Breaking The Silence: An Overview of Approaches To Detecting Cyberbullying on Social Media

Harish D[1], Jayashakthi Vishnu P[2], Manimaran M[3]
[1, 2, 3] Dept of Information Technology
[1, 2, 3] Kumaraguru College of Technology

Abstract- *Social media bullying is a rising issue, and it can have serious and lasting effects. It's crucial to identify social media cyberbullying to stop harm and advance a secure and encouraging online community. In this paper, we give an overall information of the different approaches and techniques that are currently being used to detect online bullying. We discuss the benefits and drawbacks of the current detection techniques and point out potential directions for further study. The paper covers a variety of detection techniques, including hybrid, machine learning(ML), natural language processing(NLP), social-network analysis, and keyword-based approaches. We also take into account several contextual variables, such as cultural and societal variations, privacy issues, and ethical considerations, which can have an impact on the detection of cyberbullying. Overall, this evaluation offers a thorough overview of the present status of social media cyberbullying detection and suggests important areas for further study and improvement.*

Keywords- Cyber Bullying, Machine learning, Deep learning, NLP, Social Media, Review, survey.

I. INTRODUCTION

On social media, cyberbullying is an issue that is on the rise and may have a big, lasting impact. The psychological discomfort, anxiety, despair, and other mental health issues that cyberbullying victims may face can have a severe impact on their personal and academic lives. Cyberbullying can potentially result in self-harm or suicide in extreme circumstances. Thus, it is crucial to identify cyberbullying on social media in order to stop harm and advance a secure and encouraging online community.

This review's objective is to present a summary of the many approaches and procedures that are currently being utilized to identify online bullying on social media. The review attempts to highlight the benefits and drawbacks of the current detection techniques as well as to point out potential directions for further study. A variety of detection techniques, including keyword-based, machine learning-based, NLP-based, social network analysis-based, and hybrid approaches, are included in the review's scope. The review also takes into account several contextual variables, such as cultural and societal differences, privacy issues, and ethical considerations, which might affect the identification of cyberbullying. The review's ultimate goal is to provide a thorough overview of the state of cyberbullying detection on social media and to highlight important topics for further research and development.

II. LITERATURE REVIEW

Shah et.al [1] (2020), in their proposed solution, pre-processing entails two steps: gathering data from Twitter API and Kaggle dataset, labelled by values 0 and 1, and cleaning the data in CSV format. English has many verb tenses, nouns, and other grammatical distinctions. When a word is lemmatized, it is reduced to its root form using the Wordnet Lemmatizer module. Data was divided into two sections, and TFIDF was calculated to turn tweets into vectors in order to select the best classifier. The frequency of a word in a document is known as its term frequency, and the number of documents divided by the number of times a word and its logarithm appear in a document is known as its inverse document frequency. The relevance of a word increases with its IDF score.

SVC, Logistic Regression(LR), Multinomial Naïve Bayes, Random Forest(RF) Classifier, and SGD Classifier comprised the various models utilized in the study. Three criteria—F1-score,precision, andrecall, were considered in the selection of the best classifier, and the model was validated and effectively detected bully tweets. The logistic regression classifier, that had 91% precision, 96% recall, 93% accuracy, 90% specificity, 87% MCC, 9% fall out, and 3% miss rate, is the most accurate classifier.

In the solution put by Paul et.al [2] (2022), frameworks based on multimodal deep learning(DL) have been offered to accurately detect cyber-bullying, with highest results coming from feature-based fusion. Using a multimodal technique based on deep learning to identify cyberbullying in the Vine dataset.A Residual-BiLSTM architecture was presented to early detect bullying based on textual modality, and a Universal Sentence Encoder was employed to provide

good text embedding with rich semantic information. High-level and low-level features are combined in residual-based architecture to enhance network performance and prevent over-fitting. Early bullying detection based on visual modality utilizing Recurrent-CNN (RCNN) and information fusion using Residual-BiLSTM and Recurrent Convolution Neural Network (RCNN). For (Feature-level) fusion and (Decision-level) fusion, the proposed framework outperforms other deep learning-based frameworks.Due to the fact that uni-modal classifying models did not generate any likely output, (Feature)-fusion performed better than (Decision)-fusion, which is why (Feature)-fusion outperformed (Decision)-fusion.

