2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA) | 979-8-3503-2848-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICT4DA59526.2023.10302244

# Multimodal Cyberbullying Detection Using Deep Learning Techniques: A Review

Immaculate Musyoka
*Department of Computer Science*
*Dedan Kimati University of Technology*
Nyeri, Kenya
immaculatemusyoka87@gmail.com

John Wandeto
*Department of Computer Science*
*Dedan Kimathi University of Technology*
Nyeri, Kenya
john.wandeto@dkut.ac.ke

Benson Kituku
*Department of Computer Science*
*Dedan Kimathi University of Technology*
Nyeri, Kenya
benson.kituku@dkut.ac.ke

***Abstract*—The rise of social networks and online communication, facilitated by internet accessibility and modern technology, has brought numerous benefits. However, it has also given rise to cyberbullying, a harmful phenomenon characterized by disclosing private information and posting hostile content to shame individuals. The repercussions of cyberbullying are severe, impacting victims' mental health, social lives, and personalities. Given the vast daily data uploads on social media, there is a pressing need for automated cyberbullying detection tools. This paper conducts a review of research in cyberbullying detection, encompassing both traditional machine learning and deep learning studies, spanning unimodal and multimodal approaches. The search involved major academic digital libraries like ACM Digital Library, IEEE Xplore Digital Library, and Springer Link, yielding 250 research articles. A selection process and redundancy checks followed, narrowing down the articles to 45 based on specific criteria: publication between 2019 and 2023, a focus on cyberbullying detection and related online risks like hate speech, use of English language data, and the development or introduction of cyberbullying detection algorithms. The significant contributions of the retained articles were identified, alongside future research directions. The paper also provides summaries of the datasets and algorithms employed. It concludes by highlighting ongoing challenges in the field to be addressed in the future.**

***Index Terms*—Cyberbullying; detection; machine learning; multimodal learning; deep learning; CNN; Neural Networks**

## I. Introduction

In today's digital age, social media platforms such as Twitter, Facebook, Instagram, and TikTok, among others, have become a daily way of engagement for many people all over the globe [1]. These platforms have become an essential part of our daily lives [2], and their benefits in terms of enhancing communication via connections, marketing, providing information, entertainment, and education cannot be understated. Certainly, they have had a significant impact on society, but they have also faced criticism and drawbacks that affect specific groups of people. One such drawback on the platforms is cyberbullying [3]. It entails using technology to harass, threaten, shame, or attack another person. Cyberbullying occurs frequently due to the anonymity of the internet, with the targeted unable to denounce the perpetrators. As a consequence, victims live in continual worry that someone may see or publish bullying material online, increasing anxiety,mental disturbance and despair levels [4]. Additionally, cyberbullying actions have negative effects on victims' mental health and, in rare cases, even lead to suicide. Low self-esteem, family problems, academic challenges, criminal activity, and school violence are among other problems associated with cyberbullying [5]. According to a survey conducted by Pew Research Center in 2017, 41% of the Americans reported to have faced cyberbullying with 66% reporting to have witnessed people being bullied [6]. Additionally in the following year, the Center also conducted survey on cyberbullying incidences and found that 59% of teenagers experienced different forms of bullying such as calling of names, spreading of false rumors and even physical threatening [7].The bullying trend continues to thrive every day with the victims being exposed to continual suffering. Therefore, given the adverse effects it brings and its continued spread, it is essential to devise mechanisms to automatically detect and prevent it.

In the recent past, Machine Learning (ML) techniques have shown promising results in their ability to recognize patterns of bullying behavior [8]. Some of the devised techniques utilize convectional supervised learning abilities of Naive Bayes, Support Vector Machines, and Logistic Regression among others [9]. These conventional approaches uses statistical algorithms to identify patterns in the data. At the same time, they require selecting and extracting relevant features of bullying behavior from the data, which in most cases, is always labelled [10]. To detect abusive or bullying information from the data, conventional models makes use of natural language processing (NLP) to analyze the text [11]. Despite their abilities, the approaches have huge limitations in handling large amounts of data and extracting complex features from it,for example context,semantics and object detection. To overcome these challenges, Deep Learning (DL), which is a branch of ML is used. The DL methods allows for automatic feature extraction which makes them able to handle large amounts of data and extract complex features available in images or textual data. In comparison to the Conventional approaches, the DL methods perform better with improved performances and detection capabilities of cyberbullying incidences [12].

