





Article

Unraveling Cyberbullying Dynamis: A Computational Framework Empowered by Artificial Intelligence

Liliana Ibeth Barbosa-Santillán ^{1,*} , Bertha Patricia Guzman-Velazquez ^{2,*} , Ma. Teresa Orozco-Aguilera ^{2,*} 
and Leticia Flores-Pulido ^{3,*} 

¹ Departamento de Sistemas, Universidad de Guadalajara, Guadalajara 44100, Mexico

² Instituto Nacional de Astrofísica, Óptica y Electrónica, Tonantzintla, Puebla 72840, Mexico

³ Facultad de Ciencias Básicas, Ingeniería y Tecnología, Universidad Autónoma de Tlaxcala, Apizaco 90300, Mexico

* Correspondence: ibarbosa@cucea.udg.mx (L.I.B.-S.); bpguzman@inaoep.mx (B.P.G.-V.); toa@inaoep.mx (M.T.O.-A.); leticia.flores.p@uatx.mx (L.F.-P.)

Abstract: Cyberbullying, which manifests in various forms, is a growing challenge on social media, mainly when it involves threats of violence through images, especially those featuring weapons. This study introduces a computational framework to identify such content using convolutional neural networks of weapon-related images. By integrating artificial intelligence techniques with image analysis, our model detects visual patterns associated with violent threats, creating safer digital environments. The development of this work involved analyzing images depicting scenes with weapons carried by children or adolescents. Images were sourced from social media and spatial repositories. The statistics were processed through a 225-layer convolutional neural network, achieving an 86% accuracy rate in detecting weapons in images featuring children, adolescents, and young adults. The classifier method reached an accuracy of 17.86% with training over only 25 epochs and a recall of 14.2%. Weapon detection is a complex task due to the variability in object exposures and differences in weapon shapes, sizes, orientations, colors, and image capture methods. Segmentation issues and the presence of background objects or people further compound this complexity. Our study demonstrates that convolutional neural networks can effectively detect weapons in images, making them a valuable tool in addressing cyberbullying involving weapon imagery. Detecting such content contributes to creating safer digital environments for young people.

Keywords: cyberbullying; CNN; deeplearning



Academic Editors: Arkaitz Zubiaga and Heming Jia

Received: 27 November 2024

Revised: 23 December 2024

Accepted: 16 January 2025

Published: 22 January 2025

Citation: Barbosa-Santillán, L.I.; Guzman-Velazquez, B.P.; Orozco-Aguilera, M.T.; Flores-Pulido, L. Unraveling Cyberbullying Dynamis: A Computational Framework Empowered by Artificial Intelligence. *Information* **2025**, *16*, 80. <https://doi.org/10.3390/info16020080>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Cyberbullying is defined as the deliberate and repeated use of information and communication technologies (ICT) to intimidate, harass, humiliate, or harm others. Unlike traditional bullying, cyberbullying presents unique challenges due to its pervasive and often anonymous nature, making it especially devastating. The constant availability of digital platforms, accessible 24 h a day, means that victims cannot find refuge even in the privacy of their own homes. This relentless presence inflicts emotional wounds that, although invisible, are as painful as physical ones and can have profound and lasting consequences on the mental health and well-being of victims. The immediacy of cyberbullying, combined with the aggressors' ability to conceal their identity, amplifies the psychological damage and instills a sense of constant fear. Because of these severe effects, efforts to combat cyberbullying have multiplied, and artificial intelligence (AI)-based detection

mechanisms are emerging as one of the most promising strategies to tackle this pervasive problem. According to [1], current challenges in using Artificial Intelligence (AI) to detect cyberbullying include the difficulty of identifying irony and sarcasm in online content, which is one of the primary obstacles. The nuanced forms of communication can often mask abusive behavior in a challenging way for AI algorithms to interpret accurately. Furthermore, cyberbullying is a multimodal phenomenon involving not only text but also images and videos. This diversity in content complicates automatic detection, as different modalities may not individually exhibit abusive characteristics yet collectively contribute to bullying behavior. Another significant challenge arises from the need to consider repetition and context in defining cyberbullying. Unlike straightforward abusive language, cyberbullying may depend on the repeated nature of the behavior or context provided by previous interactions, making it difficult for AI to assess a single message in isolation accurately. Additionally, identifying exclusion in cyberbullying adds another layer of complexity. For instance, tagging people to exclude others deliberately may be subtle and misinterpreted by algorithms, complicating the detection process and potentially resulting in false positives or negatives. These challenges highlight the complexity of using AI to address cyberbullying, underscoring the need to improve current tools and approaches. Cyberbullying detection using AI and machine learning has gained significant attention due to the rise of social media usage. Multiple studies have explored supervised machine learning approaches to identify cyberbullying content automatically. One such study, conducted by [2], presents a method to identify cyberbullying in social networking platforms using machine learning techniques. The primary objective is to develop a model to effectively detect cyberbullying in social network conversations. The dataset of 1608 conversations, with 804 identified as instances of cyberbullying and 804 as non-cyberbullying, was accomplished. The study evaluated two classifiers: Support Vector Machines (SVM) and Neural Networks (NN). The Neural Network achieved an accuracy of 92.8%, while the SVM reached 90.3%. The results indicated that the Neural Network outperformed the SVM in terms of accuracy and F-score, suggesting that the proposed approach can enhance cyberbullying detection and contribute to a safer use of social networks. Researchers have primarily concentrated on analyzing textual data from social media platforms. However, as cyberbullying has expanded from text-based forms to include images, detection has become more challenging due to the absence of clear textual representation in visual content. Some researchers have proposed incorporating image analysis into detection methods to address this. In an article by [3], the authors address the growing problem of cyberbullying on social networking platforms, where visual content is particularly prevalent. The authors propose an innovative technique called CNBD (Combinational Network for Bullying Detection), which integrates a picture transformer model (BEiT) with a multilayer perceptron (MLP) network. Additionally, they incorporate image description generation (Image Captioning) and Optical Character Recognition (OCR) to extract text from images, improving accuracy in cyberbullying detection. The results demonstrate that CNBD achieves an accuracy of 98.23%, surpassing previous methods. In [4], the study examines cyberbullying through images. The study aimed to identify the visual characteristics associated with cyberbullying and demonstrate their efficacy in detecting offensive content. A dataset of 19,300 images about cyberbullying was assembled through the use of keywords extracted from victim accounts, thereby depicting actual scenarios. Five key visual factors were identified: body posture, facial emotion, objects present, gestures, and social aspects, which differ from those typically associated with traditional forms of offensive content. A multimodal classification model was developed using these factors, achieving an accuracy of 93.36% in detecting cyberbullying. Furthermore, the study demonstrated that 39.32% of cyberbullying images can evade current detectors, underscoring the need for more sophisticated methodologies.

