Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 40, January 2025 – www.caeaccess.org

CBDS-ConvNet: A Cyber-Bullying Detection Model using Convolutional Neural Network

Ayodeji O. Akinwumi
Department of Computer Science, Faculty of Computing, Adekunle Ajasin University, P.M.B 001, Akungba, Nigeria

Ayokunle O. Ige
Department of Computer Science, Faculty of Computing, Adekunle Ajasin University, P.M.B 001, Akungba, Nigeria

Joy R. Obafemi
Department of Computer Science, Faculty of Computing, Adekunle Ajasin University, P.M.B 001, Akungba, Nigeria

Olatunde D. Akinrolabu
Department of Computer Science, Faculty of Computing, Adekunle Ajasin University, P.M.B 001, Akungba, Nigeria

Bolu O. Akingbesote
Department of Computer Science, Faculty of Computing, Adekunle Ajasin University, P.M.B 001, Akungba, Nigeria

ABSTRACT
In recent years, the increasing reliance on the internet and the integration of social media into daily life have led to significant advancements in various aspects of human activities. However, these developments have also facilitated unethical behaviors, with cyberbullying emerging as a critical concern. Traditional machine learning models for cyberbullying detection face challenges such as vulnerabilities to adversarial attacks and difficulty capturing nuanced or complex contextual information, often resulting in misclassifications. To address these limitations, this research introduces **CBDS-ConvNet**, a Convolutional Neural Network-based model designed for real time cyberbullying detection and prevention. The model is structured into five key layers: Data Collection, Data Preprocessing, Training, Cyberbullying Detection, and Performance Evaluation. Data from platforms such as Mendeley, Kaggle, and GitHub were utilized, with preprocessing ensuring the text data was clean and suitable for training. The model achieved an accuracy of 77.65%, precision of 56.26%, recall of 63.86%, and an F1 score of 60.20%, outperforming some other machine learning approaches. To further evaluate the robustness of the developed model, it was tested on a synthesized dataset, achieving an accuracy of 91%, precision of 89%, recall of 81%, and an F1 score of 85%. This research shows the capacity of CNNs in tackling the dynamic and complex nature of social media interactions. By enabling real-time cyberbullying detection, the CBD-ConvNet system provides a robust framework for safer online environments, thereby advancing research efforts in the field of cyberbullying prevention.

General Terms
Machine Learning, Deep Learning, Cybersecurity, Pattern Recognition, Social Media Analysis.

Keywords
Cyberbullying Detection, Convolutional Neural Networks (CNN), Real-time Detection, Text Classification, Online Safety.

1. INTRODUCTION
Over the years, internet usage has experienced a swift and widespread increase in almost every country globally, leading to notable improvements in various human activities [1] [2] [3] [4]. According to [5], the number of internet users in the world is over 5 billion people which is over 64% of the world's population. It can then be deduced that the Internet has become an essential part of people's life today. With the continued rise in internet usage, social media has also become an integral part of our daily lives. Social media, with over 3 billion worldwide, has revolutionized communication by providing the capacity to interact and exchange information with everyone at any time [6][7][8]. Recognized as one of the most impactful advancements of the 21st century [9], social networks enable simultaneous engagement with a large number of people, shaping the way people connect and share information.

Many human activities, including education, business, entertainment, and governance, have been incorporated into social media networks. They have become immensely popular as several millions users have used them as either communication tools or as real-time, dynamic data sources [10]. However, despite the advantages of the social media, it has its difficulties and challenges. There are a lot of illegal and unethical activities being orchestrated within the infrastructure of various social media platforms. Cybercriminals now utilize these platforms in committing different types of cybercrimes such as phishing, spamming, hacking, and in particular cyberbullying.

According to [11] [12] [13], Cyberbullying is the process of using the internet, cell phones or other devices to send or post text or images intended to hurt or embarrass another person. It is also the use of technology to bully someone or a group of people, which includes threats, abusive words, sexual remarks, hate speech etc. [14] [15]. Additionally, cyberbullying is formally characterized as the consistent transmission of hostile or aggressive messages by individuals or groups, aiming to cause harm, embarrassment, or distress to others [16]. Bullying is a forceful action which is done in order to physically or mentally hurt someone or a group of people over a repeated period of time [17]. Bullying has existed in society throughout history; however, the emergence of the internet has provided bullies with a new and opportunistic tool. [18]. Cyberbullying has increasingly become a crime being perpetuated on the internet and in other digital spaces, particularly on social media sites and the anonymity that the internet can provide also gives penetrators additional incentive to get involved in cyberbullying [19]. The prevalence of cyberbullying has also escalated due to this increased accessibility of technology and internet services, particularly since the COVID-19 pandemic [20], and given the reliance of youths on digital platforms, an increase in cyberbully incidents seems unavoidable. Instances

of cyberbullying include rumors posted on social media or sent by e-mail; embarrassing videos or pictures; insulting, intimidating and abusive messages posted on social networks among others. Cyberbullying have devastating effects on its victims such as depression, sadness, lack of sleep, fear, low self-esteem, and anxiety and in extreme cases, the victims may commit suicide [17].

There have been several instances of cyberbullying all around the world, from both the more developed countries to the developing countries. For instance, [21] reported the case of three Canadian children who committed suicide after being taunted by others online. Another 13-year-old female was also bullied online to the point that she hung herself in her closet [22]. The situation has not been different in Nigeria, as cyberbullying is becoming more prevalent in with the increasing use of computers and the internet [11]. According to a research carried out at a Nigerian university, approximately 50% of the students have been victims of one form of cyberbullying or the other [23]. Furthermore, a range of between 48% - 57% undergraduates have been bullied through the various existing social media platforms. With these rise of cyberbullying in online social interactions, particularly among the youth [24], coupled with the devastating consequences it has on its victims, it is crucial to urgently find solutions to detect its occurrence in real time and hence to prevent it. It is important to note that while there are existing research studies on the effects and prevalence of cyberbullying, there is a notable gap in the practical application of monitoring social networks to identify and address cyberbullying activities [25]. Therefore, one way for researchers to close this gap is to use machine learning to detect and classify offensive language in text [26] [27]. This approach thus motivated the researchers to adopt the concept of convolutional neural network which is a class of machine learning for the detection of cyberbullying.

Machine learning is a rapidly expanding technology that enables computers to automatically learn from previously known data. It is a method that uses raw data to create a model that automatically classifies appropriate actions. [18]. Several machine learning algorithms, including Decision Tree, Random Forest, and Support Vector Machine, among others, have been used for cyberbullying detection. However, these machine learning models are vulnerable to adversarial attacks where small, imperceptible perturbations to the input data can cause the model to misclassify instances. Also, traditional machine learning models often struggle to capture nuanced contextual information, leading to false positives or false negatives. Recently, there has been the emergence of deep learning algorithms which are being used effectively to solve classification and pattern recognition problems [28] [29] [30] [31]. When applied, they have produced accurate results comparable to and in some cases surpassing human expert performance [29]. Examples of such a deep learning algorithm is Convolutional Neural Network (CNN). CNN has been extensively used primarily in the field of computer vision, such as image classification [32] [33] , object detection [34], speech recognition [35], vehicle recognition [36], facial expression recognition [37] [38] among many others. While the performance of CNN in terms of accuracy and efficiency in the aforementioned areas has achieved great success, however, in the area of cyberbullying, its performance has not been fully exploited. This research is aimed at making a contribution in this area. Thus, the aim of this study is to apply Convolutional Neural Network to develop a CNN-based Cyberbullying Detection System (CBDS-ConvNet) for detecting and preventing cyberbullying. In this study, Twitter was chosen as the case study due to its substantial daily data generation and its increasing recognition as a platform susceptible to cyberbullying attacks over the years.