In several studies conducted, deep learning based cyberbullying detection has proven to be vital and successful. This can be evidenced in the works of [13] where the authors employed deep learning techniques to detect cyberbullying of the Arabic

dataset. The study used Feed Forward deep Neural Network to detect cyberbullying on data collected from Twitter and manually annotated and as a result, performance accuracy of 97.5% was achieved for the binary classification. Similarly, deep learning based Convolution Neural Network(CNN) was adopted by [14] to detect cyberbullying from tweets collected from the Twitter platform with an accuracy of 95%. Despite the state of the art results produced by the deep learning automation methods, textual based cyberbullying was the most widely studied area, despite both text and images co-existing in the cyberbullying activities across the social media platforms. As a result, this paper aims to conduct a review of different methods and modalities used for the detection of cyberbullying across different platforms. Given the contextual understanding of this research, we will be guided by the following research questions:-

1) What are the different machine learning techniques employed for detection of cyberbullying incidences in unimodal and multimodal context?
2) What are the strengths and limitations of the selected machine learning techniques in their abilities to identify bullying incidences?

To address the research questions, this paper is organized as follows: Section II describes the methodological approach utilized in analyzing review of different works. Section III describes state of the art of cyberbullying and Section IV describes the summary and potential future works from this area.

## II. Methodology

A systematic review was conducted on different ML methods employed in the detection of cyberbullying incidences. A search was conducted on significant academic databases like ACM, IEEE Xplore, Springer, and Google Scholar to obtain variety of research papers that were relevant to the study.These digital libraries provide access to a wide range of reputable, peer-reviewed journals. The search was accomplished by combining a number of different cyberbullying detection-related terms. The papers were vetted using the Mendeley to ensure that proper review was acquired. This required removing studies published outside the years 2019 through 2023 to ensure information gathered is recent, publications lacking proof of algorithmic detection, and papers lacking material or information relevant to cyberbullying were filtered. However, due to limited nature of papers considering multimodal cyberbullying detection, papers with other forms of online harassment that might depict cyberbullying incidences were considered. Out of the pool of 250 research papers, 45 papers were retained after thorough filtering based on the criteria of the years, relevancy and whether algorithms were present or not. The retained papers were examined to learn more about how the authors selected the data they used and why, how the data was annotated, what attributes were employed in the model's creation, which machine learning models were used to detect cyberbullying, and how the model was evaluated.

## III. Literature Survey

Cyberbullying has become a widespread problem in the digital era, seriously harming people all over the world. Cyberbullies discover new ways to harass their victims as internet platforms develop, making it harder and harder to identify and stop such damaging activity. In the recent times, traditional based cyberbullying has dominated the research world. However, the approach is limited to capturing the nuanced nature on the online harassment which in most cases, involves use of both multimodal data such as text and images. To better understand the best detection methods for cyberbullying instances, this review aims to explore the concepts of both unimodal and multimodal cyberbullying detection, their potential benefits, current challenges and future directions to be taken.

### *A. Unimodal based Cyberbullying detection*

The occurrence of cyberbullying incidences in the unimodal context involves use of one modality like text, images or even video or audio clips. Over the years, classical machine learning approaches have been used in the detection of cyberbullying incidences using either textual or imagery data. These classical techniques unfolds in two ways: As supervised techniques and as unsupervised techniques. In supervised learning, all the techniques employed requires use of labelled instances as either cyberbullying or non-cyberbullying which is made possible with the help of human annotators. Once the dataset is identified, relevant features are extracted from the textual data as word frequencies, sentimental analysis or syntactical patterns and later a supervised learning algorithm is trained on the labelled data where the algorithm is expected to learn features and patterns of cyberbullying behavior. These algorithms includes Naive Bayes, Support Vector Machines, Random Forests and Neural Networks. On the other hand, unsupervised techniques requires unlabelled data for training the algorithms and are characterized with abilities to uncover anomalies that are unknown. The algorithms for this technique includes the clustering techniques. Both techniques are important in the detection of cyberbullying incidences and have been experimented by different researchers.