The rise of social media platforms has amplified the occurrence of cyberbullying, posing significant risks to children and adolescents who are particularly vulnerable to online harassment and exposure to violent content [5–7]. To mitigate these risks, researchers have increasingly focused on leveraging artificial intelligence (AI) for automated detection and prevention of harmful online behaviors [8–10]. Convolutional neural networks (CNNs) have emerged as a powerful tool in this domain, enabling the identification of weapons in images, which is a critical component in addressing cyberbullying involving threats of violence [11–14]. Studies have shown that CNNs can effectively handle the complexity and variability of image data, thus providing robust solutions for cyberbullying detection [15–17]. The growing body of research underscores the need for efficient and accurate monitoring systems to enhance AI models for social good [18–20]. This paper builds upon previous work by developing a CNN-based framework to detect weapon imagery on social media, contributing to safer digital environments [21–23]. In contrast to the previously referenced studies, which employ text analysis, text extraction from images, and a combination of text analysis and images to detect cyberbullying, our approach is exclusively focused on image analysis. In the present study, the visual characteristics of the images are extracted directly and subsequently used to perform a classification to detect cyberbullying occurrences.

Table 1 shows an additional comparative table to the previously mentioned state-of-the-art works, detailing the method used, the type of data, and the evaluation percentage of each. These works are then compared to our research, where it is noted that only three of the nine works listed in the comparative table focused on detecting bullying through images, while the rest used text data sets for detection. Of the three works utilizing image detection, only Kumar’s work surpasses our percentage with a 98% accuracy compared to our 86%, an average percentage within the comparative table. Kumar’s convolutional neural network model is a hybrid or composite approach, in contrast to the model used in our research, which derives its strength from many hidden layers.

Table 1. Comparative Table of Cyberbullying Detection Methods.

Author and Year	Title	Method and Accuracy	Data Sets
Zhang, et al., 2016 [24]	Cyberbullying detection with a pronunciation-based convolutional neural network.	Pronunciation-based convolutional neural network with 96%	Two cyberbullying datasets collected from Twitter and Formspring.me (only text)
Vijayakumar et al., 2021 [25]	Multimodal cyberbullying detection using hybrid deep learning algorithms.	Hybrid model (CNN and LSTM) with 85%	Images and text extracted from GitHub and Kaggle and tested with Telegram real-time data (Kaggle Toxic Comment Classification Challenge dataset)
Dewani, et al., 2021 [26]	Cyberbullying detection: advanced pre-processing techniques and deep learning architecture for Roman Urdu data.	Recurrent neural networks (RNN) with 85%	Urdu and Roman Urdu instances (only text)
Raj et al., 2022 [27]	An application to detect cyberbullying using machine learning and deep learning techniques.	CNN-BiLSTM (Bidirectional Long Short-Term Memory) with 95%	The acquired labeled data in 3 languages
Aldhyani et al, 2022 [28]	Cyberbullying identification system based on deep learning algorithms.	Hybrid deep learning architecture consisting of convolutional neural networks integrated with Bidirectional Long Short-Term Memory networks (CNN-BiLSTM) and single BiLSTM models with 99%	A binary class dataset with 115,864 samples and a multiclass dataset with 39,869 samples
Kumar et al., 2022 [14]	Multimodal cyberbullying detection using capsule network with dynamic routing and deep convolutional neural network.	CapsNet-ConvNet model (deep neural networks) with dynamic routing with 98%	Separating text from the image using Google Lens of Google Photos App
Khafajeh, 2024 [29]	Cyberbullying Detection in Social Networks Using Deep Learning.	Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and transformer models like Bidirectional Encoder Representations from Transformers (BERT) with 87.3%	11,000 Facebook comments labeled as clean or cyberbullying
Baiganova et al., 2024 [30]	Hybrid Convolutional Recurrent Neural Network for Cyberbullying Detection on Textual Data.	Hybrid neural network architecture combining Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) with 87%	Identification of suicidal tendencies within the textual milieu of Reddit’s digital content
Gutiérrez-Batista, et al., 2024 [31]	Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model.	96.25% with Linear Regression	3 datasets: Bullying V3.0 (3889 documents), hate-speech (10,360 documents), and Myspace (2029 documents)
Our Work	Unraveling Cyberbullying Dynamics: A Computational Framework Empowered by Artificial Intelligence.	Convolutional neural network with 86%	30 images from Depositphotos dataset

This paper is structured as follows: We begin with Section 2, which provides background information and outlines our proposed architecture. In Section 3, we introduce Addressing Cyberbullying Dynamic Images. In Section 4, we discuss the experiments conducted. Section 5 presents the results obtained from these experiments. Lastly, we conclude the paper with our final remarks.

2. Unraveling Cyberbullying Dynamis

The proposed method for Unraveling Cyberbullying Dynamis consists of three stages:

1. **ROI Extraction:** In this initial stage, an ROI (Region of Interest) extraction process is applied to limit the analysis to specific regions that may contain weapons. Focusing on the ROI makes the analysis more effective and targeted, enhancing the accuracy of detecting cyberbullying-related content.
2. **Segmentation:** The second stage involves segmentation using the previously extracted ROI. This segmentation step is crucial as it lays the groundwork for a comprehensive analysis and understanding of the dynamics of cyberbullying. Precise segmentation ensures that the relevant areas are accurately identified for further examination.
3. **Neural Convolutional Network:** In the final stage, a neural convolutional network is employed to train and test the model. The data is split into two sets: one for training the model and the other for testing its performance. The use of a neural convolutional network enables efficient analysis, evaluation, and prediction within the proposed method, leveraging advanced AI techniques for enhanced detection capabilities.

This multi-stage methodology provides a comprehensive approach to unraveling Cyberbullying Dynamis, leveraging artificial intelligence techniques for enhanced analysis and understanding as shown in Figure 1.

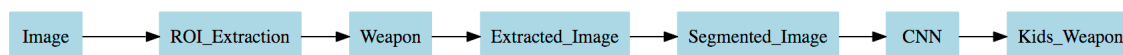


Figure 1. Unraveling Cyberbullying Dynamics.

2.1. ROI Extraction

Region of Interest (ROI) extraction is critical in detecting and analyzing weapons within digital platforms. This process involves identifying and isolating key areas within images that are likely to contain instances of weapons. In images, ROIs might include specific shapes, colors, or patterns characteristic of various types of weapons. Effective ROI extraction enables more accurate and efficient analysis, allowing for the timely identification and mitigation of potential threats.