Kumar et.al [3] (2022) ,in their research, to determine whether a person is a bully in cyberspace or not, cyberbullying detection entails data gathering, preprocessing, juicing out and pickingimportant attributes, and constructing a cyber-bullying classifying algorithm. Tokenization, case transformation, and data cleansing are all parts of data preparation. Data preprocessing for ML models involves several crucial stages, including feature extraction, padding, outlier removal, stemming/lemmatization, label encoding, and outlier removal. Because of its high time complexity, the BOW model is ineffective for large datasets, but TF-IDF and Word2Vec enhance classifier performance. Whereas N-gram is an expanded version of BOW to capture semantic links, Word2vec employs neural-networks to identify semantic relationships among words.Feature selection methods are used to pick importantattributes without lessening the prediction level of models, such as Chi-Square, Information Gain(IG), and Pearson's Correlation(PC). These methods filter out less feasible attributes to enhance classification accuracy.

By the use of supervised or unsupervised learning methods, ML is utilized to identify and predict cyberbullying. Reliable models are built using Naive Bayes (NB), SVM, LR, Decision Trees (DT), and Random Forest (RF). By calculating the distance or similarity distance amongthe train and test data, KNN is a nonparametric and instance-based classification paradigm that is used to categorize unknown instances. Cyberbullying is detected by deep neural networks employing CNNs, RNNs, LSTMs, and BLSTMs. RNNs recognize data sequence patterns, CNNs learn from the patterns in the data, and LSTMs and BLSTMs are employed to address gradient-vanishing and short-term memory issues. A bidirectional network like BLSTM can store data from the past as well as the future.

In the research work of Alotaibi et.al [4] (2021), they proposed multichannel deep learning framework combines the power of three advanced deep learning(DL) models: the transformer block, Bidirectional GRU(BiGRU), and CNN-architecture. It can juice out significant attributes and provide reliable outcomes. The transformer outperforms NLP deep models with a spatial dropout layer, bidirectional RNN layer, global-average pooling layer, and global-maximum pooling layer by using the attention method without RNN. The hidden state is calculated in both ways by each GRU cell. In order to handle the input jointly, this research offers a multichannel deep learning model that makes use of three networks: CNN, BiRNN, and a transformer block. The network is optimized by using the Adam optimizer and a binary cross-entropy loss function.The proposed method produced 0.87 precision, 0.85 recall, 0.86 F-score when identifying non-toxic and 0.89 F1-score when identifying toxic data, and 0.88 accuracy when train and test data is separated into ratio 3:1.

B.A.H. Murshed et.al [5] (2022), in the study proposed the DEA-RNN model, which involves attribute extraction and attribute picking, pre-processing and data cleaning, and classification. 32 cyberbullying keywords and 10,000 randomly chosen tweets are included in the data collection. Over the course of 1.5 months, three human annotators annotate the data. Due to the imbalance issue between bullying and non-bullying tweets, the suggested model makes use of Synthetic Minority Oversampling Technique(SMOTE) to oversample the minority class (bullying Tweets).A hybrid DL method known as DEA-RNN is used to detect cyberbullying on the Twitter platform. The exploration process is improved by changing the Accumulative-Fitness AF(A+k)j to match features with fewer iterations. RNN training time can be cut down, and the sluggish pace of convergence limitation can be overcome, using parameter optimization.The outcomes of cyberbullying are validated based on various input dataset percentage 6:4, 7:3, and 9:1, in 3 scenarios respectively.

In the solution put up by M.M. Islam et.al [6] (2020), their study gave a solution to automate the identification of posts on social media belonging to bullying by considering 2 attributes: BoW and TF-IDF. Four machine learning algorithms were used to identify bullying text. The framework for detecting cyberbullying is divided into two sections: NLP and ML. While ML uses machine learning techniques to identify bullying communications, NLP requires preparing and cleaning datasets. Based on the Bayes theorem, Naive Bayes is a powerful ML algorithm that makes predictions based on the likelihood of an item. A supervised learning model called Random Forest(RF) uses the predictions from every single created tree and bases its decisions on the maximum number of votes. In n-dimensional space, the Support Vector Machine(SVM), a supervised ML technique, has the capacity to discriminate classes in a single way.