2. RELATED WORKS

Among the several existing social media networks, Twitter (now known as X) currently ranks as one of the leading platforms [39][40]. It is characterized by its short message limit (now 280 characters) and unfiltered feed, with an average of 500 million tweets posted per day [41], and as reported by [42], there are over 192 million daily active users on Twitter; these figures show how popular Twitter is today. Twitter is also listed as one of the top five social media platforms where users experience cyberbullying [43]. Given the devastating effects of cyberbullying on victims, it is vital to develop effective methods for detecting and hence preventing it. There have been some measures that have been taken by Twitter to mitigate the issue of cyberbullying, some of which include filtering inappropriate messages from people without a profile picture, banning individuals who use abusive language, among others. Also, the Nigerian government expressly criminalized cyberbullying in the Cybercrime (Prohibition, Prevention, etc.) Act of 2015 [11]. These measures have performed to a degree of success, but still, cyberbullying perpetration on the platform has not decreased, and the challenge of detecting cyberbullying in real time still persists [44]. This may be due to the massive amount of users Twitter has, and to manually identify bullying messages over the huge network is difficult. There should be an automated system where such bullying messages can be detected automatically and in real time, thereby taking appropriate action.

Chatzakou et al. [45] cited many obstacles to detecting cyberbullying, which includes user heterogeneity, the ephemeral nature of the problem, the anonymity provided by social media, and numerous bullying forms other than harsh words. In their work, the authors took into account user, textual, and network characteristics to detect cyberbullying. The text was classified as bully or non-bully using supervised machine learning methods. According to the authors in [46], the use of harsh language has grown in recent years. A dataset was obtained from comments on Yahoo News Platform and Finance in order to develop a supervised classification approach. For classification, a framework called Vowpal Wabbit was utilized, and a supervised classification algorithm based on Natural Language Processing (NLP) features was developed. The developed model achieved an F-Score of 0.817. In [47], different algorithms were compared for the detection of cyberbullying on social media. Support Vector Machine (SVM), Naive Bayes, Random Forest, and K-Nearest Neighbor, were some of the algorithms compared. According to the authors, the SVM algorithm proved to be the best of all based on parameters such as accuracy, precision, and recall.

Zhao et al. [48] proposed a cyberbullying detection framework in which word embedding was utilized to create a list of insulting terms and assign weights to obtain bullying features. Support Vector Machines (SVM) were used as the main classifier with a recall of 79.4 percent. Also, Chen et al. [49] suggested a new approach called Lexical Syntactic Feature, in which SVM was employed as the classifier, and the developed model obtained 77.9% precision and 77.8% recall. Another method was presented by [50] based on two classifiers: Naive Bayes and SVM. The dataset used was obtained from Kaggle. The results revealed that the Naive Bayes classifier had

an average accuracy of 92.81% while the SVM with a poly kernel had an accuracy of 97.11%. The authors however, did not mention the training or testing size of the dataset used.

Another study [51] proposed a prototype system for monitoring social networking sites and detecting bullying incidents. The method used was to collect and store bullying words in a database, then use Twitter API to capture tweets and match their content to the bullying content already captured. If a bullying incident is found, an email will be sent to the police department in charge of internet offenses. However, this model is yet to be implemented. The authors in [52] proposed utilizing supervised machine learning to detect cyberbullying on Twitter. The study used two distinct feature extraction approaches with a variety of machine learning algorithms. The Sequential Minimal Optimization (SMO) classifier achieved the greatest accuracy (68.47 percent) among the other machine learning algorithms. Al-garadi et al. [53] used CNN to identify cyberbullying on Twitter. Human intelligence was used to label the training data, and the GloVe approach was used to construct word embeddings for each word. The set of word embeddings that was produced was then fed into the CNN algorithm for classification. This approach offers several advantages, including the removal of feature extraction and selection, as well as excellent accuracy.

Given the existing body of research, this study seeks to contribute to the ongoing efforts in enhancing cyberbullying detection. The approach involves the utilization of convolutional neural networks (CNNs) to identify instances of cyberbullying. Additionally, the study aims to elevate model performance and enhance classification outcomes by systematically optimizing various parameters within the convolution layers.

3. METHODOLOGY

A Convolutional Neural Network (CNN)-based Cyberbullying Detection System (CBDS-ConvNet) is proposed for the purpose of detecting and preventing cyberbullying in real-time. While CNNs have traditionally been utilized primarily in image classification tasks, recent research has demonstrated their versatility and effectiveness in a wide range of text mining applications. Numerous studies, including those found in [54][55][56][57][58][59][60], among others, have successfully employed CNNs for various tasks such as sentiment analysis, text classification, and entity recognition. Their work shows the remarkable potential of CNNs to capture hierarchical and contextual patterns in textual data, making it a powerful tool for the identification of harmful behaviors like cyberbullying within online platforms. The proposed CBDS-ConvNet leverages these capabilities to improve the accuracy and efficiency of cyberbullying detection, thereby contributing to the development of safer online environments.

The architecture of the proposed system is in five layers as depicted in Figure 1. They are the Data Collection layer, the Data Preprocessing Layer, the Training Layer, the Cyberbullying Detection Layer, and the Performance Evaluation layer.

3.1 Data Collection

The significance of data sourcing in this research was paramount, as the quality and diversity of the data play a critical role in the performance of the detection system. In line with this, tweets were extracted from three distinct and reputable sources to ensure a comprehensive dataset containing both bullying and non-bullying texts. The data sources are:

1. https://data.mendeley.com/datasets/jf4pzyvnpj/1
2. https://www.kaggle.com/datasets/dataturks/dataset-for-detection-of-cybertrolls
3. https://github.com/dhavalpotdar/detecting-offensive-language-in-tweets

3.2 Data Preprocessing

A comprehensive data cleaning process was initiated to eliminate noise, irrelevant content, and duplicate tweets. The preprocessing steps include:

1. **Tokenization:** which involves breaking down sentences or paragraphs into smaller units called tokens, resulting in a list of separated words.
2. **Lowercasing Text:** This entails converting the list of words obtained from tokenization to lowercase, ensuring uniformity in text representation.
3. **Punctuation Marks / Special Characters / Stop Words Removal:** Removing punctuation marks, special characters, and stop words to enhance the meaningfulness of the data for classifiers.
4. **Word Embedding:** Representing each word as a real value vector through word embedding, essential for the CNN model, which typically processes data in image format. This involved transforming the message feed data into a numerical format, with each word represented by a real-value vector.