Naive Bayes and K-nearest neighbors(KNN) algorithms were used in the detection of cyberbullying incidences in the study conducted by [15]. The authors used dataset that was collected from Facebook, bearing comments or posts made by the online users. To select important features for training their classifiers, $x^2$ (chi-square test) procedure was used. The processed dataset was then passed through KNN and Naive Bayes classifiers achieving an accuracy of 73% and 72% respectively. Similarly, [16] used Support Vector machine (SVM), Naive Bayes (NB), Random Forest(RF) and ensemble methods to detect sarcasm-based cyberbullying on textual dataset that was freely accessed online. The preprocessing of their data involved computation of sentiment scores, profanity features and sarcasm detection counts of features like exclamation marks, interjections and question marks. Passing the output to the classifiers aforementioned, SVM and

ensemble methods achieved higher accuracy outperforming others with an average accuracy of 79%. To advance on that, [17] identified cyberbullying based on sarcasm and irony using tweets obtained through query oriented approach. The dataset was pre-processed and features scaled using TF-IDF(Term Frequency-Inverse Document Frequency) technique. Using SVM, RF and KNN methods, their model achieved F1-score of 88.9%.

TF-IDF and Word2Vec feature extraction techniques were used on a dataset comprising of 37,373 tweets by [18]. The authors here decided to experiment with different algorithms such as Logistic Regression (LR), Light Gradient Boosting (LGB), Stochastic Gradient Descent (SGD), RF, AdaBoost, Naive Bayes and SVM. Amongst the classifiers, SGD achieved highest precision of 96.8% and SVM achieved highest recall of 100%. Comparing SVM and Neural Networks (NN), [19] evaluated textual cyberbullying data collected from Kaggle and preprocessed using TF-IDF and sentiment analysis approaches. On the examination of their approach, NN achieved a better accuracy of 92.8% while SVM achieved 90.3%.

### *B. Multimodal based cyberbullying detection*

Multimodal cyberbullying is the use of various communication channels that results in flaming, impersonation, trolling, cyber-stalking, and even sexting. These communication channels may involve use of text, images and other forms of multimedia data [35]. Through analysis of combined modalities for detecting cyberbullying, it is possible for researchers to capture the context, intent and the impact brought about by the online harassment more effectively as well as have more comprehensive understanding of the cyberbullying instances.

Adoption of multimodal strategies to cyberbullying is highly beneficial. One such benefit of the approach is the ability to achieve contextual understanding of cyberbullying incidences. Text based detection systems often struggle to identify the nuanced nature of bullying in social medial platforms like sarcasm, irony and implicit threats. Through incorporation of images, the multimodal based systems can capture subtleties that might be missed in plain texts, and as result, a more comprehension on the intent and emotional impact of the the message can be accurately assessed [36]. In addition to that, use of multimodal approaches can lead to provision of better systems that can be able to detect bullying content from non-bullying content. Furthermore, integration of multiple modalities in detecting online harassment improves the adaptability of the methods to the changing tactics of cyberbullies [37].

Despite the potential benefits of multimodal detection, several challenges exist that make the approach a bit difficult. Some of these difficulties include the complexity of collecting and annotating large corpus that encompasses instances of cyberbullying across multiple modalities, the inadequate nature of human experts to aid in annotation and finally, the difficulties in representing multimodal data, integrating and developing suitable algorithms to automatically detect cyberbullying incidences [38]. Current approaches and Techniques for multimodal cyberbullying detection include use of deep learning techniques and fusion strategies which are discussed below.

*1) Deep learning techniques:* Deep Learning (DL) techniques makes use of Artificial Neural Networks (ANN) with several hidden layers to automatically learn hierarchical representations of complex data. Different deep learning techniques have been used in the detection of cyberbullying based activities. Some of the available researches are discussed: CNN-BiLSTM was developed to detect cyberbullying on multilingual datasets of English, Hindi and Hinglish obtained from Twitter [25].With this model, an accuracy of 87% was achieved. [26] developed a multi-class based cyberbullying based on the Graph Convolutional Network (GCN).Their work involved analyzing the targets of cyberbullying based on ethnic groups, age, gender and religion. The data collected was pre-processed by analyzing the semantic correlations and similarities.Due the imbalanced nature of their data, Dynamic Query Expansion (DQE) was used to address the issue and extract data from Twitter. [27] fined tuned BERT models using tweets collected using crowd sourcing technique. The tweets composed of hate speech sentiments, which was the kind of cyberbullying in the picture. However, with a total of 85948 tweets, the hostile group was tiny with data being skewed and this necessitated use of Focal loss and cross entropy loss functions to be used to classify the sample data. The proposed model by the authors achieved an F1-score of 91%.