In the research of Neelakandan et.al [7] (2022), this study presents a novel attribute subset picking with DL-based bullying detection and segregation(FSSDL-CBDC) model on online forums. For improved classification performance, a technique called binary coyote optimization-based attribute subset picking(BCO-FSS) is used to pick a collection of attributes. To find and categorize Cyber Bullying in online forums, one uses the salp-swarm algorithm(SSA) with deep-belief network(DBN). In order to address the problems that deep layered networks face with training, such as slow-learning, becoming stuck in local minima due to poor attributepicking, and requiring a large amount of training data, deep belief networks (DBNs) were developed.The (SSA) method is used to identify and categorize bullying in online forums, while the (BCO)approach is used to select a set of attributes for enhancedsegregating efficiency. A high number of simulations were run on a (benchmark) dataset for the proposed FSSDL-CBDC method's exploratory better detecting presentation.

Priyadharshini et.al [8] (2023), in their research, Natural Language Processing requires data pre-processing to filter out irrelevant material and transform it into a format suitable for further computing. To convert data into numerical values, GloVe embeddings and word vectorization are utilized. Pre-trained language models Word2Vec and GloVe are used to build embeddings for better word representation. This study introduces three innovative feature subset selection methods: a binary coyote optimization-based feature subset selection(BCO-FSS) methodology to pick a set of attributes for increased segregating efficiency; and a salp-swarm algorithm(SSA) with deep-belief network(DBN) to identify and segregate cyberbullying in social networks.

The research work of T. Ahmed et.al [9] (2021), provides an innovative strategy for identifying cyberbullying on social media sites that relies on transformer-based structures and an attention mechanism. The average F1-score for the suggested architecture on the Fine-Grained Cyberbullying Dataset(FGCD) was 95.59% for five classes, 90.65% for six classes, and 87.3% on the Twitter-parsed dataset, outperforming the provided ML and DNN baselines by a huge margin in alternate cases. These results provide insight on how well transformer-based models detect bullying and pave the door for further studies to address this critical online problem.

The research work of Desai et.al [10] (2021), revealed a strategy to identify hate speech before it is classified as cyberbullying on social media by taking into account the sentence's syntactic, semantic, and ironic characteristics. In order to comprehend the meaning of both contexts, BERT is a bi-directional model that has been pre-trained on un-marked textual data from left and right directions. Three embeddings are used by the bidirectional BERT model to comprehend a word's meaning in both the left and right contexts.The input processing and prediction results from the testing, the classification report rooted on the test data, and the confusion matrix rooted on the test data are the most crucial details in this article. When used on the same dataset, the accuracy of the (SVM) and Naive Bayes(NB) models was 0.7125 and 0.5270, respectively. When utilized with the Twitter dataset for sentiment analysis, the BERT model performed better than typical machine learning models on comparable datasets, with accuracy of 0.9190.

In a solution put up by Prasad et.al [11] (2019), investigates an IR framework called the Semantic-Enhanced Marginalized Denoising Auto-Encoder(SEMdae) stems the expressions of a content as part of the sorting process. It is crucial to maintain each word's double-route junction in the (IR)-architecture as well as its stemmed form, s(W). The W, s(W) connection, which can infer s(W), the stemmed structure, is a better technique. Stopwords in a language can be identified and eliminated using the stemming calculation. A content corpus' vocabulary can be organized by recurrence, and terms that should be eliminated can be selected by going down the rundown.Moreover, stopwords may be eliminated either prior or post stemming findings is joined. The four structures are "am/is/are/be/being/been," "have/has/having/had," "do/does/doing/did," and "have/has/having/had."This paper mentions the content-based(CB) bullying location issue by creating a semantic-upgraded underestimated denoising autoencoder and word embeddings to refine harassing word records. It has been tested viadouble cyber-bullying corpora from online forums and is being further improved by taking into account the word request in DMs.