2.2. Segmentation

Segmentation is a vital step in the detection and analysis of weapons within digital platforms. This process involves partitioning images into distinct regions likely to contain weapons. For images, segmentation might involve delineating specific shapes, colors, or patterns characteristic of various weapon types. Effective segmentation enables more precise and efficient analysis, facilitating the timely identification and mitigation of potential threats.

2.3. Neural Convolutional

Convolutional Neural Networks (CNNs) play a critical role in detecting and analyzing weapons within digital platforms. This process involves applying CNNs to analyze and interpret visual data, such as images, to identify the presence of weapons. CNNs are adept at recognizing patterns and features, such as specific shapes, textures, and colors,

characteristic of various types of weapons. By learning these features through training, CNNs can effectively differentiate between harmless objects and potential threats. The use of CNNs enables more accurate and efficient analysis, allowing for the timely identification and mitigation of possible dangers.

Algorithm 1, designed for detecting weapons-related content, includes several key functions: `HandleCyberbullyingObject` processes objects to identify potential weapons; `HandleCyberbullyingFace` uses facial recognition to detect faces linked to threatening behavior involving weapons; and `HandleCyberbullyingPattern` identifies patterns of weapon-related activities through CNN. While these functions typically operate independently within the same iteration to avoid interference, their results are combined to form a comprehensive assessment. Outputs from functions like `IsCyberbullyingText`, `IsCyberbullyingFace`, `IsCyberbullyingContent`, and `IsCyberbullyingPatternDetected` are integrated to provide an understanding of the content, indicating a higher likelihood of threat when multiple aspects are flagged. The functions `IsCyberbullyingPatternDetected` and `HandleCyberbullyingPattern` operate without explicit arguments by utilizing the global context generated from previous function outcomes, which include processed objects, recognizing faces, and detected text patterns, ensuring a detailed detection process.

Algorithm 1 Addressing Cyberbullying in Dynamic Images

```

1: procedure ADDRESSCYBERBULLYINGDYNAMICIMAGES(image_sequence)
2:   for each frame in image_sequence do
3:     labels ← DetectObjects(frame)
4:     for each object in labels do
5:       if ObjectContainsText(object) then
6:         text ← ExtractTextFromObject(object)
7:         if IsCyberbullyingText(text) then
8:           HandleCyberbullyingObject(object)
9:         end if
10:      end if
11:      if ObjectContainsFaces(object) then
12:        faces ← ExtractFacesFromObject(object)
13:        for each face in faces do
14:          if IsCyberbullyingFace(face) then
15:            HandleCyberbullyingFace(face)
16:          end if
17:        end for
18:      end if
19:      if ObjectContainsSensitiveContent(object) then
20:        if IsCyberbullyingContent(object) then
21:          HandleCyberbullyingObject(object)
22:        end if
23:      end if
24:    end for
25:    if IsCyberbullyingPatternDetected() then
26:      HandleCyberbullyingPattern()
27:    end if
28:  end for
29: end procedure

```

3. Experiments

The collection of images includes scenes in two contexts: (a) where an adolescent or young person carries the weapon, and (b) where a young person or adolescent and a weapon appear within the same scene. A careful selection of the data corpus is made, which includes controlled scenes where the weapons are dark in color, as opposed to the

bright colors typically associated with toy weapons. It is important to note that there are also black toy weapons.

The data is anonymized to protect individual identities, and robust security measures are implemented to safeguard the data against breaches. Privacy protection is a core priority, with measures in place to ensure data minimization and access control, thereby limiting the data to what is absolutely necessary and allowing access only to authorized individuals. The societal implications of the model's application are significant, as it aims to enhance safety and well-being by detecting potential threats in environments where children and adolescents are involved. This proactive approach helps to mitigate risks and prevent harmful incidents, thereby fostering a safer community. However, addressing potential biases and considering the impact on different societal groups can significantly enhance the credibility and relevance of the study. Although this aspect is beyond the scope of the article, our work is specifically focused on detecting weapons in images to prevent harmful incidents.

During the experiment, a specific image size of 640 was utilized, and the images were normalized to ensure consistency. A deterministic approach, with a ratio mask of 4, was adopted to optimize the model's performance. To assess the model's effectiveness, the dataset was split into 10.71% for testing, 71.42% for training, and 17.36% for validation. This splitting enables thorough evaluation and validation of the model's capabilities. After 25 epochs of training, the model achieved promising results, with an accuracy of 86% and a recall of 14.2%. These results reflect the model's proficiency in capturing and recognizing patterns within the images. The model was executed on a Tesla T4 GPU, which took approximately 27 min to complete. This GPU, equipped with impressive computational power, boasting 15,102 CUDA cores and 8.4 GigaFLOPS, facilitated the efficient processing of the model's calculations. The model, which comprises 225 layers, utilizes a substantial number of parameters and gradients. It consisted of 3,011,043 parameters and 3,011,027 gradients, indicating its complexity and ability to capture intricate details within the data. A learning rate of 0.01 was employed to guide the model's optimization process. This rate dictated the step size taken during the model's learning process, influencing the convergence and overall performance. The optimization employed a gradient-based method, specifically a convolutional neural network (CNN). This method allowed the model to learn and extract meaningful features from the images, enabling accurate classifications and predictions. In summary, the experiment utilized a normalized image size 640 and employed a deterministic optimization process with a ratio mask of 4. The dataset was divided into three segments: 10.71% for testing, 71.42% for training, and 17.36% for validation. After 25 epochs of training, the model demonstrated an accuracy of 86% and a recall of 14.2%. The training process was executed on a Tesla T4 GPU, which completed the task in 27 min. The model architecture was extensive, consisting of 225 layers with a substantial parameter count of 3,011,043 and a gradient count of 3,011,027. Optimization was guided by a learning rate of 0.01.

An intricate segmentation algorithm is employed within the given set of images to identify and isolate a weapon's presence accurately. This cutting-edge technique enables the system to distinguish the weapon from its surroundings, yielding precise results across the multiple images under consideration. The results of the image recognition process applied to these images are presented in Figure 2.