BERT model was used in the detection of cyberbullying incidences with posts from different sources: 100000 posts from Wikipedia, 16000 posts from Twitter and approximated 12000 posts from Formspring [28]. The performances of BERT on these datasets was compared with CNN, LSTM and BiLSTM with BERT outperforming them with F1-score of 92% on Formspring data, 91% on Wikipedia data and 94% on Twitter data. Pre-trained BERT model was used with single linear neural networks [29]. The model was tested on 12773 Formspring labelled comments and also 111864 Wikipedia debate comments manually annotated achieving state of the art results of 98% on Formspring data and 96% on Wikipedia data.

*2) Deep learning fusion based cyberbullying detection:* Other than deep learning models, cyberbullying in multimodal data can be accomplished with multimodal fusion strategies which are responsible for combining information (features) from multiple modalities either in classification or regression tasks. The fusion approach comes in two ways: early fusion and late fusion [39]. These ways are summarized in the Fig. 1 and Fig. 2 respectively. During early fusion, features of the two modalities (text + images) are concatenated into single vector which is used as the input patter for the final classifier at the early stages of processing before applying the features to the machine learning model. The main benefit of this fusion is that it is capable of exploiting the correlations and interactions between the low level features of each modality used. In contrast, the late fusion strategy uses decision values from each modality classifiers and combines the decisions at the late stages of processing where the voting, weighting

and rule based decisions are made. The late fusion strategy is considered a better performer even though it comes with an increased cost, where the learning efforts of the model is required more compared to the early fusion strategy [33].

