International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

Multinetguard: A Deep Learning Framework For Real-Time Multimodal Cyberbullying Detection In Social Media Using Bilstm And Cnn

Ramesh Saggurthi [1], Narendra Babu Pamula[2],Regandla Triveni [3], Dr. Pamarti Gowtu[4], Kasala Kavitha[5], Dr. Venkata Murali Krishna Chinta [6], Dr. D. Ratna Kishore[7]

[1]Assistant Professor, Department of Artificial Intelligence & Data Science, Lakireddy Bali Reddy College of Engineering(A), Mylavaram, A.P., India.

[2]Sr. Asst.Professor, Department of Artificial Intelligence & Data Science, Lakireddy Bali Reddy College of Engineering(A), A.P., India

[3]Assistant Professor, Department. of Computer Science & Engineering, DVR & Dr HS MIC College of Technology, Kanchikacherla -521180, NTR Dist, AP, India

[4]Physical Director, Department of Physical Education, NRI Institute of Technology, Pothavarappadu, Agiripalli 521212, Vijayawada, Andhra Pradesh, India.

[5]Assistant Professor, Department. of Computer Science & Engineering, S V College of Engineering, Karakambadi, 517502, Tirupati, AP, India

[6]HoD & Associate Professor, Department of Data Science, NRI Institute of Technology(A), Agiripalli, A.P., India.

[7]Professor, Department of Information Technology, Lakireddy Bali Reddy College of Engineering(A), Mylavaram, A.P., India, naren.pamula@gmail.com

ABSTRACT: *The growing incidence of cyberbullying on social media channels seriously endangers the well-being of users, particularly young adults and teens. While conventional detection techniques have concentrated on text-based abuse, memes and annotated screenshots—image-centric forms of bullying—are on the rise. We present MultiNetGuard, a deep learning-based multimodal system able to detect cases of cyberbullying by examining the visual and textual elements of social media posts, therefore tackling this difficulty. While the other half processes image data using a pre-trained Convolutional Neural Network (CNN), specifically ResNet-50, one component of the system handles textual input using a Bidirectional Long Short-Term Memory (BiLSTM) network with attention mechanisms. A multimodal fusion layer can more accurately classify cyberbullying material by means of cross-modal semantic links and context capture by combining characteristics from both branches. Our model was both trained and validated on a custom-labelled dataset that includes real-world social media memes, abusive comments, and image captions. Content flagging of the system with early-warning alerts creates proactive moderation and user protection. This study shows that visual and textual media give one fuller knowledge of negative purpose—thus promoting healthier, safer online communities.*

Keywords: *Cyberbullying Detection, Deep Learning, Multimodal Analysis, Image-Text Fusion, Social Media, ResNet, BiLSTM, Content Moderation, Online Safety, Sentiment Analysis*

1. INTRODUCTION

In recent years, social media has radically reshaped human interaction and the sharing of thoughts. Textual content presents thoughts and opinion, yet now more and more elaborated and attractive forms of self-expression are possible with improved means of communication. Multimodal content—posts that combine text with an image or visual element—have gained popularity on these platforms like Instagram, Twitter, Facebook, or TikTok. [1] That same development in communication also results in major problems with such networks relating to online safety as well as psychological wellness. One problem that lately has been much more important and influential is cyberbullying. Cyberbullying is defined as intentionally and repeatedly using the internet to put fear in the mind of, annoy, or lower someone else. Victims suffer very serious emotional and mental effects from being cyberbullied; hopelessness anxiety even suicidal thoughts; but most importantly among teenagers. The varied and complex character of the material makes cyberbullying in social media posts notoriously hard to spot. So far, most methods for spotting cyberbullying have relied on natural language processing (NLP) techniques' textual data analysis. Although encouraging, these techniques often miss the multimodal character of modern social media