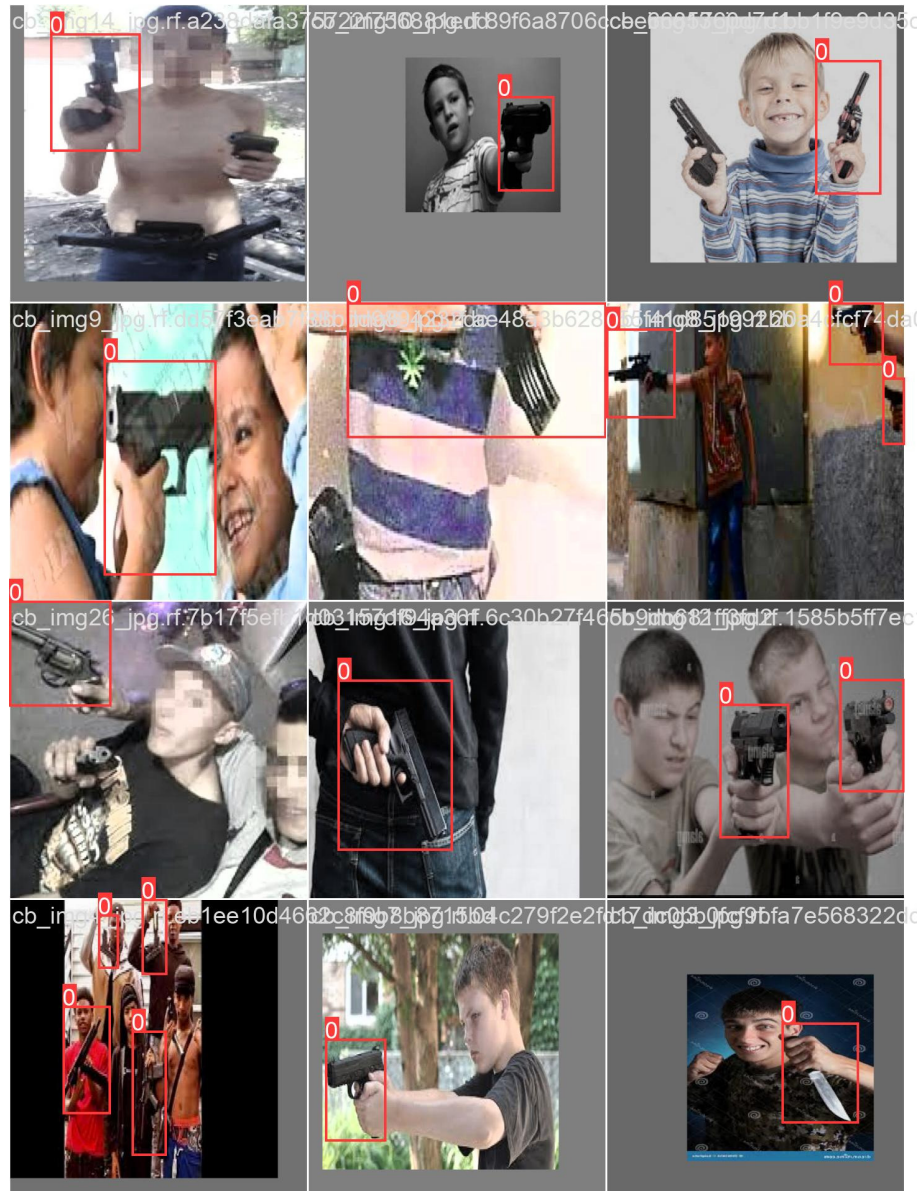


Figure 2. An intricate segmentation algorithm is employed within the given set of images to identify and isolate a weapon’s presence accurately. This cutting-edge technique enables the system to distinguish the weapon resulting from the red box from its surroundings, yielding precise results across the multiple images under consideration.

The rain figure visually represents a distinct split within the part box, serving as a notable indicator of the presence of both COS and DAL losses. This division within the figure signifies the impact and relevance of these particular losses in the context of the analyzed data or scenario (Figure 3 and Table 2).

Table 2. Summary of Model Parameters and Values.

Parameters	Values
Layers	225
Optimizer	Gradient
Learning rate	0.01
Classifier	Convolutional Neural Network

The scenario’s context consists of 201 of 225 layers with a substantial parameter count of 3,011,043, a gradient count of 3,011,027, and optimization by a learning rate of 0.01.

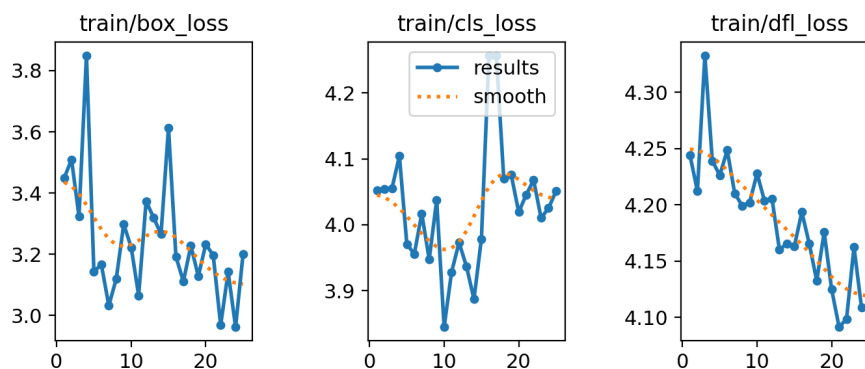


Figure 3. Represents a distinct split within the part box, serving as a notable indicator of the presence of both COS and DAL.

4. Results

The results of this project demonstrate significant advancements in the detection of weapons through digital image processing using a convolutional neural network (CNN) model. Key findings from the implementation are as follows: The overall accuracy for detecting weapons in the test dataset was 10.71%. Within the subset classification for the test data, an accuracy of 71.42% was achieved. For the validation data, a classification rate of 17.86% was obtained. These figures indicate that the CNN model exhibits a notable capability in identifying weapons, though there are variations in accuracy across different data subsets. The study achieved a precision rate of 86% and a recall rate of 14.2%. These metrics reflect the model's ability to correctly identify weapons among images containing children and adolescents. While the precision rate is high, indicating reliable detection when the model predicts a gun, the recall rate suggests that further work is needed to increase the model's capability to detect all relevant instances of weapons. Another noteworthy aspect of the project was training efficiency. The training was completed with just 25 epochs, highlighting the efficiency of the learning process. This rapid training was achieved on a computer with a model 1510M, taking only 27 min and utilizing 13 s at 8.1 GLFOPS. The quick training times suggest that the model can be effectively used in real-time applications without demanding excessive computational resources. The CNN model used for this project comprised 225 layers and was tested with a learning rate of 0.01. The training process involved a gradient optimizer managing approximately 3,011,043 parameters and 3,011,027 gradients. The model's complexity and capability in handling diverse and challenging image data and showcases demonstrates the sophisticated architecture designed to tackle the problem effectively. Lastly, the CNN model successfully recognized weapons within the images where children or adolescents are present. This recognition capability is crucial for the study's aim of supporting crime prevention and safeguarding youth. Detecting weapons in these contexts can provide valuable insights and early interventions to prevent potential harm. The high precision rate illustrates the model's reliability, while the efficiency of the training process emphasizes its practical applicability. However, the recall rate suggests areas for further improvement, particularly in optimizing the model to capture more instances of weapons accurately.