Additionally, considering the data originated from three distinct sources, a meticulous data cleaning process was imperative to merge the datasets. This included label standardization, where labels such as *"Non-offensive"* and *"Offensive"* from the GitHub source were harmonized to *"Bullying"* and *"Not Bullying"*.

To ensure data integrity and eliminate redundancy, potential duplicate texts were expunged from the consolidated dataset. Target refinement followed, with the conversion of labels to binary form—designating *"bullying"* as 1 and *"Not Bullying"* as 0—enabling a clear distinction for subsequent analyses. Subsequent text preprocessing involved a multifaceted approach encompassing the removal of stop words, usernames, email addresses, URLs, punctuation marks, and redundant whitespaces. This thorough cleansing aimed at enhancing the quality and relevance of textual content for subsequent analysis. Tokenization, a pivotal step, was employed to break down sentences into individual words and assign numerical IDs using a tokenizer. This transformative process facilitated subsequent neural network input, aligning with the exigencies of text classification and sentiment analysis tasks.

To ensure uniformity in sequence lengths, the tokenized sequences underwent padding to a standardized length of sixty (60) as shown in Figure 2. This decision was informed by a meticulous distribution plot analysis of sequence lengths, ensuring an optimal balance between information retention and

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 40, January 2025 – www.caeaccess.org

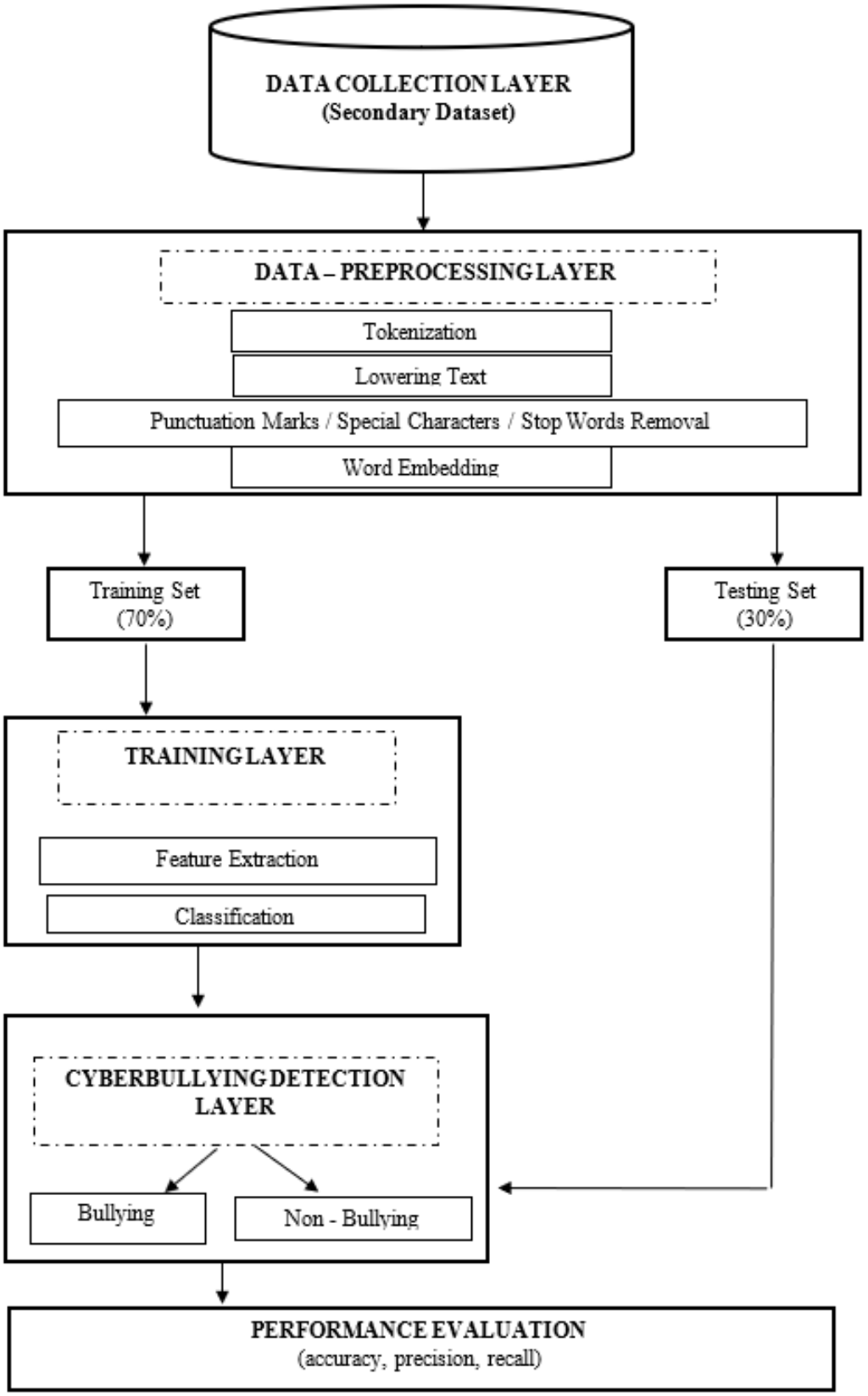

Figure 1: System Design of the CNN based Cyberbullying Detection Model (CBDS-ConvNet)

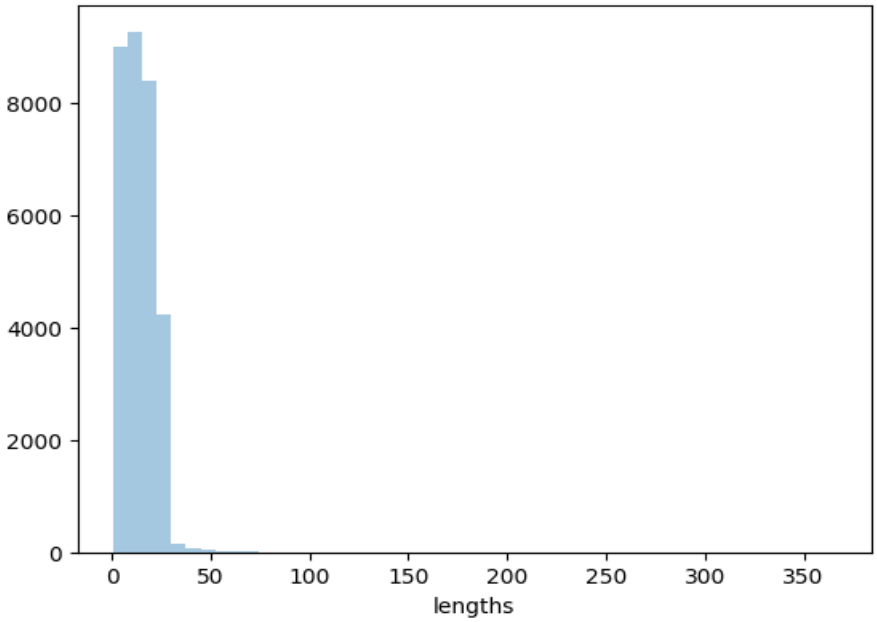

Figure 2: System Design of the CNN based Cyberbullying Detection Model (CBDS-ConvNet)

computational efficiency. Furthermore, the dataset was divided into three segments:

1. **Train Dataset:** 70% of the dataset was allocated for training the model. The objective is for the model to learn patterns and features present in the data.
2. **Validation Dataset:** 15% of the data was set aside to assess the model's performance during training without exposing it to the test data. This allows for the identification of potential overfitting and tuning of hyperparameters.
3. **Test Dataset:** Reserved for evaluating the final performance of the trained model which takes 15% of the dataset.