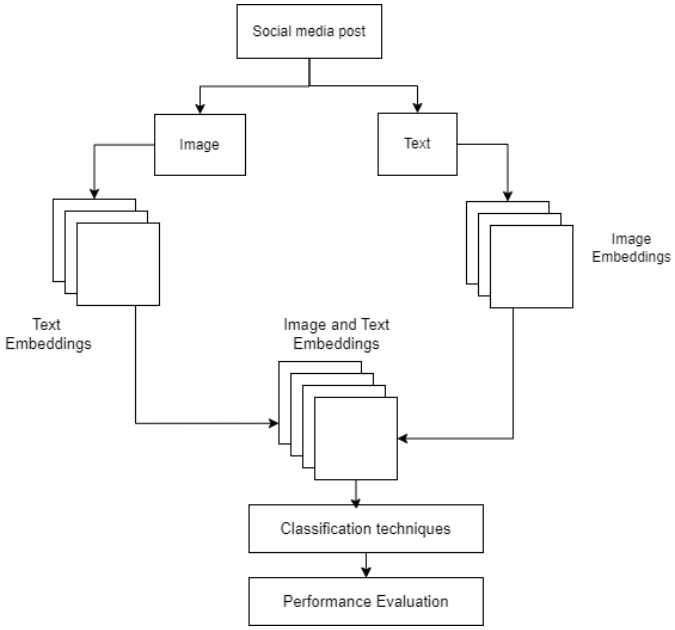


Fig. 1. Early fusion

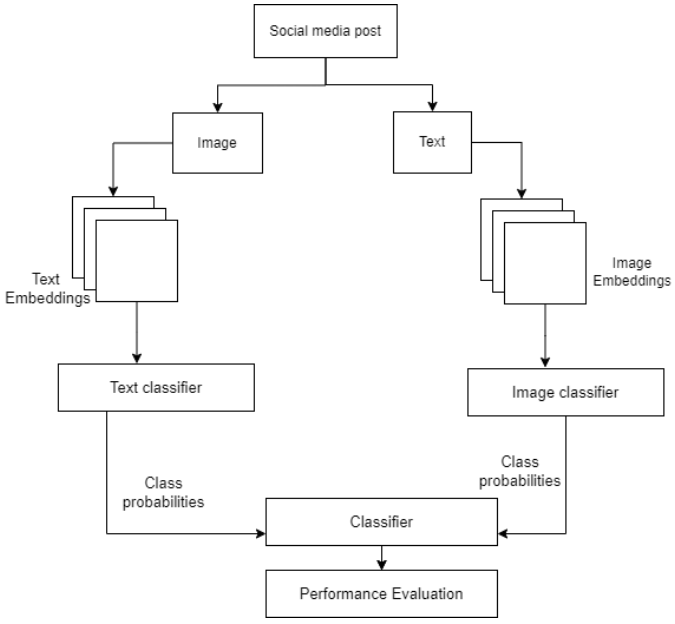


Fig. 2. Late fusion

Studies considering the combination of modalities have a better edge in comparison to individual modalities as they provide a real world occurrence of cyberbullying issues. Such scenarios are evidenced in different research works: A unified multi-modal approach was used in the detection of cyberbullying on three social media sites: Facebook, Twitter and Instagram [31]. With a total of 2100 sample data collected from these sites, their features was extracted with the use of TF-IDF where early fusion approach was used to aid in acquiring the combined features of images and text. During the training, binary cross entropy loss and ReLU activation functions were used resulting to a model with a 68% weighted average.

In addition to that, an hybrid model of CNN and LSTM was also adopted in the detection of cyberbullying activities in the text and image data by [32]. Each model was trained independently and then late fusion technique used to combine the two using datasets from Github and Kaggle. To evaluate the model, Telegram data was used with an accuracy of 85% and 86% for text and image respectively. Similarly, late fusion approach was adopted by [33]. In this case, 10,000 text and image posts from Youtube, Instagram and Twitter were used with two models: CapsNet and ConvNet used for predicting bullying contents for text and image respectively before being combined together in the final stages achieving AUC-ROC of 0.98%.

Deep learning models were used to detect cyberbullying incidences from images and texts constituting of 2100 records [40]. The textual data included 884 records of bullying incidences and 1216 non-bullying. For the image data, only 464 incidences were bullying while the rest non-bullying. The authors in this study expertly combined their textual and image data to have a combined corpora of 1481 bullying instances and 619 non-bullying. Different pre-processing techniques were applied for text (lemmatization, stemming, and stop words removal) and images (segmentation and enhancements). In the fusion stages, RoBERTa was used to extract features of textual contents and XCeption for image data before the features being fused at later stages with Light Gradient Boosting machine learning approach as the classifier. The approach produced recall and F1-score of 92% and 86%. Also, [41] explored the capabilities of deep learning approaches to detect offensive hate speech, fake news and cyberbullying incidences for Bengali memes data and text. The researchers used Conv-LSTM and XLM-RoBERTa models for the textual data achieving F1-scores of 78% and 82% respectively. As of memes dataset, the authors used ResNet-152 and DenseNet-161 models yielding F1 scores of 78% and 79%, respectively. In the later stages of processing, the researchers late fusion of XLM-RoBERTa + DenseNet-161 performed the best, yielding an F1 score of 83%.

To identify offensive content in social media sites (Reddit, Facebook, Twitter and Instagram), [42] created MultiOFF dataset consisting of images and textual data embedded in them. This data was cleaned and pre-processed through creation of vector spaces for both images and text. With the help of CNN, features from the images were extracted and combined with vector sequences of the text counterpart. To increase the size of the data, pre-trained embeddings generated were generated with the help of hybrid versions of VGG16 (LSTM + VGG16, BiLSTM + VGG16, CNNText + VGG16). The generated joint embeddings were passed through

a classifier of LSTM model using the late fusion strategy. [43] proposed a novel task of detecting hate speech content in multimodal settings. A dataset comprising of 150,000 images embedded with text was collected, pre-processed and trained on three different architectures: Feature Concatenation Model (FCM), Spatial Concatenation Model (SCM) and Textual Kernel Model (TKM) all coming from the deep networks of CNN + RNN. In the evaluation of their model, TKM model achieved the best performance with an F-score of 70%.