The model has significantly improved in the latest training session, achieving a "Training 4" state. During this period, the Box Loss, a crucial metric that measures the error in the bounding box predictions, was recorded at 3.4. This value indicates that while the model performs well, there is room for further refinement and optimization to minimize prediction inaccuracies and enhance overall performance. Additionally, in the fourth training session, the highest peak of Box Loss was 3.85, with the lowest point recorded in session 24, as shown in Figure 4a.

The model experienced a Cls Loss of 3.85, indicating a certain level of classification error during the tenth training session. Furthermore, between the 16th and 17th training sessions, the Cls Loss remained consistently high, suggesting that the model faced challenges in reducing classification errors during that period. The error highlights areas where further refinement and optimization are needed to improve the model’s classification performance, as shown in Figure 4b.

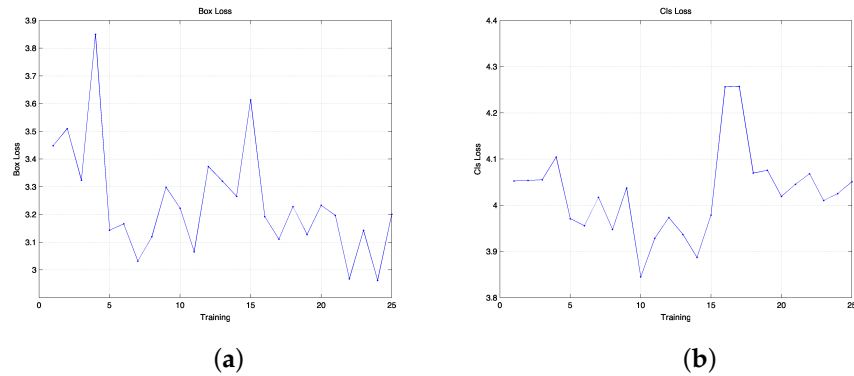


Figure 4. (a) illustrates the fluctuations in Box Loss over different training sessions, with the highest peak recorded at 3.85 during the fourth session and the lowest point observed in session 24. (b) demonstrates the trend of Cls Loss across the training sessions. A noticeable peak of 3.85 occurs during the tenth session, and the Cls Loss remains consistently high between the 16th and 17th sessions.

During the training sessions, the Dal Loss exhibited an irregular pattern on its path to stabilization. The model faced fluctuations indicated in this crucial metric, affecting consistency and progress toward achieving more stable and reliable performance, as shown in Figure 5a.

In training session 15, the mAP50(B) reached its peak, which means that during this session, the mean Average Precision at 50% overlap threshold for the B category (mAP50(B)) achieved its highest value. This metric indicates that the model performed its best in precision for object detection tasks involving category B at this particular point in the training process. The model’s performance, as measured by the mean Average Precision at thresholds ranging from 50% to 95% for the B category (mAP50-95(B)), was inconsistent or fluctuating until it eventually reached its highest value during the 15th training session, as shown in Figure 5b.

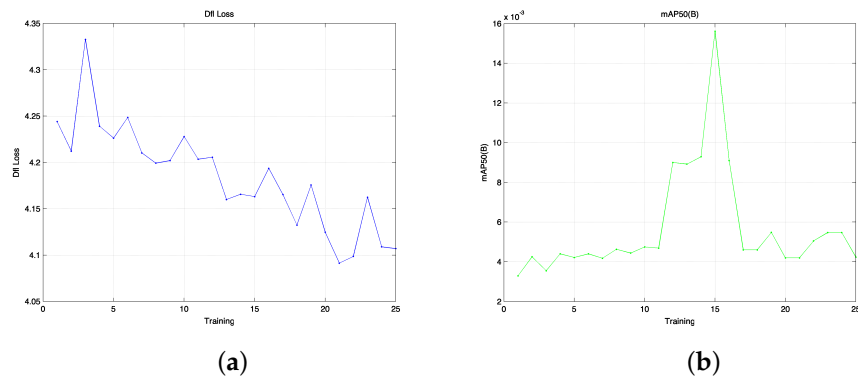


Figure 5. (a) shows the Dal Loss throughout various training sessions, exhibiting an irregular pattern as it moves towards stabilization. This fluctuation indicates challenges in achieving consistent and stable performance. (b) displays the model’s performance based on the mean Average Precision at thresholds ranging from 50% to 95% for the B category (mAP50-95(B)). The performance was inconsistent or fluctuating until it reached its peak during the 15th training session.

The mAP50-95(B) exhibited unstable behavior until it peaked during training session 15. The performance of the model, as measured by the mean Average Precision at thresholds ranging from 50% to 95% for the B category (mAP50-95(B)), was inconsistent or fluctuating until it eventually reached its highest value during the 15th training session, as shown in Figure 6a. The ValBoxLoss gradually decreased until training session 25, reaching a value of 3.8. The validation box loss, a metric used to evaluate the performance of the model's bounding box predictions, progressively declined and reached a value of 3.8 by the 25th training session, as shown in Figure 6b.

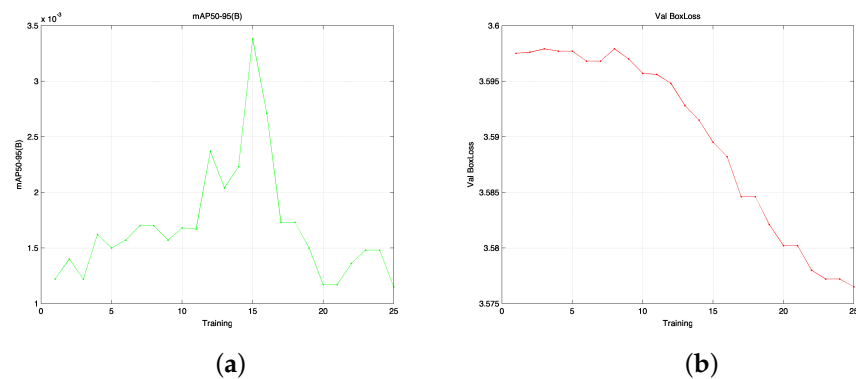


Figure 6. (a) shows that the mAP50-95(B) performance metric was unstable until it reached its peak during the 15th training session. (b) displays that the ValBoxLoss metric gradually decreased and reached a value of 3.8 by the 25th training session.