3.3 The Architecture of the Cyberbullying Detection System (CBDS-ConvNet)

In line with the layers of the CNN algorithm, the architecture of the model is carefully designed to process text data for cyberbullying detection, using each layer to extract different levels of features and patterns. The key layers in the architecture are as follows:

1. Embedding Layer: This layer transforms the textual data into numerical vectors, capturing semantic relationships between words.
2. Convolutional Layers: These layers employ convolution operations to detect spatial patterns and hierarchical features in the input data. Multiple convolutional layers allow the model to learn intricate representations.
3. Max Pooling Layers: Max Pooling is applied to downsample the spatial dimensions, focusing on the most relevant features and reducing computational complexity.
4. Dropout Layers: Integrated to mitigate overfitting, dropout layers introduce randomness by temporarily deactivating a proportion of neurons during training, preventing reliance on specific features.
5. Flatten Layer: Data is flattened into a one-dimensional vector, preparing it for input into the subsequent dense layers.
6. Dense Layers: Comprising two dense layers with 96 and 48 neurons, respectively, these layers enable the model to learn high-level abstractions. Rectified Linear Unit (ReLU) activation functions are applied to introduce non-linearity.
7. Additional Dropout Layers: Further dropout layers with different dropout probabilities (0.5 and 0.125) contribute to regularization, enhancing the model's generalization capabilities.
8. Output Layer: A single neuron with a sigmoid activation function is employed for binary classification, distinguishing between cyberbullying-related and non-related texts.

Key hyperparameters, including dropout rates (0.5 and 0.125) were carefully tuned to ensure robust generalization across every phase of the datasets.

3.4 Training Process

The training process utilized a stochastic gradient descent optimizer with binary cross-entropy as the loss function. The model was trained for 30 epochs with a batch size of 64. An early stopping criterion/function was applied to prevent overfitting while training.

3.5 Evaluation

The last phase of the workflow entails the following critical metrics:

- Accuracy: Measures overall classification correctness
- Precision, Recall, and F1-Score: Evaluate the balance between correctly predicted positive cases and the ability to detect cyberbullying content.

- ROC-AUC Score: Assesses the model's ability to distinguish classes

4. RESULTS AND EVALUATION

This section presents the results and evaluation of the proposed **CBD-ConvNet** system. The model was assessed using accuracy, precision, recall, and F1 score to evaluate its effectiveness in detecting cyberbullying.

4.1 Implementation Details

The model was trained and evaluated on a system with an Intel core i7 processor having a 2.5 GHz speed under 1 processor with 4 cores and 16 GB RAM with 6 MB L3 cache and 256 KB L2 cache. All the programming was with python using Keras. Scikit Learn and the Pandas library were used for data preprocessing and visualization. The model underwent training for 20 epochs, where an epoch represents one complete iteration through the entire training dataset. A batch size of 64 was used, indicating that the model updates its weights after processing 64 instances. These training specifics are essential in balancing the learning process, preventing overfitting, and achieving convergence.

The detailed explanation provided here aims to illuminate the choices made in constructing the model, offering insights into how it learns and generalizes from textual data.

4.1.1 Validation Data Accuracy across Epochs

The number of epochs was set to 20 to ensure the model had sufficient time to learn and stabilize while monitoring for overfitting. From the training and validation data, the performance across the 20 epochs shows the following trends:

1. **Epoch 1-5:** The model shows rapid progress during the early epochs, with the training loss significantly decreasing from 1.38 to approximately 1.21. Training accuracy rises from 53.8% to 72.8%, and recall also improves noticeably, indicating that the model is learning to identify relevant patterns in the data. Validation accuracy increases concurrently, reaching 80.4%, demonstrating strong early generalization.
2. **Epoch 6-10:** As training progresses, improvements in metrics continue at a slower pace. Training accuracy reaches 78.4% by epoch 10, and the F1 score shows consistent growth, highlighting the model's ability to balance precision and recall. Validation accuracy peaks at 81.1% during this phase, with stable F1 scores, reflecting the model's growing ability to refine its learning.
3. **Epoch 11-15:** During these epochs, training accuracy climbs steadily, reaching over 83.6%, with consistent improvements in precision, recall, and F1 score. Validation metrics maintain stability, with accuracy averaging above 80%, suggesting that the model continues to perform reliably on unseen data.
4. **Epoch 16-20:** In the final epochs, the model exhibits high training accuracy, reaching 88.9% by epoch 20, with strong F1 scores indicating balanced performance. Validation metrics remain stable, with an average accuracy of 80.7% and a corresponding F1 score of 59.9%. This stability demonstrates that the model has converged effectively without a significant decline in performance metrics.

4.2 Results

4.2.1 Results on Validation Data

The graph in Figure 3 tracks the accuracy trends for both the training and validation datasets across different epochs, showing the degree of correctness in its predictions over the training period. Figure 4 displays the loss trends for the training and validation datasets over epochs. Loss indicates the model's error, and this graph helps understand how well the model is minimizing its error during training. Lower loss values indicate improved model performance. Precision trends across epochs for different learning rates are depicted in Figure 5. Precision measures the accuracy of positive predictions, and this graph offers insights into the model's ability to make accurate positive classifications as training progresses. In Figure 6, the recall trends across epochs for different learning rates are presented. Recall measures the model's ability to capture instances of the positive class, providing information on how well the model identifies cyberbullying instances throughout the training process.

4.2.2 Performance Evaluation

To assess the effectiveness of the proposed CNN-based detection model, an experiment was undertaken to evaluate its performance using the test data. The performance metrics employed for the evaluation include Accuracy, Precision, Recall, and F1 Score.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1-Score = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad (4)$$

Where True Positive (TP) are instances correctly classified as "Bullying", True Negative (TN): are instances correctly classified as "Not Bullying", False Positive (FP) are instances incorrectly classified as "Bullying" when they are actually "Not Bullying" (Type I error) and False Negative (FN) are instances incorrectly classified as "Not Bullying" when they are actually "Bullying" (Type II error).