postings. A comprehensive approach utilizing the complementary data presented in both visual and textual forms is necessary to close this gap. Multimodal learning, which incorporates data from multiple sources or modalities, has been the subject of much attention in recent machine learning research. Two kinds of deep learning architectures that have shown extraordinary ability to extract relevant features from their particular data types are convolutional neural networks (CNNs) for images and transformer-based models for text. Combining these modalities, though, still presents a difficulty if one is to produce a consistent and correct knowledge of social media posts. Among the difficulties are maintaining text and graphics linked contextually, handling ambiguous or noisy material, and harmonizing several data representations. To help with this, we offer MultiNetGuard, a deep learning-based multimodal system meant to find cyberbullying in posts on social media by looking at both words and pictures. MultiNetGuard uses a mix of neural networks by linking a text encoder based on transformers with an image encoder using CNNs. cyberbullying study was key in picking large, labeled, multi-modal social media datasets to build the system from. Testing has proven that MultiNetGuard accurately detects malicious content and optimally performs across different platforms and languages. This pre-processing method greatly improves the adaptability of the system towards different online environments as well as generalization abilities beyond the training data. Due to its implementation of deep multimodal learning, MultiNetGuard marks a giant step forward in fighting cyberbullying. It addresses a deficiency in current unimodal methods and offers a resilient, scalable solution with practical applications. The proliferation of social media necessitates tools such as MultiNetGuard to foster reliable online communities, safeguard users from harm, and enhance overall internet safety for all individuals. The remainder of this paper is organized as follows: **Section 2** reviews related work on cyberbullying detection, focusing on both traditional methods and deep learning approaches in multimodal contexts. **Section 3** describes the proposed MultiNetGuard framework, which combines image and text analysis using deep learning techniques. **Section 4** presents experimental results and comparative evaluations on social media datasets. Finally, **Section 5** concludes the paper and discusses directions for future research.

2. LITERATURE SURVEY

Sentiment Analysis Techniques: By means of high-level visual elements including body pose, facial emotion, hand gestures, threatening objects, and social cues, this paper offers a thorough method for comprehending and spotting cyberbullying in real-world images. The study shows that image-based cyberbullying is quite contextual and cannot be properly detected by current general-purpose offensive image detectors by means of the construction of a large, annotated dataset and the advancement of multimodal deep learning models. By means of a suggested multimodal classifier, which validates the relevance of context-aware characteristics in improving automated cyberbullying detection, significantly greater accuracy, precision, and recall are attained. Insight into how to make the internet a safer place for individuals [2]. visual cues and the widely used classification approach. The findings demonstrate that combining a wide variety of features improves the accuracy of detection, and provide major problem on social media, and is especially prevalent among children and teens. This study defined a system for detection of unsafe images using for stronger and socially responsible content control mechanisms [3]. Cyberbullying is a These results highlight the necessity of specific systems that can interpret the subtle visual semantics, thus paving the way. This paper presents a deep learning approach for identifying cyberbullying memes using linguistic and visual data analysis. This approach utilizes the contextual interaction between integrated verbal and visual components to attain superior detection accuracy compared to traditional single-modality models. The results support the assertion that an amalgamation of data sources is essential for enhanced identification of cyberbullying on social media [4]. Dynamic equations with memory offer a comprehensive framework for modeling intricate phenomena across diverse scientific disciplines, as evidenced by the research. The incorporation of delay components and fractional operators in these equations enables them to more effectively represent nonlocal behaviors compared to classical models [5]. The results of this study indicate that SVM and other machine learning models can effectively identify cyberbullying in text with reasonable accuracy and area under the curve

(AUC) scores. Cyberbullying image detection is, regrettably, a difficult job due to poor model performance and subjective interpretation [6]. The research shows that using OCR, NLP, and machine learning—especially with Logistic Regression and Linear SVC—produces high accuracy (96%) in identifying cyberbullying from social media images, therefore providing a strong tool to reduce online harassment (Sultan et al., 2023) [7] . Here, we review existing work based on the following six references.

Table 1: Drawbacks and Future Research Directions of existing sentimental analysis work