ClsLoss exhibited irregular behavior until the 16th training session, eventually showing an upward trend by the 25th. The classification loss (ClsLoss), which measures the model's classification predictions' performance, showed inconsistent values until the 16th training session. However, by the 25th training session, it began to show a rising trend, as shown in Figure 7a.

DflLoss decreased until the 25th training session. The distribution focal loss (DflLoss), which assesses a model's performance, decreased its value until the 25th training session, indicating improvement in the model's performance over those training sessions, as shown in Figure 7b.

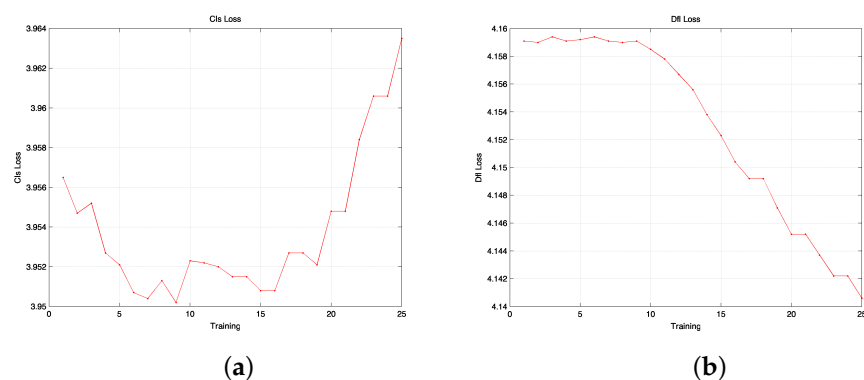


Figure 7. (a) shows that ClsLoss exhibited irregular behavior until the 16th training session, eventually showing an upward trend by the 25th session. (b) displays that DflLoss decreased until the 25th training session, indicating an improvement in the model's performance.

Lr/pg0 exhibited parabolic behavior throughout the training process. The learning rate (Lr) or parameter group zero (pg0) displayed a trend or pattern that resembles a parabolic shape, typically indicating a rise and fall or vice versa, during the training process, as shown in Figure 8a.

The learning rate or parameter group one ($lr/pg1$) gradually decreases during training. The learning rate or parameters associated with group one are progressively reduced during training. As shown in Figure 8b, to fine-tune the model and improve performance by making more minor adjustments as training progresses.

The validation figure displays a split within the part box, highlighting COS and DAL losses. This division signifies their significance and impact, suggesting the need for further analysis and evaluation as shown in Figure 9.

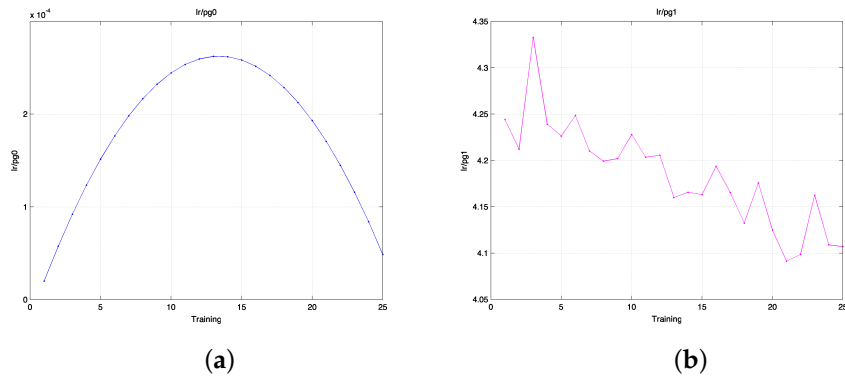


Figure 8. (a) The learning rate or parameter group one ($lr/pg1$) gradually decreases during training, enabling finer model adjustments to improve performance. (b) The validation figure shows a split within the part box, highlighting the significance and impact of COS and DAL losses, suggesting further analysis and evaluation.

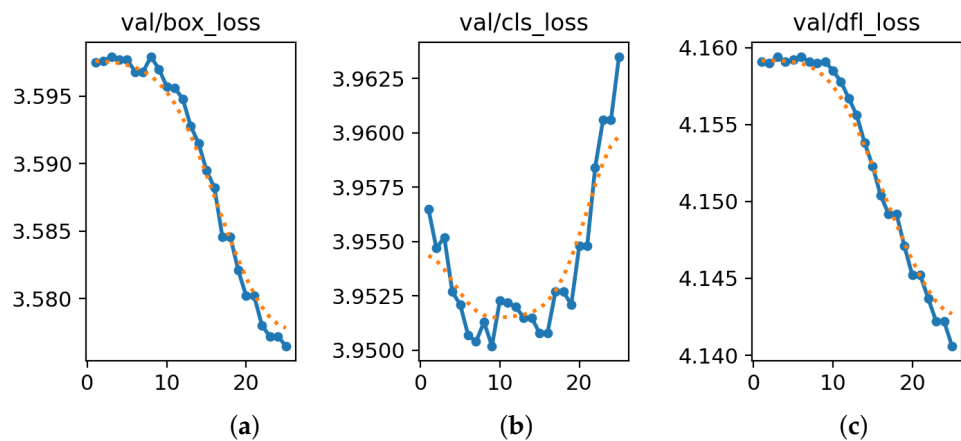


Figure 9. (a) The learning rate or parameter group zero ($lr/pg0$) exhibited a parabolic behavior throughout the training process, indicating a rise and fall pattern. (b) The learning rate or parameter group one ($lr/pg1$) gradually decreased during training to fine-tune the model and improve performance through minor adjustments. (c) The validation figure displays a split within the part box, highlighting the significance of COS and DAL losses and suggesting the need for further analysis and evaluation.

The precision figure represents the accuracy and exactness of the analyzed dataset. It showcases the level of detail and correctness in the results, indicating the reliability and quality of the information presented.

The figure reminds us of the mean Average Precision at 50 (mAP50). It is a benchmark for evaluating the accuracy and performance of the model’s object detection or recognition capabilities. By referencing mAP50, the figure highlights the model’s ability to correctly identify and locate objects with a certain level of confidence within the given dataset, as shown in Figure 10.

