Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

A Survey on Cyberbullying Predictive Model using Deep Learning Techniques

P. Maragathavalli [1], A. Thanushri [2], Seru Neha Lakshmi Gayathri[3], Anjana B K[4], Hima Asok[5]

Department of Information Technology, Puducherry Technological University, Puducherry, India
Email: [1]marapriya@ptuniv.edu.in, [2]thanushri.a@pec.edu, [3]nehalakshmi@pec.edu, [4]anjana.bk@pec.edu, [5]hima.asok@pec.edu

Abstract

Cyberbullying, a pervasive issue in the current digital age, has prompted the need for advanced predictive models to identify and mitigate online harassment. This survey paper explores the landscape of cyberbullying severity level prediction using an ensemble-based deep learning approach for multimodal datasets. Delving into the realm of deep learning techniques and investigating their efficacy in discerning subtle patterns indicative of cyberbullying behaviour, the survey encompasses a comprehensive review of existing ensemble methodologies, highlighting their strengths and weaknesses in cyberbullying prediction. Diverse datasets, model architectures, and evaluation metrics employed in relevant studies are analysed, aiming to provide a thorough understanding of the current technological status. Additionally, difficulties and potential avenues for upcoming studies are discussed, fostering advancements in the development of robust predictive models to combat cyberbullying effectively. Researchers, practitioners, and policymakers looking for insights on the changing field of cyberbullying prevention using ensemble-based deep-learning methodologies will find this survey to be a valuable resource.

Keywords: Cyberbullying, Predictive Model, Online Harassment, Deep Learning Techniques, Ensemble Methodologies, Multimodal Data, Severity Level.

DOI: https://doi.org/10.36548/jtcsst.2024.2.001
Received: 03.03.2024, received in revised form: 08.04.2024, accepted: 22.04.2024, published: 03.05.2024

1. Introduction

This research delves into predicting the severity of cyberbullying using an ensemble-based deep learning approach that analyses diverse data modalities-text, image, audio, and video. By integrating these modalities, the model gains a comprehensive understanding of cyberbullying nuances. This multimodal approach surpasses the limitations of unimodal analyses, providing a nuanced perspective crucial for accurate severity prediction and effective preventive measures is shown in figure 1 which is referred from [13]. Our survey contributes to cyberbullying prevention by advocating an advanced, holistic model that considers the intricate relationships among various data types. Figure 1. Depicts the cyberbullying severity classification task

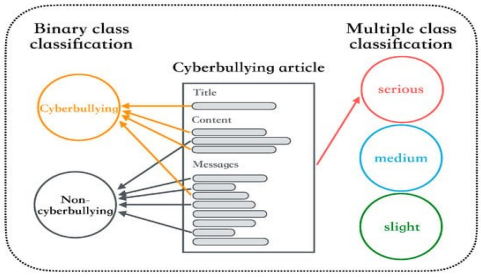

Figure 1. Cyberbullying Severity Classification Task [13]

1.1 Cyberbullying

The deliberate and frequent use of digital technologies to hurt people online is known as cyberbullying. Cyberbullies display a range of traits, including power and control, relational aggressiveness, and cognitive-affective impairments. The consequences of cyberbullying are significant for both the bullies and the victims, making it a critical public health concern [1]. Cyberbullying involves the use of computers, cell phones, and other electronic devices to inflict harm on others through various forms such as flaming, harassment, stalking, impersonation, outing, and exclusion, Table 1 gives a clear idea of cyberbullying types. The use of technology allows perpetrators to humiliate or threaten victims by sending or posting messages or photos to a third party or public forum [2]. Table 1. Illustrates the types of cyberbullying

Table 1. Types of Cyberbullying

Category	Description	Example
Flaming	Sending verbally abusive messages containing insults, threats, and harassment.	Sending a message like "You're the worst! Nobody likes you!"
Harassment	Repeatedly sending offensive messages to intimidate or distress someone.	Sending someone constant messages mocking their appearance or interests.
Denigration	Spreading rumours or gossip to damage someone's reputation or social standing.	Posting false rumours about someone online to make them look bad.
Exclusion	Deliberately excluding someone from online groups or activities.	Removing someone from a group chat or deliberately not inviting them to an online event.
Impersonation	Pretending to be someone else online to embarrass or damage their reputation.	Creating a fake account to spread lies about someone or to trick their friends.
Outing/ Doxing	Sharing private or embarrassing information about someone online without their consent.	Posting someone's phone number or address online without their permission.
Trickery	Deceiving someone into revealing personal information or embarrassing themselves online.	Luring someone into a video call and then recording them in a compromising situation.
Cyberstalking	Repeatedly sending unwanted messages or threats, or monitoring someone's online activity to cause fear or distress.	Sending someone constant messages even after they've asked you to stop.