Nirmal et.al [12] (2021), in their study used various ML models, in which, the Naive Bayes Model is a conditional probabilistic classifier that employs naïve independent presumptions between the various features while using the Bayes theorem. A supervised learning approach called the SVM Model leverages complex statistical learning theory to get around the dimensionality problem. The DNN Model is a supervised learning approach that finds the best hyperplane by using kernel functions to maximize the margin of training data. Using data sets obtained from Myspace, Kongregate, and Slashdot datasets, a linear kernel was applied. A linear kernel applied to Lib SVM with tenfold cross validation and a false positive rate of 28 in 294 occurrences and false negative rate of 12 in 10184 instances were obtained.The training was enhanced by using the weighted TF IDF model. While

employing bigrams instead of unigrams, SVM with a linear kernel had an accuracy of 83.3%. Deep neural networks have multiple layers that work together to execute calculations, and many times the hidden-layer activation functions h(l)(x) are the same at every level.

R. Jayadurga et.al [13] (2023), in their research, used Deep Learning to classify cyberbullying. To improve text detection, pre-processing entails cleaning up text and deleting spammy information. The approaches for extracting features from texts are TF-IDF and Word2Vec. Word2Vec vectorizes text by creating attribute vectors for each word using a neural network with two layers. It searches for words using the CBOW model and searches for texts that might be close to each word using the skip-gram model. BOW and TF-IDF algorithms are two examples of text categorization techniques, and TF-IDF features can be used to identify cyber-bullying in online forums.By enabling the identification of relationships between features in the past and present through the use of social media, DL offers special techniques for analyzing user behaviour. Understanding the state of a society and its constituent parts is crucial. SVM performed poorly in terms of accuracy and precision, indicating it may not be the ideal option for detecting cyberbullying. In contrast, LR, SGD, LGBM, and SVM classifiers all attained comparable accuracy and precision. Although SVM accuracy was poor due to the large dataset and F-measure analysis, SGD virtually outperformed LR despite the error not being reduced to the same degree.

J. Yadav et.al [14] (2020), proposed a method for detecting cyberbullying, based on a deep neural network(DNN) called Transformer, a revolutionary neural-network architecture rooted on a self-attention mechanism. A recently created BERT pre-trained model is employed, with the job and dataset being fine-tuned. When given text as input, BERT creates contextualized embeddings for that text. For classification, a single linear neural-network layer is applied above BERT. A BERT model's primary function is to produce word and sentence embeddings for model input. The final embeddings are produced using 12 layers of transformer in the BERT base model. Since 2018, BERT, a pre-trained model that creates contextualized embeddings for classifier input, has been employed in Google search engines. To create the final embeddings, 12 layers of transformer encoders are used. The BERT model is a bully detection paradigm that uses two unique tokens to begin and end phrases ([CLS] and [SEP]), pads shorter sentences with [PAD] tokens, and explicitly distinguishes between sentence tokens and [PAD] tokens with the "attention-mask." BERT is a lowercase English language pre-trained model that uses 12-layers of transformer encoders to encode the linguistic information. The proposed model has improved accuracy over the existing models, with the Formspring dataset having 78.5% accuracy and the Wikipedia dataset having 96% accuracy.

P. Ratadiya et.al [15] (2019), in their study used, pre-processing, text encoding, model training and prediction to put a textual sample data into single or multiple categories of profanity, preserving features of the entire sequence. By converting text into model-friendly data using pre-trained word embeddings, sequence padding improves performance. To arrive at the final forecast, four models are combined into an ensemble. The model architecture employs a multi-headed self-attention method to reduce complexity and memory needs, but because of the lack of recurrence, there is no information on the sequence's order. Although multi-head attention functions associate queries and key-value pairs with output, positional encodings reveal the absolute and relative positions of tokens. The outputs of numerous attention functions are computed for h distinct projections of the queries, keys, and values using scaled dot product attention. To decrease the number of parameters and enhance binary cross-entropy loss, average pooling and dropout are utilized.The proposed approach outperformed the Bi-GRU based model on the majority of classes across all metrics. Attention ensemble model provides better results than conventional methods due to class imbalance and complexity analysis of various layers.

Aurpa et.al [16] (2022), proposed a Framework, in which, before training their model, they pre-processed the dataset's comment texts. With the use of transformer-based learning and pre-processed comments, they train their model so that it can recognize offensive comments. Finally, they adjusted the model using various hyperparameter values. This process improves the model's ability to predict classes.The pre-trained models employed in the study are BERT and ELECTRA; both include transformer-based architecture, a number of layers, hidden sizes, and parameters.The outcomes of our tests demonstrate that, given a particular learning-rate of (2e-04), both the BERT and ELECTRA models perform better, with the Troll-class having maximum true positive rate. Both model's accuracy and loss remain constant over time. For the learning-rate of (2e-04), the BERT-(Base) and ELECTRA-(Base) models fared better than others, while the ELECTRA-large model also did well.