Additionally, a comprehensive evaluation of the classification performance of the model was conducted using a confusion matrix. The confusion matrix, depicted in Figure 7, provides a detailed breakdown of the model's predictions, including True Positives, True Negatives, False Positives, and False Negatives. The confusion matrix is a 2x2 table used to evaluate the performance of a classification model. It is composed of four key components, which are TP, TN, FP, FN. Based on the confusion matrix, the following values were derived:

1. *True Positives (TP) = 1559*
2. *True Negatives (TN) = 5718*
3. *False Positives (FP) = 1212*
4. *False Negatives (FN) = 882*

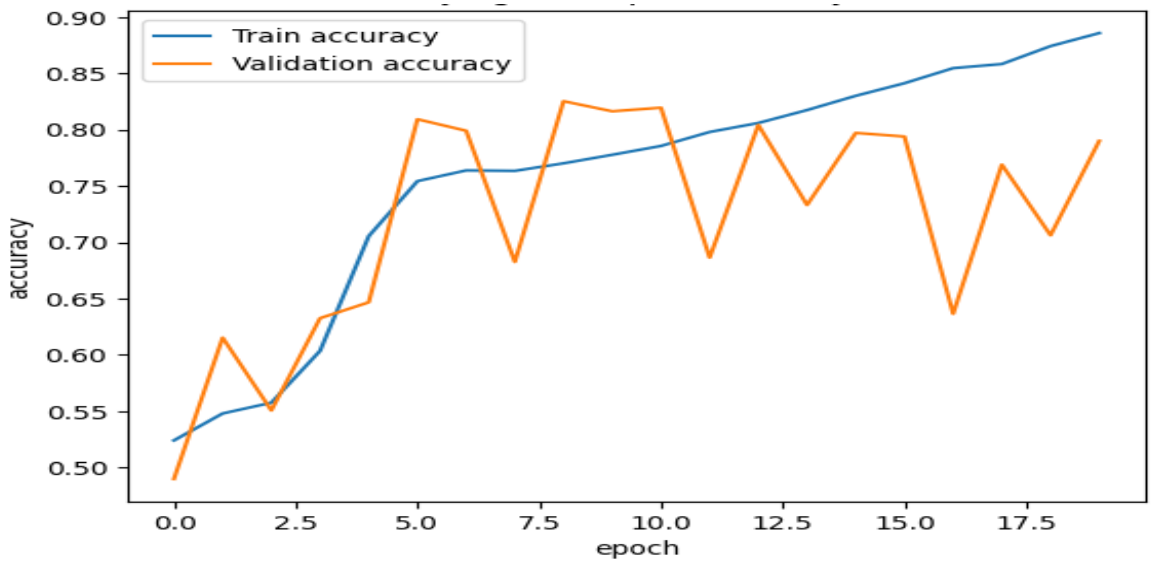

Figure 3: Accuracy vs. Epoch for Training and Validation Datasets

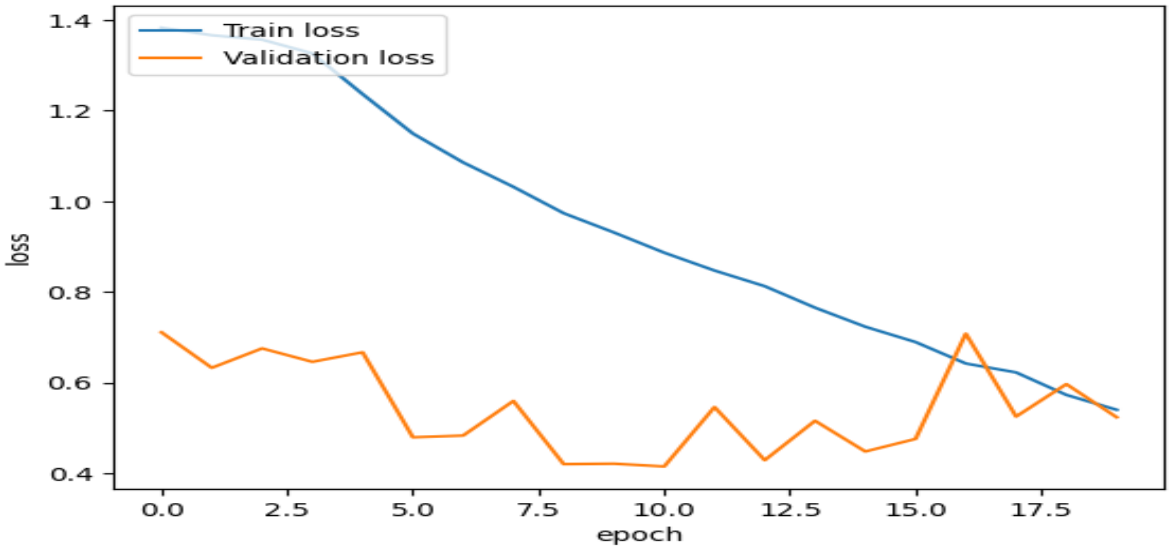

Figure 4: Loss vs. Epoch for Training and Validation Datasets

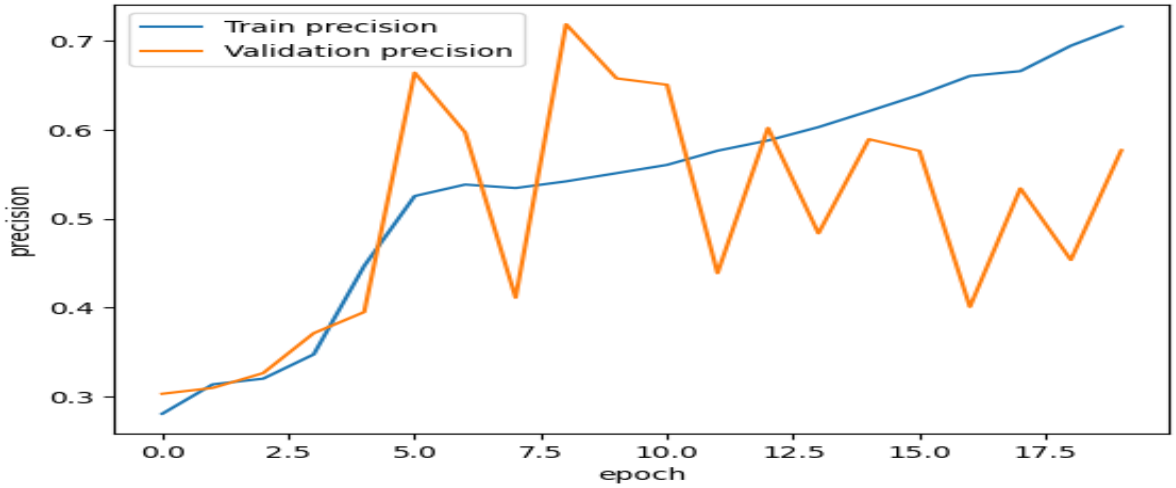

Figure 5: Precision vs. Epoch for Training and Validation Datasets

Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 7 – No. 40, January 2025 – www.caeaccess.org