1.2 Deep Learning Techniques

Deep learning is an artificial intelligence method that enables computers to process data in a way inspired by the human brain. It has been extensively studied and used in a number of

fields, including the categorization of text, images, audio, and video which is tabulated in Table 2. Convolutional Neural Networks (CNNs), one of the deep learning approaches, have been widely employed for image classification because of its high classification accuracy and efficient feature extraction capabilities [3]. Deep learning models including BERT, auto-encoder, sparse coding, constrained Boltzmann machine, and deep belief networks have been used to text categorization [4]. Similarly, for audio and video classification, deep learning models, particularly CNNs, have achieved significant success in automatically classifying videos [2]. Overall, deep-learning techniques have emerged as powerful tools for text, image, audio, and video classification, offering improved accuracy and efficiency in these domains. Table 2. illustrates a comparison of deep learning techniques for cyberbullying detection.

Table 2. A Comparison of Deep Learning Techniques for Cyberbullying Detection

Technique	**Description**	**Suitability for Cyberbullying Detection**	**Considerations**
Convolutional Neural Networks (CNNs)	Efficiently learn features from text, especially for short messages.	Well-suited	Captures local patterns but might miss long-range context.
Long Short-Term Memory Networks (LSTMs)	Capture long-range dependencies in text, crucial for understanding context.	Excellent	Particularly effective for capturing long-range context in cyberbullying messages.
Recurrent Neural Networks (RNNs)	Handle sequential data well, suitable for analysing text flow.	Potentially useful	Can be less powerful than LSTMs for cyberbullying detection.
Generative Adversarial Networks (GANs)	Used for data generation tasks.	Needs further exploration	Potential application in generating synthetic data, but direct use in detection unclears.

Radial Basis Function Networks (RBFNs)	Primarily used for function approximation problems.	Not directly applicable	Not designed for text analysis tasks.
Multilayer Perceptrons (MLPs)	Basic building blocks of deep learning.	Potentially useful with additional techniques	Can be used in conjunction with word embeddings for cyberbullying analysis.
Self-Organizing Maps (SOMs)	Used for data visualization and dimensionality reduction.	Preprocessing technique	Useful for data exploration before feeding it into other models.
Deep Belief Networks (DBNs) and Restricted Boltzmann Machines (RBMs)	Building blocks for deeper architectures.	Indirect application	Used as components of complex models for cyberbullying detection.
Autoencoders	Learn compressed representations of data.	Preprocessing technique	Useful for extracting relevant features from text data before feeding it into other models.

1.3 Ensemble Methodologies

Ensemble is a process in which various learning algorithms are combined to enhance their overall performance. This technique, used in regression, classification, and reinforcement learning, results in a reliable model with improved predictive performance. Ensemble learning improves accuracy by aggregating the predictions of weak learners, resulting in strong learners that make accurate predictions is shown in figure 2 which was referred from [15] and Table 3 highlights the ensemble techniques. Ensemble techniques may be broadly classified into two categories: generative methods, which actively improve the variety and accuracy of base learners, and non-generative methods, which merge learning machines. In a number of fields, ensemble approaches have proved effective in increasing classification accuracy. Table 3.

highlights the strengths and weakness of the ensemble techniques with respect to cyberbullying detection

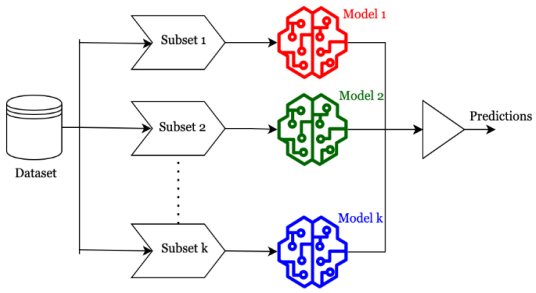

Figure 2. Ensemble Model [15]