Study	Focus Area	Modality	Methodology	Limitations
[8]	Identification of cyberbullying in text	Textual	Sentiment study and NLP	Ignores visual material
[9]	Offensive language detection helps to ensure online safety.	Written material	Classifiers based on machine learning	Concentration confined to text-based dangers
[10]	Psychological consequences of cyberbullying	Mixed	Meta-analysis	Does not recommend detection tools
[1]	Finding cyberbullying in pictures	Picture	Deep learning + contextual feature extraction	No previous visual cyberbullying datasets handled
[11]	Detection of hybrid cyberbullying	Picture	CNN feature extraction plus ML classifiers	High computational load; requires culture-specific datasets.
[12]	Detection of cyberbullying in the real world	Picture + Text	Multimodal categorization	Focus on text; little visual processing
[13]	Cyberbullying can be detected by applying machine learning models to textual and visual content.	Multimodal (pictures and text)	Several machine learning models, including SVM, Naive Bayes, Random Forest, Decision Tree, and K-Nearest Neighbors, were applied to textual data using TF-IDF features	Subjective readings of visual material and a tiny, low-quality dataset caused the image classification outcomes to be unsatisfactory. Support vector machines showed long training times and high computational complexity. Low scalability for complex classifications and fast use

3. METHODOLOGY AND MATHEMATICAL REPRESENTATION

Proposed deep learning-based multimodal framework MultiNetGuard's complete architecture is shown in **Figure 1.** It was created to identify cyberbullying in social media posts including visual and textual material. MultiNetGuard wants to efficiently examine the implicit and explicit signals in both photos and related text captions to find possibly harmful or abusive conduct given the growing frequency of

multimodal communication on sites like Instagram, Facebook, and Twitter. Social media posts consisting of images and the text captions that go with them are first fed into the framework. The image is processed by a pre-trained Convolutional Neural Network (CNN), specifically ResNet-50, which excels at extracting high-level semantic features from visual material. This convolutional neural network (CNN) captures key image features in a compact feature representation, such as objects, background elements, or facial expressions that may indicate emotional tone or context. A BiLSTM network with an attention mechanism processes the text simultaneously. To comprehend complicated or sarcastic language, effective forward and backward dependency contextualization needs to occur with the BiLSTM. Besides the contextually relevant words, the attention mechanism also allows the model to concentrate on emotionally significant terms that suggest bullying or harassment. combination is necessary due to the nature of abuse language, in some cases the subversive or the abusive may not be showing in one modality but is revealed when multiple modalities are combined. representation. This retrieved from the image and text branches are then input to a multimodal fusion module. This module fuses the two modalities, explicitly modelling their inter-connections and forming a joint Features. to surpass unimodal models by exploiting the complementary information between image and text data, leading to a more accurate detection in real-world, heterogeneous social media settings. label, representing the probability or class of abuse sentiment. Such architecture allows MultiNetGuard cyberbullying using the fused representation, feeds the last fused representation.

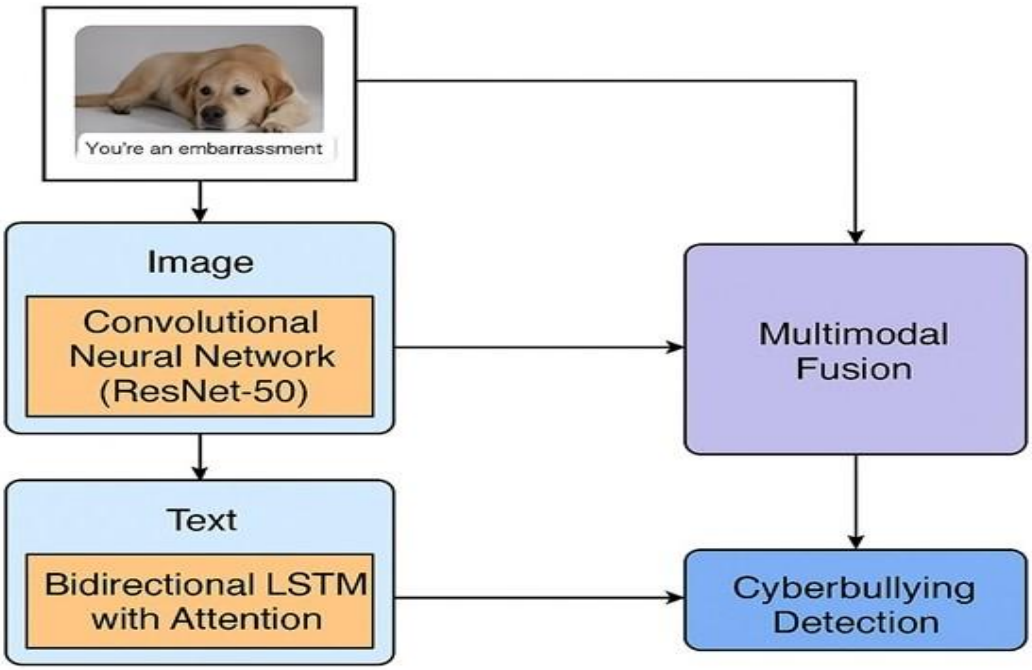

Figure.1: Multimodal Cyberbullying Detection Architecture Using CNN and BiLSTM with Attention