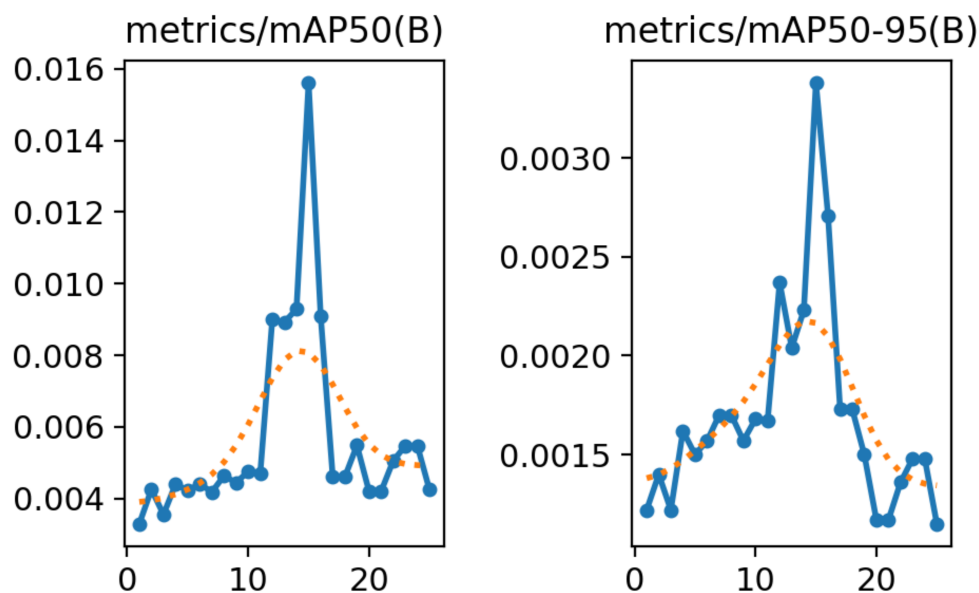


Figure 10. The figure illustrates the model's object detection performance evaluated by mean Average Precision at 50 (mAP50), showcasing its accuracy and ability to identify and locate objects within the dataset.

5. Discussions

The findings from this research highlight several important aspects regarding the use of convolutional neural networks (CNNs) for detecting weapons in images, especially those involving children and adolescents. First and foremost, the overall accuracy and reliability of the detection system are promising, with a precision of 86% and a recall of 14.2%. These results, though varied across different subsets, indicate that the CNN model is capable of discerning weapons with a notable level of accuracy, which is crucial for effective monitoring and prevention strategies. One point of discussion revolves around the complexity of the dataset used. The images contain various weapon types, sizes, orientations, and contextual backgrounds, which pose significant challenges for accurate detection. Despite these complexities, the CNN model achieved substantial accuracy, demonstrating its robustness and adaptability. This suggests that further refinement and expansion of the dataset could enhance the model's performance even more. Another discussion point pertains to the model's computational efficiency. Completing training in just 27 min on a specific computer setup and achieving such results within 25 epochs underscores the efficiency of the convolutional neural network. This rapid training and execution process is advantageous for practical applications where time and computational resources may be limited. Furthermore, the project underscores the importance of leveraging advanced machine learning techniques for social good. The research successfully detects weapons in images shared on social networks, providing a valuable tool for law enforcement and social service agencies. This can help in early intervention, preventing youth exposure to firearms and potentially reducing the risk of violence. The ethical aspect of this research, focusing on social good, is a significant highlight. However, the precision and recall rates indicate room for improvement. While precision is relatively high, the recall rate suggests that not all instances of weapons are being detected. It calls for further investigation into optimizing the model parameters, refining the dataset, and possibly integrating additional features or techniques to enhance detection capabilities. Lastly, this research opens avenues for exploring the application of CNNs to other types of harmful content on social networks, such as cyberbullying or drug-related imagery. The methodology and findings can be a foundation for developing comprehensive monitoring systems to safeguard youth from

online dangers. In conclusion, while the study presents promising results and practical implications, ongoing refinement and exploration are essential. The ultimate goal is to develop a more comprehensive and robust system to effectively support crime prevention efforts, protect youth, and contribute to safer digital environments. The potential of CNNs for improving safety and monitoring strategies is a crucial takeaway from this research. In future work, a more diverse dataset should incorporate multimodal detection approaches (e.g., combining text and image analysis) to capture a broader range of potential threats.

6. Conclusions

The development of this project aims to support the monitoring and detection of cyberbullying, with a particular focus on identifying weapons in images to prevent children and young people from being harmed, harassed, or exposed to violent content online. Our findings show a 10.71% accuracy within the test dataset for weapon detection in the corpus of images considered. For the test subset classification, we achieved 71.42%, and for validation, we obtained a 17.86% classification rate. The training set achieved higher accuracy percentages, providing a significant level of reliability in detecting weapons in the image corpus. During the evaluation, we achieved 86% precision and 14.2% recall for weapon recognition in images featuring children and adolescents. This performance was completed within 27 min on a computer model 1510M with 13 s at 8.1 GLFOPS. The training involved 25 epochs, reflecting a relatively fast training process compared to more sophisticated models of similar complexity. The convolutional neural network model included 225 layers and employed a learning rate of 0.01 through a gradient optimizer, managing around 3,011,043 parameters and 3,011,027 gradients. These results demonstrate the model's robustness and efficiency in detecting weapons in images, contributing to safer online environments for young people.

Author Contributions: L.I.B.-S.: software, validation, formal analysis, investigation, writing—original draft, writing—review and editing, visualization. B.P.G.-V.: writing—review and editing, conceptualization, methodology, software. M.T.O.-A.: methodology, software, validation, writing—review and editing, visualization. and L.F.-P.: validation, writing—original draft, writing—review and editing, resources, project administration, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are available from the corresponding authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Milosevic, T.; Van Royen, K.; Davis, B. Artificial intelligence to address cyberbullying, harassment and abuse: New directions in the midst of complexity. *Int. J. Bullying Prev.* **2022**, *4*, 1–5. [[CrossRef](#)] [[PubMed](#)]
2. Hani, J.; Nashaat, M.; Ahmed, M.; Emad, Z.; Amer, E.; Mohammed, A. Social media cyberbullying detection using machine learning. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 703–707. [[CrossRef](#)]
3. Pericherla, S.; Ilavarasan, E. Overcoming the Challenge of Cyberbullying Detection in Images: A Deep Learning Approach with Image Captioning and OCR Integration. *Int. J. Comput. Digit. Syst.* **2024**, *15*, 393–401. [[CrossRef](#)] [[PubMed](#)]
4. Vishwamitra, N.; Hu, H.; Luo, F.; Cheng, L. Towards understanding and detecting cyberbullying in real-world images. In Proceedings of the 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), Virtual, 14–17 December 2021.