Table 3. Highlighting the Strengths and Weakness of the Ensemble Techniques with Respect to Cyberbullying Detection

Technique	Description	Strengths for Cyberbullying Detection	Weaknesses for Cyberbullying Detection
Bagging (Bootstrap Aggregation)	Trains multiple models on random subsets of the data with replacement. Final prediction is based on majority vote.	Improves accuracy by leveraging diverse models. Reduces variance and overfitting.	Can be computationally expensive, especially for complex models. Interpretability can be challenging due to averaging predictions.
Boosting (e.g., AdaBoost, XGBoost)	Sequentially trains models, where each subsequent model focuses on the errors of previous models.	Highly accurate, especially for imbalanced datasets. Improves performance on previously misclassified instances.	More prone to overfitting than bagging if not carefully tuned. Interpretability can be challenging due to the sequential model building process.
Stacking	Trains a meta-model on the predictions of	Captures complementary strengths of base	Interpretability can be even more challenging as it involves a meta-

	multiple base models.	models, potentially leading to higher accuracy. Offers flexibility in choosing diverse base models.	model on top of base models. May require more computational resources compared to simpler ensembles.
Voting Ensembles (Majority Vote, Weighted Vote)	Combines predictions from multiple models with the final prediction decided by the most frequent prediction or a weighted vote considering model confidence.	Simple and easy to implement.	Doesn't necessarily improve over the best individual model's performance. Limited improvement in accuracy compared to more advanced ensembles.

2. Related Work

The paper [5] involves Exploratory Data Analysis (EDA), Machine Learning, Deep Learning, and a Hybrid Ensemble method which outperforms traditional classifiers, emphasizing the significance of native (Bangla) language research in combating cyberbullying. The paper [6] proposes a novel approach to detecting cyberbullying in images using a fine-tuned transformer-based network. The CNBD technique combines image features, Optical Character Recognition (OCR), and image captions, achieving significantly higher accuracy and precision. [7] aims to detect cyberbullying in online text, images, and videos using an Ensemble Deep Learning (EDL) approach. Utilizing datasets from various sources, including Twitter and YouTube, the study employs models like BERT for text and Ensemble CNN for images/videos. [8] employs a systematic methodology involving dataset collection from Twitter, expert labelling, and rule-based automated algorithms for cyberbullying severity detection. Machine learning models, including Random Forest and XGBoost, are used to assess severity. The paper [9] proposes a Deep learning technique with several inputs for detecting cyberbullying using social media data, considering text, images, and videos. The hybrid model achieves promising results in detecting cyberbullying events, showcasing its effectiveness with separate inputs and outputs for text and image features. [10] introduces a framework for detecting and categorizing

the severity of cyberbullying on Twitter. Leveraging features like Embedding, Sentiment, and Lexicon, machine-learning algorithms like Naïve Bayes and Random Forest are applied. [11] examines the efficacy of ensemble multi-classification techniques for identifying different forms of cyberbullying tweets with an emphasis on Twitter. XGBoost, Random Forest, and Decision Tree models are shown; ensemble methods (voting and stacking) beat classical classifiers and demonstrating increased accuracy. [12] enhances cyberbullying detection on social platforms using advanced deep learning models and multi-task learning mechanisms. Leveraging BERT pre-trained models and attention mechanisms, the framework outperforms SVM approaches, showcasing significant advancements in accuracy and emphasizing the importance of context-based word importance determination. Table 4 illustrates overview of cyberbullying detection approaches.