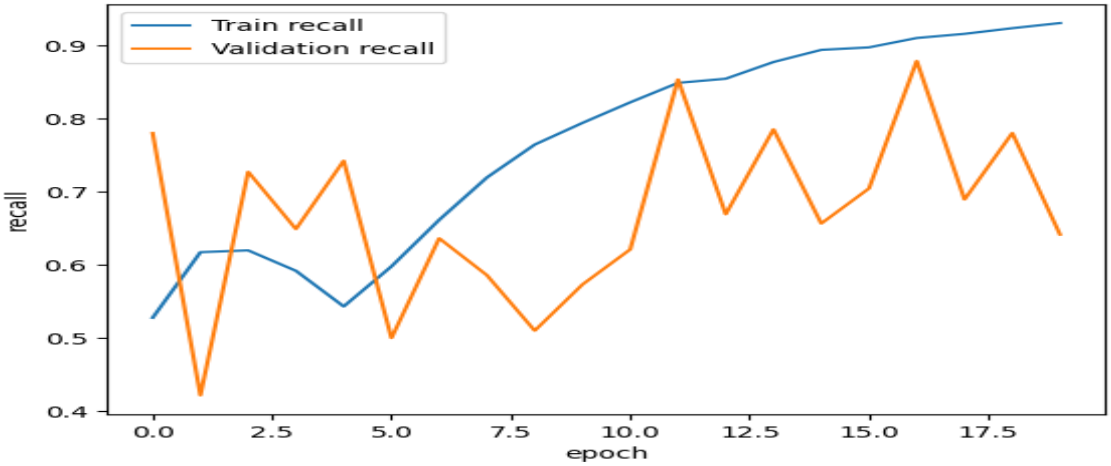

Figure 6: Recall vs. Epoch for Training and Validation Datasets

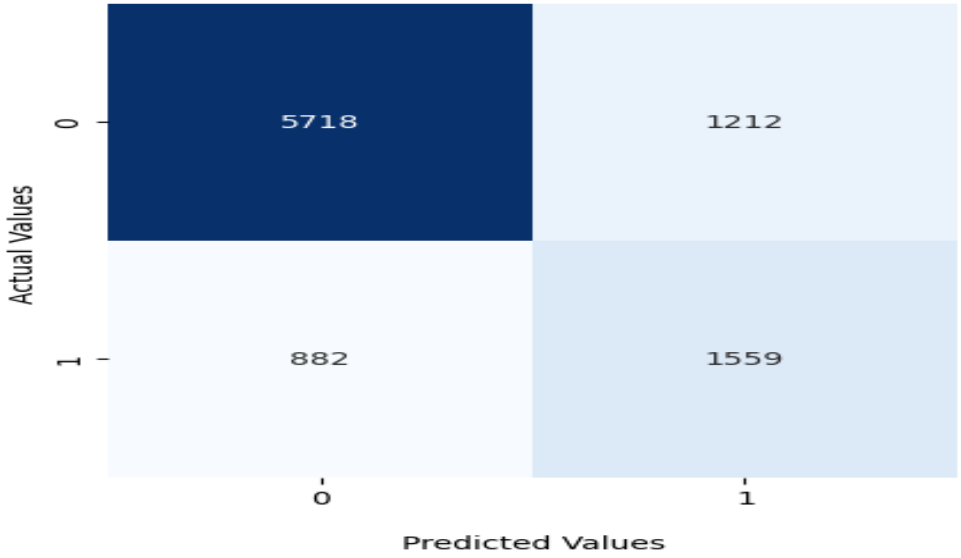

Figure 7: Confusion Matric Cyberbullying Detection Model (CBDS-ConvNet)

Using these values, the evaluation metrics were computed using Equations (1), (2), (3), and (4) above. From the results, the following observations were recorded:

1. Accuracy (77.65%): This metric represents the overall correctness of the model's predictions. In this case, the model is accurate in its classification 77.65% of the time.
2. Precision (56.26%): Precision measures the accuracy of positive predictions. In the context of cyberbullying detection, it indicates the percentage of instances predicted as "Bullying" that are actually "Bullying."
3. Recall (63.86%): Recall, also known as sensitivity or true positive rate, measures the model's ability to capture all instances of a positive class. In the context of cyber bullying, it represents the percentage of actual "Bullying" instances correctly identified by the model.
4. F1 Score (60.20%): The F1 score is the harmonic mean of precision and recall, providing a balanced metric that considers both false positives and false negatives. It is particularly useful when there is an imbalance between the classes. These outcomes are also detailed in Table 1.

Table 1: Performance of the Proposed CNN-Based Cyberbullying (CBDS-ConvNet) Detection Model

Metric	CBDS-ConvNet Model Performance (%)
Accuracy	77.65
Precision	56.26
Recall	63.86
F1 Score	60.20

4.3 Further Evaluation with other Dataset/Scenarios (Experiment B)

To evaluate the robustness of the CBDS-ConvNet model, it was tested on additional datasets, including data with synthetic noise designed to assess its resilience against adversarial inputs. This synthesized dataset comprised a total of 430 combined samples and, while distinct from the original training data, was subjected to the same preprocessing steps to ensure consistency. The results of this evaluation are presented in Figure 8 and Table 2.

Table 2: Performance metrics of the CBDS-ConvNet model on synthesized data

SN	Model	Accuracy	Precision	Recall	F1-Score
1	CBDS-ConvNet	91%	89%	81%	85%

```
classification_report:
              precision    recall  f1-score   support

           0       0.89      0.81      0.85       134
           1       0.92      0.96      0.94       296

    accuracy                           0.91       430
   macro avg       0.90      0.88      0.89       430
weighted avg       0.91      0.91      0.91       430


Confussion Matrix:
 [[108  26]
 [ 13 283]]
```

Figure 8: The CBDS-ConvNet Classification Report on the Synthesized data

The model demonstrated comparable performance metrics on the synthesized dataset, indicating its ability to generalize effectively beyond the initial experimental conditions. Key performance indicators, including accuracy, precision, recall, and F1-score, showed significant improvements, with the model achieving an accuracy of 91%, a precision of 89%, a recall of 81%, and an F1-score of 85%.

5. CONCLUSION

This study introduced and implemented a Convolutional Neural Network-based Cyberbullying Detection System (**CBDS-ConvNet)** using Twitter data. The model achieved an accuracy of 77.65%, precision of 56.26%, recall of 63.86%, and an F1 score of 60.20%. The experimentation phase involved training the CNN model over 20 epochs, optimizing convolution layers, and evaluating the model's performance on a test dataset. The data preprocessing steps included extracting tweets from multiple sources, standardizing target labels, handling duplicate texts, and cleaning the text by removing stop words, URLs, and punctuation. The tokenization process converted text into sequences of tokens, and sequences were padded to ensure uniform length for model input. To further evaluate the CBDS-ConvNet's robustness, it was tested on a synthesized dataset containing 430 samples with synthetic noise. The model achieved an accuracy of 91%, precision of 89%, recall of 81%, and an F1 score of 85%, demonstrating its ability to generalize across different data conditions. While the proposed CNN-based approach shows promise, further refinement and optimization are necessary to address the challenges posed by the dynamic nature of cyberbullying. Future research could explore ensemble models, incorporate additional features, and collaborate with social media platforms for real-time implementation and continuous improvement. This study contributes to the ongoing efforts to create safer online spaces by leveraging machine learning techniques for cyberbullying detection.

6. ACKNOWLEDGEMENT

The authors acknowledge the Almighty God for the wisdom and strength to complete this research. Sincere appreciation is also extended to the various academic repositories, journals, and research platforms that offered valuable knowledge and studies, which greatly influenced the development and direction of this work. The abundance of information provided by these resources was instrumental in shaping the depth and scope of the study.

7. REFERENCES

[1] N. S. Ruzgar, “A research on the purpose of internet usage and learning via internet,” *Turkish Online J. Educ. Technol.*, vol. 4, no. 4, pp. 27–32, 2015.