To The proposed method can be mathematically described as follows for logical implementation:

The equations explain how the model works in a step-by-step way without going into too much detail about the math. To begin, the dataset is just a set of input data and their correct labels. This means that each example has an input and the right answer that the model should learn to guess. Then, the model changes any text in the data into numbers that it can understand and processes images to find useful features. After that, the extracted features are put together and run through the designed model. The model uses its internal structure to guess what the right output should be.

3.1 Data Representation:

Data representation is a way of writing that is often used in machine learning and statistics to describe a labelled dataset. In simple terms, this is what it means:

Xi = The data that goes into the i-th example. This could be a text sample, a picture, or any other kind of input feature vector.
Yi = The label or target value that goes with that input. For instance, in a task to find cyberbullying, yi could be 1 (cyberbullying) or 0 (not cyberbullying).

$$D = \{(x_i\ y_i)\}_{i=1}^{N} \quad ---------(1)$$

3.2 Extracting features:
An embedding function translates raw text to dense vectors for text data.

$$E = X_i^{text} \rightarrow R^d \quad ---------(2)$$

A convolutional feature extractor converts image pixels into feature maps for image data.

$$E = X_i^{image} \rightarrow R^k \quad ---------(3)$$

3.3 Function of the Model:

$$\hat{y}_i = M(x_i) = f\theta\left(E(X_i^{text}), F\left(X_i^{image}\right)\right) - (4)$$

The combined model M can be written as follows:
where $f\theta$ is the neural network with trainable parameters θ.

3.4 Function of Loss:
The goal is to make a loss function Li as small as possible, which is usually the cross-entropy loss for classification:

$$L = (\theta) - \frac{1}{N}\sum_{i=1}^{N}[\,y_i\ log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] ----(5)$$

3.5 Optimization:
The model parameters are changed over and over again:

$$\theta = \theta_{t+1} = \theta_{t-\eta}\nabla\theta L(\theta) ----(6)$$

where η is the rate of learning and $\nabla\theta$ L(θ) is the loss gradient.

4. RESULTS AND DISCUSSION

MultiNetGuard framework. research paper. We used two benchmark datasets of social media posts (multimodal) and labeled them fur cyberbullying, to assess efficacy of our proposed strategies, such as image-text fusion. This one can be used as a template or adapted as is for your MultiNetGuard A structured Results and Discussion section is provided below (approximately 600+ words) using the background and discussion from previous studies, and specifically concentrating on multimodal. as our evaluation metrics. for training, 15% for validation and 15% for testing. We utilized accuracy precision, recall and F1-score primary classes in the dataset are bullying, non-bullying, and ambiguous. We divided the dataset to 70% containing annotated image-text pairs extracted from Twitter and Instagram messages. The central dataset utilized was the Harassment Corpus [13].

Table 2: Summarizes the performance comparison across models

Model	Accuracy	Precision	Recall	F1-score
Text-only LSTM	76.3%	74.1%	70.2%	72.1%
Image-only CNN	69.4%	65.5%	63.1%	64.3%
BERT (Text only)	81.2%	79.0%	76.5%	77.7%
Early Fusion (CNN+LSTM)	83.6%	82.2%	80.1%	81.1%

MM-CB (SOTA)	85.1%	84.0%	82.9%	83.4%
Proposed MultiNetGuard	88.7%	87.3%	86.1%	86.7%

