube, and Facebook was published by Elsevier facilitating enhancement of detection systems that help in identifying content across multiple platforms [7]. Used in the IEEE 2024 paper [8], the Facebook Hateful Meme Dataset contains over ten thousand images with pairs of text and images illustrating the more complicated side of hate found in memes. Studies making use of twitter datasets are always in high demand due to their ephemeral nature. A BullyingV3.0 dataset that comprises 3889 tweets and a hate speech dataset composed of 10,360 tweets used in [11] are indicative of the less frequent occurrence of abusive language in everyday social media conversations. Additional social media platforms include MySpace with 2029 posts and ASKfm with 90296 posts, while the Wikipedia Talk Pages dataset with 115,864 comments documents instances of aggressive behavior in formal conversations. AMiCA, a ASKfm dataset, comprising 112,247 posts was widely used by many of the papers that used machine learning [16]. [25] used a manually annotated dataset consisting of 111,561 video clips from YouTube, which were equally distributed as safe and unsafe.

Practical measures to mitigate online abuse incidences on a national and state level require the use of multilingual datasets such as the one in [12] consisting of English (34,200), Hindi (16,100) and Hinglish (20,300) samples. One more cognitive variation comprises studying the episodes of cyberbullying on social media across a number of sessions, which has been applied in the analysis of the growth of abusive content using datasets from Instagram (2,218 sessions) and Vine videos (970 sessions) in [13]. Specific datasets assist in addressing some forms of hate, for example, the pol 4chan dataset with 12.5 million image text pairs was used in [14] to understand the evolution of memes and another dataset comprised of 1.27 million tweets about COVID-19 related hate was explored in [15].

Multimodal content, rare negative instances in imbalanced classes, and sensitive information from personal records are some of the typical issues present in the above datasets. Nonetheless, these datasets remain important enablers of advancing the development of automatic detection systems. It is common in this area to see deep learning models perform better than non-learning approaches because of the socially, cultured and visually rich content of social media. BERT-based models process real-world inputs with minimal constraints and variations in language with F1 scores over 0.90 but are very resource intensive. The existing systems also come with hurdles such as the existence of skewed data as seen in [10], [15] and the challenges posed by multimodal content. In studies that employed meme based hate speech detection, for example, the CLIP and Inception V3 models accuracy was reported to be 87.42% and 89% respectively [6]. Although this also helps alleviate the effects of the absence of labelled data, AdaBoost was able to accomplish 90.1% accuracy with no prior training [5]. The addition of model adjustment is sufficiently important to consider in research where data is little and particularly in cases such as cyberbully detection in different contexts [30]. Session based datasets, which capture the context of entire sessions rather than isolated posts, were very few but could potentially increase accuracy [24].

4 Proposed Methodology

Our approach shown in fig. 1 employs state-of-the-art models for each content type to effectively recognize cyberbullying and hate speech. In case of text content, BERT is employed for contextualizing the aggressor's language [9][26][15] and SVM for high-dimensional text classification [1][2][3] along with Convolutional Neural Network with Bidirectional Long Short-Term Memory (CNN-BiLSTM) that provides combination of both CNN based and BiLSTM based feature comprehension [21][17]. The best model from among those found optimized against the available datasets on cyberbullying will finally be integrated into the system. In terms of content that is image-based, such as memes, the analysis takes Inception V3 at hand on strong image recognition [6][31] and models for assessing image-text correspondence to determine whether an image is offensive or not [8][14]. The final model will then be selected on how it would detect the offensive materials. For video content the system includes EfficientNet-B7 with BiLSTM for efficient processing of short videos and for superior performance [4][31].

The integration of these different high-performance models under each modality would be like the use of BERT, SVM or CNN-BiLSTM for text, while Inception V3 or an image-text correspondence model would use images, and EfficientNet-B7 with BiLSTM would process videos. Every model here thus takes care of a single type of content while serving the very clear goal of effective detection for cyberbullying and hate speech, but this resource-efficient approach allows processing each type with the most suitable model rather than developing them equally. Also, this modularity enables each of the components to be independently optimized, thereby improving performance. With those advanced models, this technique ensures that offensive content is detected accurately using fewer resources, which is a complete and practical approach to the electric base to address online toxicity. Let's face it, society has not been optimal in applying terms to describe certain behaviors.