## IV. Findings from the literature

Use of both unimodal and multimodal modes of cyberbullying detection are important in identifying different cases of cyberbullying across social media platforms with each modality having its strengths and weaknesses. For the unimodal context, the focus involves use of single modality like text or images. From the studies above, text based modality have gained a lot of attention in the unimodal context, in comparison to image based modalities achieving state of the art results as witnessed in the works of [18] and [19]. However, this modality is prone to missing some visual or auditory cues such as tone or facial expressions when text-based approach is followed. On the other hand, multimodal approach offers a more comprehensive approach to combining multiple modalities, and providing contextual cyberbullying insights. Despite the benefits, the approach requires more computational resources and complex fusion techniques.

Looking at different machine learning techniques from the review above, findings suggests conventional machine learning models are rarely used in the multimodal cyberbullying detection even though they achieve better results in unimodal context. On the other hand, deep learning based approaches like BERT are used in analyzing and extracting features for textual modalities in the multimodal aspects while deep CNN networks are mostly used in the imaging modality. In fusion aspects, late fusion is highly used compared to early fusion approaches. This shows that features of individual modalities are extracted and trained on different models separately before combined in the later stages of detecting cyberbullying incidences. Some of the work from the literature is summarized in Table I. below.

TABLE I
Summary of review

| Authors | Classifier | Performance | Source of data |
|---|---|---|---|
| [15] | Naïve Bayes<br>K-Nearest Neighbors | KNN = 73%<br>NB = 72% | Facebook |
| [16] | SVM, naive Bayes<br>Logistic Regression<br>Random Forest, and<br>ensemble method | SVM = 79%<br>LR = 78%<br>RF = 76.7%<br>NB = 76%<br>Ensemble = 79% | Twitter and<br>Formspring<br>datasets |
| [17] | Naive Bayes<br>SVM<br>random forests<br>and K-nearest<br>neighbors<br>J48<br>JRip | NB = 80.8%<br>SVM = 84.4%<br>RF = 88.3%<br>KNN = 75.3%<br>J48 = 88.3%<br>JRip = 89.9% | Twitter |
| [18] | LR<br>LGBM<br>SGD<br>RF<br>AdaBoost (ADB)<br>Naive Bayes<br>and SVM | LR = 90.57%<br>LGBM = 90.55%<br>SGD = 90.6%<br>RF = 89.84%<br>ADB = 89.30%<br>NB = 81.39%<br>SVM = 67.13 | Twitter |
| [19] | Neural Networks<br>SVM | NN = 92.8%<br>SVM = 90.3% | Kaggle |
| [25] | CNN<br>BiLSTM | BiLSTM = 87% | Twitter |
| [27] | BERT | BERT = 91% | Twitter |
| [28] | BERT | F1-score = 94% | Twitter<br>Wikipedia<br>Formspring |
| [31] | CNN | Recall = 74% | Facebook<br>Twitter<br>Instagram<br>Google |
| [32] | CNN<br>LSTM | CNN = 85%<br>LSTM = 86% | GitHub<br>Kaggle |
| [33] | CapsNet<br>ConvNet | AUC-ROC = 98% | Youtube<br>Instagram<br>Twitter |
| [40] | RoBERTa<br>XCeption<br>LGB | LGB<br>F1-score = 86% | Instagram<br>Twitter |
| [41] | Conv-LSTM<br>XLM-RoBERTa<br>RoBERTa +<br>DenseNet-161 | F1-score = 83% | Bengali<br>memes |
| [43] | CNN<br>+ RNN | F1-score = 70% | Twitter |

## V. Conclusion and Future works

From the summaries and review of the literature, researchers are shifting their focus from using classical machine learning techniques to considering use of deep learning techniques, and even use of pre-trained models like BERT. Publicly available datasets from Kaggle, Facebook, Instagram ,Form spring and Twitter being the most used sources to provide data for cyberbullying detection. To achieve better results, classical approaches are adopted to solve limited annotations problems.