In a solution put up byS. M. Kargutkar et.al [17] (2020), input, hidden, and output layer are the three layers that make up the deep learning class of neural networks. Data cleaning, or pre-processing, removes extraneous words and special characters from the data. Sequential layer, perceptron, and neural network are layers used in CNN models. The

perceptron is a solo algorithm that receives an input vector (x) of (m) values and returns either 1(yes) or 0(no) depending on whether wx+b0 holds true. If not, f(x)=0 is returned. The activation functions employed in this work to generate smooth values using nonlinear functions are the most crucial information.These include sigmoid and ReLu, which are defined as [1/ (1+e x)] and can be utilized to compute the non-linear function [z=wx+b]. Using NLTK, the incoming text is transformed into a list of word indices. The fit() function is then used to compile and train the model. Epochs, batch size, validation data, and the error brought on by the comparison step are the variables that are used. The method is then put to the test until it reaches a local or global minima.

In the research work of Ahmet et.al [18] (2020), current state of deep learning(DL)-rooted sentiment analysis algorithms for document-level, sentence-level, and aspect-based sentiment analysis for short and long text is surveyed and examined. There is a thorough explanation of DL-architectures for sentiment analysis. The investigated methods are divided into three categories: cross-domain, coarse-grain(including document and sentence level), and fine-grain(including target and aspect level). Finally, for each of the aforementioned categories, we present a summary and in-depth analysis of the surveyed studies. Long short-term memory(LSTM),gated recurrent units(GRU), convolutional neural-networks(CNNs), and attention mechanisms were all extensively studied. LSTM and CNN-rooted models battle on performance for coarse-grained sentiment analysis, although CNN provides less model-complexity and training time.A model must understand the intricate relationships between target/aspect words and opinion terms in order to do fine-grained sentiment analysis. Although CNN-rooted models have been successful at juicing out aspect, bi-directional LSTM and attention processes show the most potential. LSTM and attention models dominate efforts in cross-domain sentiment analysis. Proposed research into cross-domain strategies uncovered the utilization of joint-training, adversarial-training, and multitask-learning for domain adoption.

K. S. Alam et.al [19] (2021), in their research,in order to evaluate the three-level voting system, they used a publicly available dataset. They then used ML algorithms and ensemble approaches to identify cyberbullying in tweets. Dataset splitting was done to divide the dataset into a train and test set. Data pre-processing techniques were employed to get the data in a clean format. In order to validate the model, minimize data size using feature extraction techniques, and identify cyberbullying on social media, machine learning algorithms were applied. The voting classifiers in the DLE architecture are divided into two levels, V C1 and V C2, which compare their results to determine the ultimate classification outcome.While using TF-IDF to extract features, they were able to get 94% accuracy for SLE, 70% accuracy for retrieving features from tweets, and satisfactory performance for other features. When Stratified K-Fold and Stratified Shuffle Split are employed, DLE model has obtained 75% accuracy for TF-IDF ('Bigram') and 83% accuracy for TF-IDF ('Bi-gram'). For accuracy, cross-validation methods have been used. With high accuracy and cross-validation methods, machine learning-based SLE and DLE models can be used to identify social bullying, making them appropriate for fact-checking websites.

Iwendi et.al [20] (2020),in order to identify insults on social media networks, the Kaggle dataset was employed. Text size was decreased, and extraneous information was removed using pre-processing and tokenization. The highest standards were used to safeguard and test tokenization. Tokens are converted into standard formats using stemming, lemmatization, stopwords, and LSTM networks. Inflectional forms are minimized, common terms are extracted, and LSTM RNNs are employed for text categorization. The proposed LSTM uses forget, output, and double input gates to increase accuracy but increases computing complexity and cost.The BLSTM model uses two LSTMs to gather knowledge about the inputs around it. A forget gate is used to remove cell-state data, and an input gate is used to add new cell-state data. When utilized with time series data, recurrent neural networks (RNNs) perform better than CNNs at solving the vanishing gradient problem. RNN is employed to predict serial data, and GRU is made to address issues with short-term memory. It has double gates: an update and a reset gate.BLSTM fared superiorto other deep learning models in terms of F1-Measure, recall, accuracy, and precision. The accuracy, recall, and F1-Measure of the LSTM model for the insult class were 68%, 80%, and 63%, respectively. While GRU AUC is 3% higher, BLSTM AUC is 5% higher than LSTM and 3% higher than RNN. In terms of precision, recall, F1-Measure, and AUC, BLSTM and RNN models performed better than others.