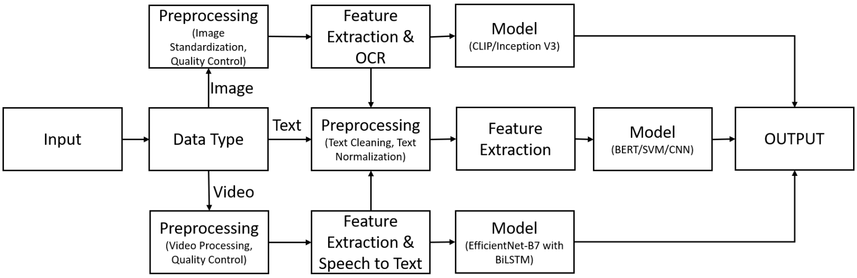
Preprocessing (Image Standardization, Quality Control)
Feature Extraction & OCR
Model (CLIP/Inception V3)
Image
Input
Data Type
Text
Preprocessing (Text Cleaning, Text Normalization)
Feature Extraction
Model (BERT/SVM/CNN)
OUTPUT
Video
Preprocessing (Video Processing, Quality Control)
Feature Extraction & Speech to Text
Model (EfficientNet-B7 with BiLSTM)

Fig. 1. Multimodal Cyberbullying and Hate Speech Detection Pipeline

5 Conclusion

In examining the body of work as presented in various research papers, it becomes evident that technological advancements in the prevention of cyberbullying and detection of hateful content can be useful to some degree, but most often put into use lose efficiency when it comes to content that is dynamic, coexisting and diverse, such as that exhibited in social media. It is apparent that traditional machine learning's approaches, although they are viewed as basic and fairly easy to use, are not able to capture the complexities that come with text, images and videos. With studies showing relatively improved detection of nuanced forms of offensive content using advanced deep learning models such as BERT and CLIP contextual and multi-modal capabilities.

6 Future Work

Analysing existing methodologies reveals gaps and opportunities for future research in the field of cyberbullying and hate speech detection. First, it is important to proceed with multimodal analysis, since existing systems mostly do not keep image and video reporting along with text. Such holistic frameworks may prove itself all the more useful in formulating detection methods. Second, the field stands to gain from much larger and diverse datasets across many languages to capture cultural nuances and social contexts that would enable models to generalize better across platforms and regions; more than this, there's definitely a need for improved techniques of data annotation to minimize bias and have representativeness. In the end, they could provide some insight into detection systems, perhaps sensing contextual and time-related features. For example, what a session-based analysis of so many different events over time, instead of individual events, might reveal as an abusive pattern. Another issue requiring continuing attention might include examining ethical questions of user privacy and reduced, unintended biases. Such possible directions might significantly strengthen the durability, scalability, and fairness of cyberbullying and hate speech detection systems.

References

1. Phanomtip, A., Sueb-in, T., & Vittayakorn, S. (2021, May). Cyberbullying detection on Tweets. In 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON) (pp. 295-298). IEEE.
2. Chakravarthi, B. R. (2022). Hope speech detection in YouTube comments. Social Network Analysis and Mining, 12(1), 75.
3. Perera, A., & Fernando, P. (2021). Accurate cyberbullying detection and prevention on social media. Procedia Computer Science, 181, 605–611.
4. Wu, C. S., & Bhandary, U. (2020). Detection of hate speech in videos using machine learning. In 2020 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE.
5. Bozyiğit, Alican, Semih Utku, and Efendi Nasibov. "Cyberbullying detection: Utilizing social media features." Expert Systems with Applications 179 (2021): 115001.
6. Roy, Pradeep Kumar, and Fenish Umeshbhai Mali. "Cyberbullying detection using deep transfer learning." Complex & Intelligent Systems 8.6 (2022): 5449-5467.
7. Kannammal, A., S. Omprakash, and J. Danesh Dheerthan. "Automated Decision Support System for Cyberbullying Detection." Procedia Computer Science 230 (2023): 760-768.
8. Arya, G., Hasan, M. K., Bagwari, A., Safie, N., Islam, S., Ahmed, F. R. A., ... & Ghazal, T. M. (2024). Multimodal Hate Speech Detection in Memes Using Contrastive Language-Image Pre-Training. IEEE Access.
9. Aliyeva, Çinare Oğuz, and Mete Yağanoğlu. "Deep learning approach to detect cyberbullying on twitter." Multimedia Tools and Applications (2024): 1-24.
10. Al-Harigy, L. M., Al-Nuaim, H. A., Moradpoor, N., & Tan, Z. (2022). Building towards automated cyberbullying detection: A comparative analysis. Computational Intelligence and Neuroscience, 2022(1), 4794227.
11. Gutiérrez-Batista, Karel, Jesica Gómez-Sánchez, and Carlos Fernandez-Basso. "Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model." Social Network Analysis and Mining 14.1 (2024): 136.
12. Raj, M., Singh, S., Solanki, K., & Selvanambi, R. (2022). An application to detect cyberbullying using machine learning and deep learning techniques. SN computer science, 3(5), 401.