Despite having a lot of bullying information shared across the social media platforms, research involving multi-modal cyberbullying detection has received less attention as many researchers have explored widely on unimodal approach of using textual content to detect cyberbullying. This is however considered to be a limitation in considering cyberbullying in real world scenarios where both imagery and textual content are used. Multimodal cyberbullying detection presents a promising approach to tackle the complex and evolving nature of online harassment. By incorporating multiple modalities, the systems developed of multiple modalities have the potential to capture context, improve accuracy, and adapt to new cyberbullying tactics. However, challenges related to data collection, class imbalance, modality integration, and contextual understanding need to be addressed. With ongoing research, future researchers are recommended to adopt multimodal approaches to detecting online harassment's and use deep learning approaches to ensure high detection accuracy of cyberbullying incidences.

## REFERENCES


[1] K. E. Anderson, 'Getting acquainted with social networks and apps: it is time to talk about TikTok', Library hi tech news, 2020.

[2] Shakambhari, J. S. Raj, and S. Anantha Babu, 'Smart Cyberbullying detection with Machine Learning', in Disruptive Technologies for Big Data and Cloud Applications: Proceedings of ICBDCC 2021, Springer, 2022, pp. 237–248.

[3] P. K. Roy and F. U. Mali, 'Cyberbullying detection using deep transfer learning', Complex and Intelligent Systems, vol. 8, no. 6, pp. 5449–5467, 2022.

[4] S. Hinduja and J. W. Patchin, 'Cyberbullying Identification, Prevention, and Response. 2021'. 2014.

[5] J. Wang, K. Fu, and C.-T. Lu, 'Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection', in 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 1699–1708.

[6] M. Duggan, 'Online harassment 2017', Pew Research Center: Internet, Science and Tech, 11-Jul-2017. [Online]. Available: https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/. [Accessed: 28-Jun-2023].

[7] M. Anderson, 'A majority of teens have experienced some form of cyberbullying', Pew Research Center: Internet, Science and Tech, 27-Sep-2018. [Online]. Available: https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/. [Accessed: 28-Jun-2023].

[8] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, 'Social media cyberbullying detection using machine learning', Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, 2019.

[9] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, 'Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection', Inf. Process. Manag., vol. 58, no. 4, p. 102600, Jul. 2021.

[10] M. A. Al-Garadi et al., 'Predicting cyberbullying on social media in the big data era using machine learning algorithms: review of literature and open challenges', IEEE Access, vol. 7, pp. 70701–70718, 2019.

[11] 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', in 2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), 2020, pp. 1–6.

[12] A. Veltman, D. W. J. Pulle, and R. W. De Doncker, 'The Transformer', in Power Systems, Cham: Springer International Publishing, 2016, pp. 47–82.

[13] B. Haidar, M. Chamoun, and A. Serhrouchni, 'Arabic cyberbullying detection: Using deep learning', in 2018 7th international conference on computer and communication engineering (iccce), 2018, pp. 284–289.

[14] M. A. Al-Ajlan and M. Ykhlef, 'Deep learning algorithm for cyberbullying detection', International Journal of Advanced Computer Science and Applications, vol. 9, no. 9, 2018.

[15] Nureni Ayofe AZEEZ, S. Misra, Omotola Ifeoluwa LAWAL, and J. Oluranti, 'Identification and detection of cyberbullying on Facebook using machine learning algorithms', J. Cases Inf. Technol., vol. 23, no. 4, pp. 1–21, Jan. 2022.

[16] J. Mehta, 'Cyber Bullying Detection using Machine Learning', Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 9, pp. 144–151, Sep. 2021.

[17] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, 'Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection', Inf. Process. Manag., vol. 58, no. 4, p. 102600, Jul. 2021.

[18] A. Muneer and S. M. Fati, 'A comparative analysis of machine learning techniques for cyberbullying detection on twitter', Future Internet, vol. 12, no. 11, p. 187, 2020.

[19] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, 'Social media cyberbullying detection using machine learning', Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, 2019.

[20] Nureni Ayofe AZEEZ, S. Misra, Omotola Ifeoluwa LAWAL, and J. Oluranti, 'Identification and detection of cyberbullying on Facebook using machine learning algorithms', J. Cases Inf. Technol., vol. 23, no. 4, pp. 1–21, Jan. 2022.

[21] J. Mehta, 'Cyber Bullying Detection using Machine Learning', Int. J. Res. Appl. Sci. Eng. Technol., vol. 9, no. 9, pp. 144–151, Sep. 2021.