III. CHALLEGES AND LIMITATIONS OF EXISTING METHODS

While assessing the efficacy of existing approaches for identifying cyberbullying on social media, it's crucial to take into account the obstacles and limits these methods face. Among these difficulties and restrictions are:

1. Limited accuracy: Despite improvements in machine learning and NLP methods, it can still be difficult to effectively identify cyberbullying on social media. This is partly because of the nuanced nature of social

media interactions, the intricacy of language, and the variety of forms that cyberbullying can take.

2. Lack of standardization: There is no agreed-upon definition of cyberbullying, and various detection techniques might use various standards to identify and classify cyberbullying. It may be challenging to compare and assess the efficacy of various approaches due to this lack of consistency.
3. Bias and discrimination: Detection techniques may be biassed or discriminating, especially if they rely on terms or phrases that may be connected to specific communities or groups. This may cause misleading positives or negatives, propagate discrimination, or stereotypes.
4. Contextual considerations: Social media interactions' settings have a big impact on whether or not they qualify as cyberbullying. An innocent jest among friends, for instance, might be interpreted as harmful or menacing in another situation. Current approaches could have trouble taking these contextual elements into account.
5. Privacy issues: To identify cyberbullying, detection techniques frequently need access to user data and social media messages. This can cause privacy issues, especially if the techniques are not visible or if they are applied without the subject's consent.

Overall, these issues and constraints indicate that there is still a lot of work to be done in creating reliable and efficient techniques for identifying cyberbullying on social media. These problems will need to be addressed in the next research in order to create accurate, dependable, and morally correct procedures.

IV. CONCLUSION

In this review study, we emphasized the significance of identifying cyberbullying on social media and gave a summary of the available approaches and tools for doing so. We talked about the many methods that are now in use, including hybrid, machine learning, NLP, social network analysis, and keyword-based methods. Also, we emphasized the drawbacks and shortcomings of the current approaches, such as their poor accuracy, lack of standardization, bias and discrimination, sensitivity to context, and privacy problems. We concluded by highlighting the significance of ongoing research in this field to solve these issues and provide better, more dependable approaches for identifying cyberbullying on social media.

V. FINAL THOUGHTS AND RECOMMENDATIONS FOR FUTURE WORK

There is a need for more research in this field given the difficulties and limits of current approaches. The goal of future study should be to improve the detection of cyberbullying on social media while simultaneously addressing issues with context, privacy, bias, and discrimination. Future studies could specifically concentrate on:

1. Creating more complicated machine learning and NLP-based algorithms to more accurately detect cyberbullying in intricate social media exchanges.
2. Combining different detection methods in order to increase precision and dependability while simultaneously tackling bias and context issues.
3. Creating uniform definitions and standards for cyberbullying that can be used to various detection techniques and platforms.
4. Investigating joint strategies between social media sites, researchers, and users to enhance detection techniques and handle privacy issues.

Therefore, further research is essential to establishing measures that will stop cyberbullying on social media and make the internet a safer and more welcoming place for everyone.

REFERENCES

[1] Shah, R., Aparajit, S., Chopdekar, R. and Patil, R., 2020. Machine Learning based Approach for Detection of Cyberbullying Tweets. *Int. J. Comput. Appl*, *175*(37), pp.51-56.

[2] Paul, S., Saha, S. &Hasanuzzaman, M. Identification of cyberbullying: A deep learning based multimodal approach. *Multimed Tools Appl* **81**, 26989–27008 (2022). https://doi.org/10.1007/s11042-020-09631-w.

[3] Kumar, R., Bhat, A. A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media. *Int. J. Inf. Secur.* **21**, 1409–1431 (2022). https://doi.org/10.1007/s10207-022-00600-y.