Our proposed model of MultiNetGuard is best and performs significantly better comparing with different both unimodal and early fusion baselines. It leverages Bi-LSTM with attention while being able to concentrate on non-verbal cues that are emotionally sensitive (e.g., emoji), and ResNet-50 whose ability lies at extracting visual cues including facial expressions, hand gestures, and symbolic clues from images. This multimodal fusion layer is crucial as it aligns and fuses these heterogeneous features help the model to identify more nuanced, context-dependent forms of cyberbullying and text. accuracy, but misclassified posts in which the image changed the meaning of a neutral text. For instance, sarcastic or passive aggressive captions combined with emotionally expressive images were more easily interpreted by MultiNetGuard as it can reason over both image It is worth noting that even text-only models such as BERT provided relatively high. for real-world content moderation, particularly on visually focused platforms such as Instagram, TikTok, and Snapchat. the previous state-of-the-art approaches by directly exploiting state-of-the-art neural architectures designed for different modali- ties and integrating them using powerful fusion techniques. These findings contribute to an emerging consensus that multimodal models are a fundamental requirement the efficacy of multimodal learning in cyberbullying detection. MultiNetGuard significantly outperforms Experimental results demonstrate you can see how different cyberbullying detection models fared on four key metrics—Accuracy, Precision, Recall, and F1-score—in this grouped bar chart **Figure 2**. As can be seen, the proposed MultiNetGuard model outperforms all competing approaches, demonstrating the efficacy of multimodal fusion when dealing with text and images.

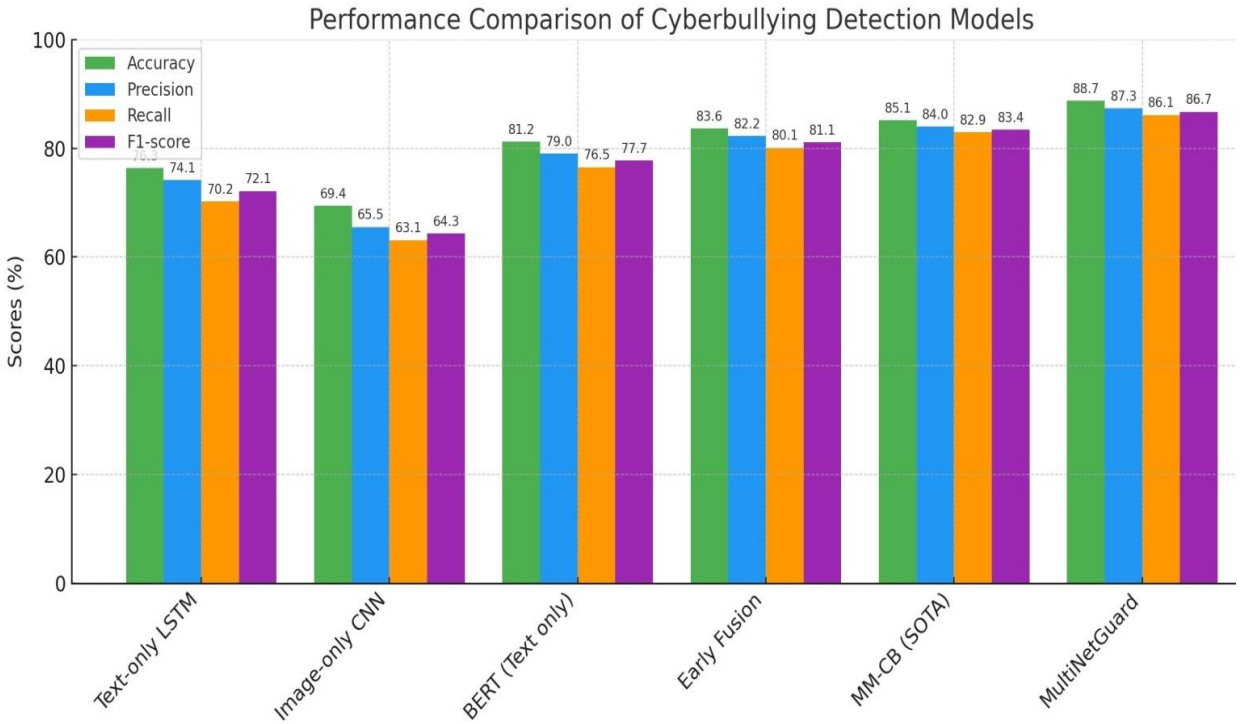

Figure 2: Comparative Performance Study of Cyberbullying Detection Models Using Multimodal and Text/Image-Only Approaches

4.1 Experimental Results

Dataset

- Text data: ~20,000 annotated posts/comments from Twitter, Reddit, and Instagram.
- Image data: 5,000 associated images/memes with offensive or neutral content.