Table 4. Overview of Cyberbullying Detection Approaches

Authors	Dataset Used	Input Modalities	Key Techniques	Results
Md. Tofael Azhmed, et al.,	Bangla dataset (10,512 data points)	Text	Hybrid Ensemble Method, EDA, TF-IDF, Machine Learning, Deep Learning	Accuracy of 85%
Subbaraju Pericherla,et al.,	Facebook, Instagram and Twitter (19,300 images)	Images	Fine-tuned transformer (BEiT), OCR, Image features, Image captions	Accuracy of 98.23%, precision of 98.05%, and recall score of 98.05%.
Zarapala Sunitha Bai,et al.,	Twitter, YouTube, Facebook	Text, Images, Videos	Ensemble Deep Learning (EDL), BERT, CNN, Oversampling, Under sampling, Synthetic data generation	EDL and DNN achieved an accuracy of 0.987, precision of 0.976, F1-score of 0.975, and recall of 0.971 for the Twitter dataset. Ensemble CNN achieved an accuracy of 0.887, precision of 0.88, F1-score of 0.88, and recall of 0.887 for the Image dataset. Ensemble CNN achieved an accuracy of 0.807, precision of 0.81, F1-score of 0.82, and recall of 0.81 for the Video dataset.
Madhura Vyawahare,et al.,	Twitter dataset	Text	Rule-based automated algorithm, Machine Learning (Random Forest, XGBoost), Dataset relabelling	Accuracy,Precision,Recall,F1-score,AUROC
Dr. Vijayakumar V,et al.,	Social media data	Text, Images, Videos	Multi-input deep learning algorithm, LSTM, CNN, Hybrid model, Real-time testing	Accuracy of 86%
Bandeh Ali Talpur,et al.,	Twitter dataset	Text	Pointwise Mutual Information, Machine Learning (Naïve Bayes, KNN, Decision Tree, Random Forest, SVM)	AUC increased from 0.894 to 0.971.
Mohammad Ilyas,et al.,	Twitter dataset - test set (20%) and a training set (80%)	Text	Ensemble classifiers (Voting, Stacking), Machine Learning (Decision Trees, Random Forest, XGBoost)	Decision Trees exhibited accuracy at 89%, followed by Random Forest at 88%, and XGBoost at 86%. Stacking showed a marginal 1% improvement in accuracy compared to voting.
Guo Xingyi,et al.,	Social media platforms - 159,571 comments	Text	BERT pre-trained model, Attention mechanisms, Deep Learning, Multi-task learning mechanisms	Precision of 0.841,Recall of 0.881,F1-score of 0.861,AUROC of 0.989

3. Comparison of Existing Methods

The results from the presented papers showcase promising advancements in the field of cyberbullying detection across various modalities and datasets. Azhmed et al. leveraged a Hybrid Ensemble Method, incorporating EDA, TF-IDF, and both Machine Learning and Deep Learning approaches, achieving an impressive 85% accuracy on a Bangla dataset. Pericherla et al [8]. delved into multimedia content from Facebook, Instagram, and Twitter, employing a Fine-tuned transformer (BEiT), OCR, and Image features, achieving remarkable results with 98.23% accuracy, 98.05% precision, and a recall score of 98.05%. Bai et al [7]. adopted an Ensemble Deep Learning (EDL) approach alongside BERT, CNN, and data manipulation techniques for text, image, and video datasets, showcasing nuanced accuracy and precision scores for each modality. Vyawahare et al [10]. employed rule-based algorithms and machine learning for Twitter data, providing a multifaceted evaluation with parameters including AUROC, F1-score, recall, accuracy, and precision.

Here's a detailed analysis of some key findings from the presented papers, Effectiveness of Hybrid Methods: Studies by Azhmed et al. [9] and Talpur et al. [5] demonstrate the success of hybrid ensemble methods that combine machine learning, deep learning, and techniques like EDA and TF-IDF. These methods achieved impressive accuracy on Bangla and English datasets, respectively. Fine-tuned Models and Multimodal Fusion: Pericherla et al. [8] achieved remarkable results using a fine-tuned transformer network (BEiT) for image classification in cyberbullying detection. This highlights the potential of utilizing pre-trained models with further adjustments for specific tasks. Additionally, Bai et al. [7] showcase the effectiveness of ensemble deep learning (EDL) that combines different deep learning models for text, image, and video data, achieving nuanced accuracy for each modality. Machine Learning for Severity Detection: Vyawahare et al. [10] employed rule-based algorithms and machine learning models like Random Forest and XGBoost for cyberbullying severity detection on Twitter data. This approach provides a comprehensive evaluation with various performance metrics. These studies collectively highlight the potency of the following approaches in extracting meaningful insights from the intricate landscape of social media content: Hybrid methodologies that combine various learning techniques, fine-tuned deep learning models leveraging pre-trained architectures, Ensemble techniques for combining predictions from multiple models, Multimodal fusion that incorporates text, image, audio, and video data for a richer

understanding. (Figure 3) illustrates a comparison of basic deep learning techniques with ensemble methods and advanced deep learning techniques, highlighting the potential for improved accuracy with ensemble approaches.