[4] Alotaibi, M.; Alotaibi, B.; Razaque, A. A Multichannel Deep Learning Framework for Cyberbullying Detection on Social Media. *Electronics* **2021**, *10*, 2664. https://doi.org/10.3390/electronics10212664.

[5] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif and H. D. E. Al-Ariki, "DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform," in IEEE Access, vol. 10,

pp. 25857-25871, 2022, doi: 10.1109/ACCESS.2022.3153675.

[6] M. M. Islam, M. A. Uddin, L. Islam, A. Akter, S. Sharmin and U. K. Acharjee, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches," 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), Gold Coast, Australia, 2020, pp. 1-6, doi: 10.1109/CSDE50874.2020.9411601.

[7] Neelakandan S, Sridevi M, Saravanan Chandrasekaran, Murugeswari K, Aditya Kumar Singh Pundir, Sridevi R, T.Bheema Lingaiah, "Deep Learning Approaches for Cyberbullying Detection and Classification on Social Media", *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 2163458, 13 pages, 2022. https://doi.org/10.1155/2022/2163458.

[8] Priyadarshini, I., Sahu, S. & Kumar, R. A transfer learning approach for detecting offensive and hate speech on social media platforms. *Multimed Tools Appl* (2023). https://doi.org/10.1007/s11042-023-14481-3.

[9] T. Ahmed, M. Kabir, S. Ivan, H. Mahmud and K. Hasan, "Am I Being Bullied on Social Media? An Ensemble Approach to Categorize Cyberbullying," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 2442-2453, doi: 10.1109/BigData52589.2021.9671594.

[10] Desai, A., Kalaskar, S., Kumbhar, O. and Dhumal, R., 2021. Cyber Bullying Detection on Social Media using Machine Learning. In *ITM Web of Conferences* (Vol. 40, p. 03038). EDP Sciences.

[11] Prasad, K.L., Anusha, P., Rao, M.S. and Rao, K.V., 2019. A Machine Learning-based Preventing the Occurrence of Cyber Bullying Messages on OSN. *International Journal of Recent Technology and Engineering*, *8*(2), pp.1861-1865.

[12] Nirmal, N., Sable, P., Patil, P. and Kuchiwale, S., 2021. Automated detection of cyberbullying using machine learning. *Int. Res. J. Eng. Technol.(IRJET)*, pp.2054-2061.

[13] R. . Jayadurga, T. . Veeramakali, M. . Ali Sohail, N. . Alangudi Balaji, K. . Kumar C., and S. L. P., “Deep Learning Based Detection and Classification of Anomaly Texts in Social Media”, *Int J Intell Syst Appl Eng*, vol. 11, no. 4s, pp. 78–89, Feb. 2023.

[14] J. Yadav, D. Kumar and D. Chauhan, "Cyberbullying Detection using Pre-Trained BERT Model," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 1096-1100, doi: 10.1109/ICESC48915.2020.9155700.

[15] P. Ratadiya and D. Mishra, "An Attention Ensemble Based Approach for Multilabel Profanity Detection," 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China, 2019, pp. 544-550, doi: 10.1109/ICDMW.2019.00083.

[16] Aurpa, T.T., Sadik, R. & Ahmed, M.S. Abusive Bangla comments detection on Facebook using transformer-based deep learning models. *Soc. Netw. Anal. Min.* **12**, 24 (2022). https://doi.org/10.1007/s13278-021-00852-x.

[17] S. M. Kargutkar and V. Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 734-739, doi: 10.1109/ICCMC48092.2020.ICCMC-000137.

[18] Ahmet, A., Abdullah, T. (2020). Recent Trends and Advances in Deep Learning-Based Sentiment Analysis. In: Agarwal, B., Nayak, R., Mittal, N., Patnaik, S. (eds) Deep Learning-Based Approaches for Sentiment Analysis. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-1216-2_2.

[19] K. S. Alam, S. Bhowmik and P. R. K. Prosun, "Cyberbullying Detection: An Ensemble Based Machine Learning Approach," 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 2021, pp. 710-715, doi: 10.1109/ICICV50876.2021.9388499.

[20] Iwendi, C., Srivastava, G., Khan, S. *et al.* Cyberbullying detection solutions based on deep learning architectures. *Multimedia Systems* (2020). https://doi.org/10.1007/s00530-020-00701-5.