[2] A. O. Akinwumi, A. O. Akingbesote, O. O. Ajayi, and F. O. Aranuwa, “Detection of Distributed Denial of Service (DDoS) attacks using convolutional neural networks,” *Niger. J. Technol.*, vol. 41, no. 6, pp. 1017–1024, 2022, doi: 10.4314/njt.v41i6.12.

[3] A. Schonfeld, D. Mcniel, T. Toyoshima, and R. Binder, “Cyberbullying and Adolescent Suicide,” *J. Am. Acad. Psychiatry Law*, vol. 51, no. 1, pp. 112–119, 2023, doi: 10.29158/JAAPL.220078-22.

[4] A. O. Akinwumi, O. L. Ogbeide, and D. F. Folorunso, “Implementing Image Steganography Techniques for Secure Data Hiding in the Development of an Android Application,” *Commun. Appl. Electron.*, vol. 7, no. 39, pp. 26–33, 2023, doi: 10.5120/cae2023652902.

[5] World Internet Stats, “World Internet Usage and Population Statistics,” Internet Usage Statistics. Accessed: Nov. 14, 2023. [Online]. Available: https://www.internetworldstats.com/stats.htm

[6] V. Malpe and S. Vaikole, “A Comprehensive Study on Cyberbullying Detection Using Machine Learning Approach,” vol. 13, no. 1, pp. 342–351, 2020.

[7] L. Minocha, P. Jain, A. Singh, and P. Pandey, “Social Media’s Impact on Business and Society: A Study,” in *8th International Conference on Advanced Computing and Communication Systems, ICACCS 2022*, 2022, pp. 2078–2081. doi: 10.1109/ICACCS54159.2022.9784959.

[8] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, “Advances in Social Media Research: Past, Present and Future,” *Inf. Syst. Front.*, vol. 20, no. 3, pp. 531–558, 2018, doi: 10.1007/s10796-017-9810-y.

[9] O. Oriola and E. Kotze, “Evaluating Machine Learning Techniques for Detecting Offensive and Hate Speech in South African Tweets,” *IEEE Access*, vol. 8, no. 2020, pp. 21496–21509, 2020, doi:

10.1109/ACCESS.2020.2968173.

[10] A. Saravanaraj, J. I. Sheeba, and S. P. Devaneyan, "Automatic Detection of Cyberbullying From Twitter," *IRACST-International J. Comput. Sci. Inf. Technol. Secur.*, vol. 6, no. 6, pp. 2249–9555, 2019, [Online]. Available: https://www.researchgate.net/publication/333320174

[11] A. O. Adediran, "Cyberbullying in Nigeria: Examining the Adequacy of Legal Responses," *Int. J. Semiot. Law*, no. 0123456789, 2020, doi: 10.1007/s11196-020-09697-7.

[12] A. Bozyi˘git, S. Utku, and E. Nasibov, "Cyberbullying detection : Utilizing social media features," *Expert Syst. Appl.*, vol. 179, no. 103411, 2021, doi: 10.1016/j.eswa.2021.115001.

[13] T. K. H. Chan, C. M. K. Cheung, and Z. W. Y. Lee, "Cyberbullying on social networking sites: A literature review and future research directions," *Inf. Manag.*, vol. 58, no. 2, p. 103411, 2020, doi: 10.1016/j.im.2020.103411.

[14] P. Ingle, R. Joshi, N. Kaulgud, A. Suryawanshi, and M. Lokhande, "Detecting Cyber Bullying on Twitter," vol. 7, no. 19, pp. 1090–1094, 2020.

[15] S. Hinduja and J. W. Patchin, "Cyberbullying: Identification, Prevention, and Response 2024 Edition," 2024. [Online]. Available: https://cyberbullying.org/Cyberbullying-Identification-Prevention-Response-2024.pdf

[16] H. Lin, P. Siarry, H. L. Gururaj, J. Rodrigues, and D. K. Jain, "Special Issue on Deep Learning Methods for Cyberbullying Detection in Multimodal Social Data," *Multimed. Syst.*, vol. 28, no. 6, pp. 1873–1875, 2022, doi: 10.1007/s00530-022-01000-x.

[17] R. S. Pawar, "Multilingual Cyberbullying Detection System," The Purdue University, 2019.

[18] J. Hani, M. Nashaat, and M. Ahmed, "Social Media Cyberbullying Detection using Machine Learning," vol. 10, no. 5, pp. 703–707, 2019.

[19] J. I. Nwufo and M. B. Nwoke, "Cyber Bullying in Contemporary Nigeria: Implications on Youths' Psychological Wellbeing," *Pract. Psychol.*, vol. 8, no. 1, pp. 167–182, 2018, [Online]. Available: http://journals.aphriapub.com/index.php.pp

[20] 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, pp. 1–13, 2022, doi: 10.1007/s42979-022-01308-5.

[21] R. Deschamps and K. Mcnutt, "Cyberbullying : What ' s the problem ?," *Can. Public Adm.*, vol. 59, no. 1, pp. 45–71, 2016.

[22] N. M. Aune, "Cyberbullying," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2019.

[23] K. C. Nwosu, E. C. Ngozi, and E. P. Eberechi, "Cyberbullying among undergraduate students in a Nigerian university: Awareness and incidence," *Rom. J. Psychol. Stud.*, vol. 6, no. 1, pp. 43–58, 2018.

[24] M. Fridh, M. Lindström, and M. Rosvall, "Subjective health complaints in adolescent victims of cyber harassment : moderation through support from parents / friends - a Swedish population-based study," *BMC Public Health*, no. February 2016, 2015, doi: 10.1186/s12889-015-2239-7.

[25] B. S. Nandhini and J. I. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques," *Procedia - Procedia Comput. Sci.*, vol. 45, pp. 485–492, 2015, doi: 10.1016/j.procs.2015.03.085.

[26] A. Muneer, "A Comparative Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter," 2020.

[27] P. Hajibabaee, M. Malekzadeh, M. Ahmadi, M. H. A. Esmaeilzadeh, and R. Abdolazimi, "Offensive Language Detection on Social Media Based on Text Classification," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA*, 2022, pp. 0092–0098. doi: 10.1109/CCWC54503.2022.9720804.

[28] R. Chauhan, K. K. Ghanshala, and R. . Josh, "Convolutional Neural Network (CNN) for Image Detection and Recognition," in *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, IEEE, 2018, pp. 278–282. doi: 10.1109/ICSCCC.2018.8703316.

[29] S. Indolia, A. K. Goswami, S. P. Mishra, and P. Asopa, "Conceptual Understanding of Convolutional Neural Network- A Deep Learning Approach," *Procedia Comput. Sci.*, vol. 132, pp. 679–688, 2018, doi: 10.1016/j.procs.2018.05.069.

[30] A. O. Ige and M. H. M. Noor, "A lightweight deep learning with feature weighting for activity recognition," *Comput. Intell.*, vol. 39, no. 2, pp. 315–343, 2022, doi: https://doi.org/10.1111/coin.12565.