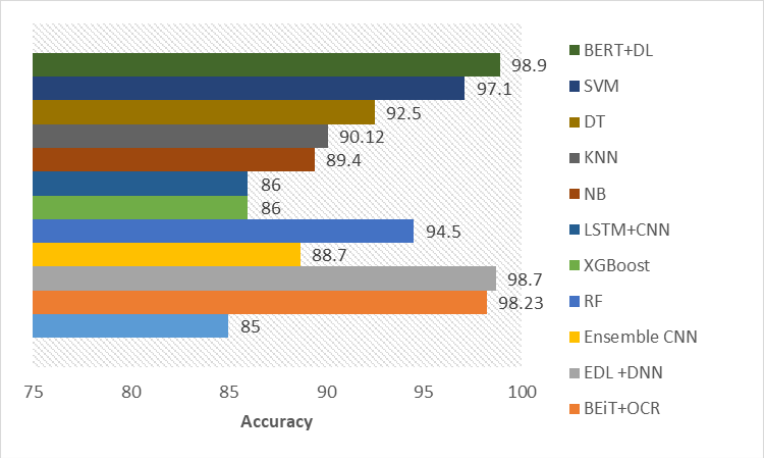

Figure 3. Comparison of the Basic Deep Learning Techniques, Machine Learning and Ensemble Techniques

4. Conclusion

The conclusion, this study has emphasized the significance of ensemble-based deep learning with multimodal data in predicting the severity of cyberbullying incidents. The integration of diverse modalities has not only improved the accuracy of severity predictions but has also provided a deeper understanding of the complex dynamics and impact of cyberbullying. Despite the progress made, there are still challenges that need to be addressed, such as the need for improved interpretability, ethical considerations in handling diverse data sources, and exploring innovative ensemble strategies for multimodal fusion. Furthermore, recommendations for real-time monitoring, post-detection actions, collaboration with online platforms, and investigating contextual influences highlight the importance of a comprehensive and evolving approach to cyberbullying prevention. By focusing on these areas, researchers can contribute significantly to the ongoing development of multimodal cyberbullying prevention strategies in the constantly changing digital landscape.

5. Future Enhancement

Future research in this area can delve deeper by exploring novel deep learning architectures specifically crafted to handle the complexities of multimodal cyberbullying data. Additionally, the inclusion of further modalities, such as network traffic data or user behavioral patterns, holds promise for a more comprehensive understanding of cyberbullying incidents. However, alongside these advancements, it is crucial to address the interpretability of ensemble models, allowing us to dissect their decision-making processes. Furthermore, techniques for handling imbalanced datasets, where cyberbullying instances are rare, need to be developed. Finally, ethical considerations surrounding potential biases within the data or the models themselves must be explored to ensure responsible implementation of deep learning in cyberbullying detection.

References

[1] Sevgi Mestci Sunerli, Füsun Gökkaya, Zahide Aliusta Denk. “Cyberbullying.” Advances in social networking and online communities book series, (2022), doi: 10.4018/978-1-6684-5426-8.ch026

[2] Belhaouari, Samir Brahim, Md Alamgir Kabir, and Adnan Khan. "On the Use of Deep Learning for Video Classification." Applied Sciences (2076-3417) 13, no. 3 (2023).

[3] Yadav, Stuti, and Manish D. Sawale. "A review on image classification using deep learning." World Journal of Advanced Research and Reviews 17, no. 1 (2023): 480-482.

[4] Sarker, Iqbal H. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions." SN Computer Science 2, no. 6 (2021): 420.

[5] Talpur, Bandeh Ali, and Declan O’Sullivan. "Cyberbullying severity detection: A machine learning approach." PloS one 15, no. 10 (2020): e0240924.