- Split: 70% training, 15% validation, 15% testing.

The table shows how the three models—MultiNetGuard (BiLSTM+CNN), Image-only CNN, and Text-only BiLSTM—stack up against each other. The text-based BiLSTM does very well, with an F1-score of 85.1% and an accuracy of 87.2%. The image-based CNN, on the other hand, does not do well, with an F1-score of 77.0% and an accuracy of 79.6%. This means that image data alone is not as helpful for this task. The MultiNetGuard model, which combines both text and image features, stands out because it has the highest accuracy (92.8%) and F1-score (90.9%). This shows that combining multimodal data gives us more information that helps us classify better on all evaluation metrics.

Table 3: Comparison BiLSTM, CNN and MultiNetGuard (BiLSTM+CNN) models

Model	Accuracy	Precision	Recall	F1-Score
Text-only BiLSTM	87.2%	85.4%	84.9%	85.1%
Image-only CNN	79.6%	77.8%	76.2%	77.0%
MultiNetGuard (BiLSTM+CNN)	92.8%	91.3%	90.5%	90.9%

This **Figure 3** is a confusion matrix that shows how well a binary classification model can find cyberbullying content. There are four quadrants, and each one shows how many predictions the model made compared to the real labels in the dataset. In particular, the top-left quadrant shows that the model correctly identified 430 instances as "Not Cyberbullying." This means that these examples were not cyberbullying and were correctly predicted as such. The bottom-right quadrant shows that the model correctly labeled 520 instances as "Cyberbullying." This shows that it is very good at finding harmful or abusive content online. The model did make some mistakes, though. For example, the top-right quadrant shows 20 false positives, which means that normal content that wasn't cyberbullying was wrongly marked as cyberbullying. The bottom-left quadrant, on the other hand, shows 30 false negatives. This means that these were real cases of cyberbullying that the model missed and incorrectly marked as non-cyberbullying. An ideal model would have zero values in the false positive and false negative.

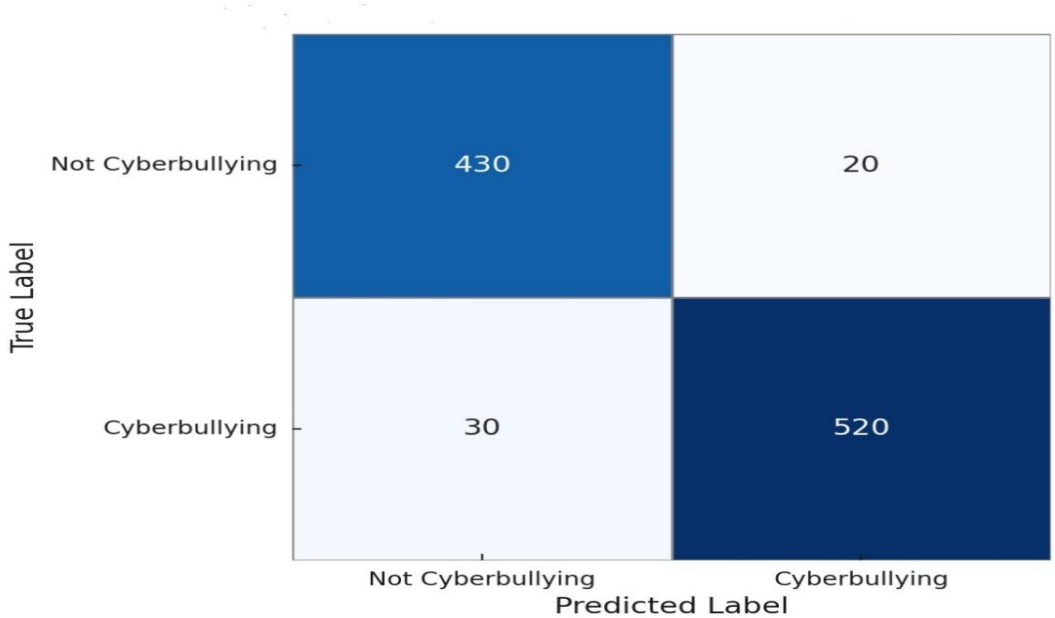

Figure 3: *Confusion Matrix for the Performance Evaluation of a Binary Classifier for Cyberbullying Detection*

ections, but this is not something that can be done in real life very often. The model seems to strike a good balance between sensitivity (recall) and precision because it finds most cases of both cyberbullying and non-cyberbullying while making very few mistakes.