13. Yi, Peiling, and Arkaitz Zubiaga. "Learning like human annotators: Cyberbullying detection in lengthy social media sessions." Proceedings of the ACM Web Conference 2023.
14. Qu, Y., He, X., Pierson, S., Backes, M., Zhang, Y., & Zannettou, S. (2023, May). On the evolution of (hateful) memes by means of multimodal contrastive learning. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 293-310). IEEE.
15. Li, M., Liao, S., Okpala, E., Tong, M., Costello, M., Cheng, L., ... & Luo, F. (2021, December). COVID-HateBERT: A pre-trained language model for COVID-19 related hate speech detection. In 2021 20th IEEE international conference on machine learning and applications (ICMLA) (pp. 233-238). IEEE.
16. 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): 55533-55560.
17. Gomez, Christopher E., Marcelo O. Sztainberg, and Rachel E. Trana. "Curating cyberbullying datasets: A human-AI collaborative approach." International journal of bullying prevention 4.1 (2022): 35-46.
18. Obaid, Mohammed Hussein, Shawkat Kamal Guirguis, and Saleh Mesbah Elkaffas. "Cyberbullying detection and severity determination model." IEEE Access 11 (2023): 97391-97399.
19. Emmery, C., Verhoeven, B., De Pauw, G., Jacobs, G., Van Hee, C., Lefever, E., ... & Daelemans, W. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Language Resources and Evaluation, 55, 597-633.
20. Han, H., Asif, M., Awwad, E.M. et al. Innovative deep learning techniques for monitoring aggressive behavior in social media posts. J Cloud Comp 13, 19 (2024). https://doi.org/10.1186/s13677-023-00577-6
21. Elsafoury, F., Katsigiannis, S., Pervez, Z., & Ramzan, N. (2021). When the timeline meets the pipeline: A survey on automated cyberbullying detection. IEEE access, 9, 103541-103563.
22. Yuan, Lanqin, Tianyu Wang, Gabriela Ferraro, Hanna Suominen, and Marian-Andrei Rizoiu. "Transfer learning for hate speech detection in social media." Journal of Computational Social Science 6, no. 2 (2023): 1081-1101.
23. Raza, Muhammad Owais, Areej Fatemah Meghji, Naeem Ahmed Mahoto, Mana Saleh Al Reshan, Hamad Ali Abosaq, Adel Sulaiman, and Asadullah Shaikh. "Reading Between the Lines: Machine Learning Ensemble and Deep Learning for Implied Threat Detection in Textual Data." International Journal of Computational Intelligence Systems 17, no. 1 (2024): 183
24. Yi, Peiling, and Arkaitz Zubiaga. "Session-based cyberbullying detection in social media: A survey." Online Social Networks and Media 36 (2023): 100250.
25. Yousaf, Kanwal, and Tabassam Nawaz. "A deep learning-based approach for inappropriate content detection and classification of youtube videos." IEEE Access 10 (2022): 16283-16298.
26. Paul, Sayanta, and Sriparna Saha. "CyberBERT: BERT for cyberbullying identification: BERT for cyberbullying identification." Multimedia Systems 28.6 (2022): 1897-1904.
27. Yadav, Jaideep, Devesh Kumar, and Dheeraj Chauhan. "Cyberbullying detection using pre-trained bert model." 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020.
28. Mansur, Zainab, Nazlia Omar, and Sabrina Tiun. "Twitter hate speech detection: A systematic review of methods, taxonomy analysis, challenges, and opportunities." IEEE Access 11 (2023): 16226-16249.
29. Iwendi, Celestine, Gautam Srivastava, Suleman Khan, and Praveen Kumar Reddy Maddikunta. "Cyberbullying detection solutions based on deep learning architectures." Multimedia Systems 29, no. 3 (2023): 1839-1852.
30. Rezvani, Nabi, Amin Beheshti, and Alireza Tabebordbar. "Linking textual and contextual features for intelligent cyberbullying detection in social media." Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia. 2020.

31. Zhou, Yi, Zhenhao Chen, and Huiyuan Yang. "Multimodal learning for hateful memes detection." 2021 IEEE International conference on multimedia & expo workshops (ICMEW). IEEE, 2021.
32. Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. "Advances in machine learning algorithms for hate speech detection in social media: a review." IEEE Access 9 (2021): 88364-88376.