[31] Z. Sun, "Pattern Recognition in Convolutional Neural Network (CNN)," *Appl. Intell. Syst. Multi-modal Inf. Anal.*, vol. 138, pp. 295–302, 2022.

[32] H. M. Z. Haque, "Aortic Valve Segmentation using Convolutional Neural Network with Skip Mechanism," *Commun. Appl. Electron.*, vol. 7, no. 29, pp. 1–5, 2019.

[33] F. O. Aranuwa and O. B. Fawehinmi, "Classification Model For Iris Images Using Convolutional Neural Network (CNN)," in *Proceedings of the 32nd Accra Multidisciplinary Cross-Border Conference (AMCBC)*, Accra, 2022, pp. 7–22.

[34] C. Szegedy, A. Toshev, and D. Erhan, "Deep Neural Networks for Object Detection," *Adv. neural Inf. Process. Syst.*, pp. 2553–2561, 2013, doi: 10.3928/19404921-20140820-01.

[35] Y. Zhang *et al.*, "Towards end-to-end speech recognition with deep convolutional neural networks," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 08-12-Sept, no. September, pp. 410–414, 2016, doi: 10.21437/Interspeech.2016-1446.

[36] X. Luo, R. Shen, J. Hu, J. Deng, L. Hu, and Q. Guan, "A Deep Convolution Neural Network Model for Vehicle Recognition and Face Recognition," *Procedia Comput.*

Sci., vol. 107, no. Icict, pp. 715–720, 2017, doi: 10.1016/j.procs.2017.03.153.

[37] A. Ucar, "Deep Convolutional Neural Networks for facial expression recognition," *Proc. - 2017 IEEE Int. Conf. Innov. Intell. Syst. Appl. INISTA 2017*, pp. 371–375, 2017, doi: 10.1109/INISTA.2017.8001188.

[38] A. E. Ebitigha, O. O. Ajayi, O. D. Akinrolabu, A. Adegbite, J. Obafemi, and J. K. Ogunleye, "Facial Appearance Analysis for Age Group Prediction Using Convolutional Neural Network," in *International Conference on Science, Engineering and Business for Driving Sustainable Development Goals*, 2024.

[39] L. C. Hon and K. D. Varathan, "Cyberbullying Detection System on Twitter," pp. 1–11, 2015, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.740.1582&rep=rep1&type=pdf

[40] Y. Liu, C. Kliman-silver, and A. Mislove, "The Tweets They are a-Changin ': Evolution of Twitter Users and Behavior," in *Proceedings of the International Association for the Advancement of Artificial Intelligence Conference on Web and Social Media*, 2014, pp. 5–15.

[41] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore, and M. Camacho-Collados, "Detecting and monitoring hate speech in twitter," *Sensors (Switzerland)*, vol. 19, no. 21, pp. 1–37, 2019, doi: 10.3390/s19214654.

[42] B. Dean, "New Twitter." Accessed: Apr. 13, 2021. [Online]. Available: https://backlinko.com/twitter-users#daily-active-users

[43] Turbofuture.com, "Cyberbullying on Social Media (What It Is and How to Avoid It) - TurboFuture." Accessed: Apr. 13, 2021. [Online]. Available: https://turbofuture.com/internet/Cyberbullying-and-Social-Media

[44] V. Balakrishnan, S. Khan, and R. A. Hamid, "Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning," *Comput. Secur.*, p. 101710, 2019, doi: 10.1016/j.cose.2019.101710.

[45] D. Chatzakou *et al.*, "Detecting Cyberbullying and Cyberaggression in Social Media ∗," *ACM Trans. Web*, vol. 13, no. 3, 2017.

[46] C. Nobata and J. Tetreault, "Abusive Language Detection in Online User Content," in *Proceedings of the 25th international conference on World Wide weW*, International World Wide Web Conferences Steering Committee, 2016, pp. 145–153.

[47] M. A. Al-garadi, M. A. Al-garadi, M. R. Hussain, and N. Khan, "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms : Review of Literature and Open Challenges," *IEEE Access*, vol. PP, no. May, p. 1, 2019, doi: 10.1109/ACCESS.2019.2918354.

[48] R. Zhao, A. Zhou, and K. Mao, "Automatic Detection of Cyberbullying on Social Networks based on Bullying Features," in *Proceedings of the 17th international conference on distributed computing and networking*, ACM, 2016, pp. 1–6.

[49] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," *Int. Conf. 2012 Int. Confernece Soc. Comput. (SocialCom*, pp. 71–80, 2012, doi: 10.1109/SocialCom-PASSAT.2012.55.

[50] S. M. Isa and L. Ashianti, "Cyberbullying Classification using Text Mining," in *1st International Conference on Informatics and Computational Sciences (ICICoS)*, IEEE, 2017, pp. 241–246.

[51] R. Sugandhi, A. Pande, A. Agrawal, and H. Bhagat, "Automatic Monitoring and Prevention of Cyberbullying Automatic Monitoring and Prevention of Cyberbullying," *Int. J. Comput. Appl.*, vol. 144, no. 8, pp. 17–19, 2016, doi: 10.5120/ijca2016910408.

[52] B. Irena and E. B. Setiawan, "Fake News (Hoax) Identification on Social Media Twitter using Decision Tree C4.5 Method," *RESTI J. (System Eng. Inf. Technol.*, vol. 4, no. 4, pp. 711–716, 2021, doi: 10.29207/resti.v4i4.2125.

[53] M. A. Al-garadi, K. D. Varathan, and S. D. Ravana, "Computers in Human Behavior Cybercrime detection in online communications : The experimental case of cyberbullying detection in the Twitter network," *Comput. Human Behav.*, vol. 63, pp. 433–443, 2016, doi: 10.1016/j.chb.2016.05.051.

[54] M. Wang, H. Li, M. S. Hospital, W. Jiang, and Q. Liu, "gen CNN: A Convolutional Architecture for Word Sequence," in *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing: Association for Computational Linguistics., 2015, pp. 1567–1576. doi: 10.3115/v1/P15-1151.

[55] M. A. Al-Ajlan and M. Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 9, pp. 199–205, 2018.

[56] X. Zhang, J. Zhao, and Y. Lecun, "Character-level convolutional networks for text classification," *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 649–657, 2015.

[57] R. Johnson and T. Zhang, "Effective use of word order for text categorization with convolutional neural networks," *NAACL HLT 2015 - 2015 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, no. 2011, pp. 103–112, 2015, doi: 10.3115/v1/n15-1011.

[58] P. R. Vardhani, Y. I. Priyadarshini, and Y. Narasimhulu, "CNN Data Mining Algorithm for Detecting Credit Card Fraud," *SpringerBriefs Appl. Sci. Technol.*, pp. 85–93, 2019, doi: 10.1007/978-981-13-0059-2.

[59] R. Dai, "Text Data Mining Algorithm Combining CNN and DBM Models," *Mob. Inf. Syst.*, vol. 2021, pp. 1–7, 2021.

[60] J. Peng and S. Huo, "Application of an Improved Convolutional Neural Network Algorithm in Text Classification," *J. Web Eng.*, vol. 23, no. 03, pp. 315–340, 2024, doi: 10.13052/jwe1540-9589.2331.