[6] Dr. Vijayakumar V, Dr. Hari Prasad D, Adolf P. “Multi-Input Deep Learning Algorithm for Cyberbullying Detection.” International Journal for Research in Engineering Application & Management, (2021), ISSN: 2454-9150 Vol-07, Issue-05, DOI:10.35291/2454-9150.2021.0451

[7] Zarapala Sunitha Bai, Sreelatha Malempati. “Ensemble Deep Learning (EDL) for Cyber-bullying on Social Media.” International Journal of Advanced Computer Science and Applications, (2023), Vol. 14, No.7.

[8] Pericherla, Subbaraju, and E. Ilavarasan. "Overcoming the Challenge of Cyberbullying Detection in Images: A Deep Learning Approach with Image Captioning and OCR Integration." International Journal of Computing and Digital Systems 15, no. 1 (2024): 393-401.

[9] Md. Tofael Ahmed, Afroza Sharmin Urmi,et al.“Cyberbullying Detection Based on Hybrid Ensemble Method using Deep Learning Technique in Bangla Dataset.” International Journal for Research in Engineering Application & Management, (2022), Vol. 14, No. 9.

[10] Vyawahare, Madhura, and Sharvari Govilkar. "Identifying Severity of Cyberbullying Using Scalable Labeled Multi-Platform Dataset." International Journal of Intelligent Systems and Applications in Engineering 10, no. 4 (2022): 201-210.

[11] Abdulkarim Faraj Alqahtani1 and Mohammad Ilyas. “An Ensemble Based Multi-Classification Machine Learning Classifiers Approach to Detect Multiple Classes of Cyberbullying.” Mach. Learn. Knowl. Extr. (2024), 6, 156 – 170. https://doi.org/10.3390/make6010009.

[12] Guo Xingyia and Hamedi Mohd Adnana. “Potential cyberbullying detection in social media platforms based on a multi-task learning framework.” International Journal of Data and Network Science 8 ,(2024) 25–34, doi: 10.5267/j.ijdns.2023.10.021.

[13] Wu, Jheng-Long, and Chiao-Yu Tang. “Classifying the Severity of Cyberbullying Incidents by Using a Hierarchical Squashing-Attention Network” Applied Sciences 12, (2022), no. 7: 3502. https://doi.org/10.3390/app12073502.

[14] Hasan, Md. Tarek, Md. Al Emran Hossain, Md. Saddam Hossain Mukta, Arifa Akter, Mohiuddin Ahmed, and Salekul Islam. “A Review on Deep-Learning-Based Cyberbullying Detection”, Future Internet 15, (2023), no. 5: 179. https://doi.org/10.3390/fi15050179.

[15] https://intuitivetutorial.com/2023/05/12/ensemble-models-in-machine-learning/

[16] Chung, Jetli & Teo, Jason. Single classifier vs. ensemble machine learning approaches for mental health prediction. Brain Informatics,(2023),10.10.1186/s40708-022-00180-6.

Author's biography

Dr. P. Maragathavalli, she received her B.E. degree in CSE from Bharathidasan University, MTech. and Ph.D. degree in CSE from Pondicherry University. She joined Pondicherry Engineering College in 2006 and currently working as Professor in the Department of Information Technology. She has published several research papers in various refereed journals and international conferences. Her area of interest includes Security Testing, Optimization Techniques, Machine Learning and Deep Learning Techniques, Blockchain Technologies and Cyber Security.

Thanushri A, B.Tech (IT) student in Puducherry Technological University, currently immersed in the study of deep learning. Proficient in C, C++, Java, and Python, with a keen interest in applying these skills to advance the field of artificial intelligence.

Seru Neha Lakshmi Gayathri, B.Tech (IT) student in Puducherry Technological University, currently immersed in the study of deep learning and machine learning techniques.

Hima Asok, B.Tech (IT) student in Puducherry Technological University, currently immersed in the study of deep learning, machine learning techniques and technologies contributing to the growth of artificial intelligence.

Anjana B K, B.Tech (IT) student in Puducherry Technological University, currently immersed in the study of deep learning and machine learning techniques. Proficient in C, C++, and Python, with a keen interest in applying these skills to advance the field of artificial intelligence.