5. Yuvaraj, N.; Chang, V.; Gobinathan, B.; Pinagapani, A.; Kannan, S.; Dhiman, G.; Rajan, A.R. Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification. *Comput. Electr. Eng.* **2021**, *92*, 107186. [[CrossRef](#)]
6. Cheng, L.; Guo, R.; Silva, Y.N.; Hall, D.; Liu, H. Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. *ACM/IMS Trans. Data Sci.* **2021**, *2*, 1–23. [[CrossRef](#)]
7. Orrù, G.; Galli, A.; Gattulli, V.; Gravina, M.; Micheletto, M.; Marrone, S.; Nocerino, W.; Procaccino, A.; Terrone, G.; Curtotti, D.; et al. Development of Technologies for the Detection of (Cyber) Bullying Actions: The BullyBuster Project. *Information* **2023**, *14*, 430. [[CrossRef](#)]
8. Ige, T.; Adewale, S. AI powered anti-cyber bullying system using machine learning algorithm of multinomial naïve Bayes and optimized linear support vector machine. *arXiv* **2022**, arXiv:2207.11897. [[CrossRef](#)]
9. Verma, K.; Davis, B.; Milosevic, T. Examining the Effectiveness of Artificial Intelligence-Based Cyberbullying Moderation on Online Platforms: Transparency Implications. In Proceedings of the 23rd Annual Conference of the Association of Internet Researchers, Dublin, Ireland, 2–5 November 2022.
10. Simpson, J.D. Applications of Artificial Intelligence and Graphy Theory to Cyberbullying. Graduate Thesis, Missouri State University, Springfield, MO, USA, 2020.
11. Salawu, S.; He, Y.; Lumsden, J. Approaches to automated detection of cyberbullying: A survey. *IEEE Trans. Affect. Comput.* **2017**, *11*, 3–24. [[CrossRef](#)]
12. Biswas, R.; Ganguly, K.; Das, A.; Saha, D. Securing Social Spaces: Harnessing Deep Learning to Eradicate Cyberbullying. *arXiv* **2024**, arXiv:2404.03686.
13. Almomani, A.; Nahar, K.; Alauthman, M.; Al-Betar, M.A.; Yaseen, Q.; Gupta, B.B. Image cyberbullying detection and recognition using transfer deep machine learning. *Int. J. Cogn. Comput. Eng.* **2024**, *5*, 14–26. [[CrossRef](#)]
14. Kumar, A.; Sachdeva, N. Multimodal Cyberbullying Detection Using Capsule Network with Dynamic Routing and Deep Convolutional Neural Network. *Multimed. Syst.* **2022**, *28*, 2043–2052. [[CrossRef](#)]
15. Sultan, T.; Jahan, N.; Basak, R.; Jony, M.S.A.; Nabil, R.H. Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition. *Int. J. Intell. Syst. Appl.* **2023**, *15*, 1. [[CrossRef](#)]
16. Kargutkar, S.M.; Chitre, V. A study of cyberbullying detection using machine learning techniques. In Proceedings of the 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 11–13 March 2020; pp. 734–739.
17. Shrimali, S. A Natural Language Processing and Machine Learning-Based Framework to Automatically Identify Cyberbullying and Hate Speech in Real-Time. In Proceedings of the 2022 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA, 30 September–2 October 2022; pp. 1–5.
18. Dani, H.; Li, J.; Liu, H. Sentiment informed cyberbullying detection in social media. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, 18–22 September 2017; pp. 52–67.
19. Sasikumar, K.; Nambiar, R.K.; Rohith, K. Unmasking Cyberbullies on Social Media Platforms Using Machine Learning. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; pp. 1–7.
20. Soni, D.; Singh, V. Time reveals all wounds: Modeling temporal characteristics of cyberbullying. In Proceedings of the International AAAI Conference on Web and Social Media, Palo Alto, CA, USA, 25–28 June 2018; Volume 12.
21. Islam, M.R.; Bataineh, A.S.; Zulkernine, M. Detection of Cyberbullying in Social Media Texts Using Explainable Artificial Intelligence. In Proceedings of the International Conference on Ubiquitous Security, Zhangjiajie, China, 28–31 December 2023; pp. 319–334.
22. Chandra, N.; Khatri, S.K.; Som, S. Cyberbullying detection using recursive neural network through offline repository. In Proceedings of the 2018 7th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 29–31 August 2018; pp. 748–754.
23. Nandhini, B.S.; Sheeba, J. Cyberbullying detection and classification using information retrieval algorithm. In Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015), Unnao, India, 6–7 March 2015; pp. 1–5.
24. Zhang, X.; Tong, J.; Vishwamitra, N.; Whittaker, E.; Mazer, J.P.; Kowalski, R.; Dillon, E. Cyberbullying detection with a pronunciation-based convolutional neural network. In Proceedings of the 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), Anaheim, CA, USA, 18–20 December 2016; pp. 740–745.
25. Vijayakumar, V.; Prasad, D.D.H.; Adolf, P. Multimodal cyberbullying detection using hybrid deep learning algorithms. *Int. J. Appl. Eng. Res.* **2021**, *16*, 568–574.
26. Dewani, A.; Memon, M.A.; Bhatti, S. Cyberbullying detection: Advanced preprocessing techniques and deep learning architecture for Roman Urdu data. *J. Big Data* **2021**, *8*, 160. [[CrossRef](#)] [[PubMed](#)]

27. Raj, M.; Singh, S.; Solanki, K.; Selvanambi, R. An application to detect cyberbullying using machine learning and deep learning techniques. *SN Comput. Sci.* **2022**, *3*, 401. [[CrossRef](#)] [[PubMed](#)]
28. Aldhyani, T.H.; Al-Adhaileh, M.H.; Alsubari, S.N. Cyberbullying identification system based deep learning algorithms. *Electronics* **2022**, *11*, 3273. [[CrossRef](#)]
29. Khafajeh, H. Cyberbullying Detection in Social Networks Using Deep Learning. *Int. Arab J. Inf. Technol. (IAJIT)* **2024**, *21*, 245–255. [[CrossRef](#)]
30. Baiganova, A.; Toxanova, S.; Yerekeshova, M.; Nauryzova, N.; Zhumagalieva, Z.; Tulendi, A. Hybrid Convolutional Recurrent Neural Network for Cyberbullying Detection on Textual Data. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 834–842. [[CrossRef](#)]
31. Gutiérrez-Batista, K.; Gómez-Sánchez, J.; Fernandez-Basso, C. Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model. *Soc. Netw. Anal. Min.* **2024**, *14*, 136. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.