[22] Z. L. Chia, M. Ptaszynski, F. Masui, G. Leliwa, and M. Wroczynski, 'Machine Learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection', Inf. Process. Manag., vol. 58, no. 4, p. 102600, Jul. 2021.

[23] A. Muneer and S. M. Fati, 'A comparative analysis of machine learning techniques for cyberbullying detection on twitter', Future Internet, vol. 12, no. 11, p. 187, 2020.

[24] J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer, and A. Mohammed, 'Social media cyberbullying detection using machine learning', Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 5, 2019.

[25] M. Raj, S. Singh, K. Solanki, and R. Selvanambi, 'An application to detect cyberbullying using machine learning and deep learning techniques', SN Comput. Sci., vol. 3, no. 5, p. 401, Jul. 2022.

[26] J. Wang, K. Fu, and C.-T. Lu, 'SOSNet: A graph convolutional network approach to fine-grained cyberbullying detection', in 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 2020.

[27] M. Behzadi, I. G. Harris, and A. Derakhshan, 'Rapid Cyber-bullying detection method using Compact BERT Models', in 2021 IEEE 15th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2021.

[28] S. Paul and S. Saha, 'CyberBERT: BERT for cyberbullying identification', Multimed. Syst., vol. 28, no. 6, pp. 1897–1904, Dec. 2022.

[29] J. Yadav, D. Kumar, and D. Chauhan, 'Cyberbullying Detection using Pre-Trained BERT Model', in 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020.

[30] F. Elsafoury, S. Katsigiannis, S. R. Wilson, and N. Ramzan, 'Does BERT pay attention to cyberbullying?', in Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event Canada, 2021.

[31] K. Kumari, J. P. Singh, Y. K. Dwivedi, and N. P. Rana, 'Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach', Soft Comput., vol. 24, no. 15, pp. 11059–11070, Aug. 2020.

[32] Vijayakumar and H. P. D. Adolf, 'Multimodal cyberbullying detection using hybrid deep learning algorithms', Int. J. Appl. Eng. Res., vol. 16, no. 7, p. 568, Jul. 2021.

[33] A. Kumar and N. Sachdeva, 'Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network', Multimed. Syst., vol. 28, no. 6, pp. 2043–2052, Dec. 2022.

[34] P. K. Roy and F. U. Mali, 'Cyberbullying detection using deep transfer learning', Complex Intell. Syst., vol. 8, no. 6, pp. 5449–5467, Dec. 2022.

[35] 'Multimodal essay', Scribd. [Online]. Available: https://www.scribd.com/document/546479189/multimodal-essay. [Accessed: 03-Jul-2023].

[36] S. Kim, A. Razi, G. Stringhini, P. J. Wisniewski, and M. De Choudhury, 'A human-centered systematic literature review of cyberbullying detection algorithms', Proc. ACM Hum. Comput. Interact., vol. 5, no. CSCW2, pp. 1–34, Oct. 2021.

[37] U. Shrivastav, P. Srivastava, and V. Tripathi, 'A Survey on Cyberbullying Detection Techniques', in 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2023, pp. 628–633.

[38] M. S. Jahan and M. Oussalah, 'A systematic review of hate speech automatic detection using natural language processing', Neurocomputing, vol. 546, no. 126232, p. 126232, Aug. 2023.

[39] X. Liu, Z. Wang, and L. Wang, 'Multimodal fusion for image and text classification with feature selection and dimension reduction', J. Phys. Conf. Ser., vol. 1871, no. 1, p. 012064, Apr. 2021.

[40] S. Pericherla and Ilavarasan, 'Cyberbullying detection on multi-modal data using pre-trained deep learning architectures', Ing. Solidar., vol. 17, no. 3, pp. 1–20, Sep. 2021.

[41] M. R. Karim, S. K. Dey, T. Islam, M. Shajalal, and B. R. Chakravarthi, 'Multimodal hate speech detection from Bengali memes and texts', arXiv [cs.CL], 19-Apr-2022.

[42] S. Suryawanshi, B. R. Chakravarthi, M. Arcan, and P. Buitelaar, 'Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text', in Proceedings of the second workshop on trolling, aggression and cyberbullying, 2020, pp. 32–41.

[43] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, 'Exploring Hate Speech Detection in Multimodal Publications', in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1459–1467.