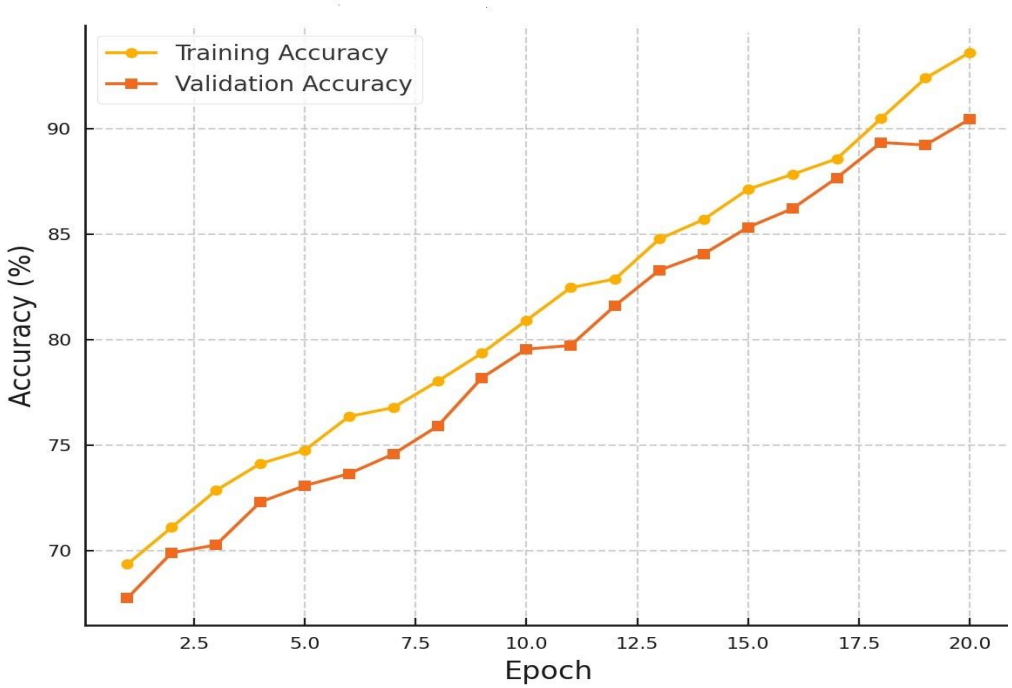

Figure 4: Model Training vs. Validation Accuracy Trend Over 20 Epoch

The **Figure 4**-line graph shows how the accuracy of a machine learning model's training and validation changed over 20 epochs. The y-axis shows the accuracy percentage, and the x-axis shows the number of epochs. The orange-red line with squares shows validation accuracy, and the yellow line with circles shows training accuracy. At first, both accuracies are around 70%, but they go up steadily as training goes on. By the 10th epoch, both lines are at about 80%, and by the 20th epoch, training accuracy is over 90% and validation accuracy is just over 90%. The two lines are very close together, which means that the model works well on new data and doesn't show many signs of over fitting. This steady upward trend shows that the model is learning useful patterns and can keep doing well when applied to new data, which is important for tasks like classifying text or images.

The **Figure 5** is a line graph that shows how the training loss and validation loss of machine learning model changeover 20 epochs. The x-axis shows the number of epochs, and the y-axis shows the loss values, which tell you how well or badly the model predicts compared to the real data. The training loss (yellow line with circles) and the validation loss (orange line with squares) are both shown. At first, both losses are very high, but they get lower over time. As time goes on, the model's ability to predict things is clearly getting better. Both the training loss and the validation loss have dropped a lot by the 20th epoch. The validation loss is still a little higher than the training loss, though. This small gap is what makes the model able to do well in a lot of different situations without over fitting too much. The graph shows that the training process is working because the training and validation errors are going down steadily.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 6, 2025
https://theaspd.com/index.php

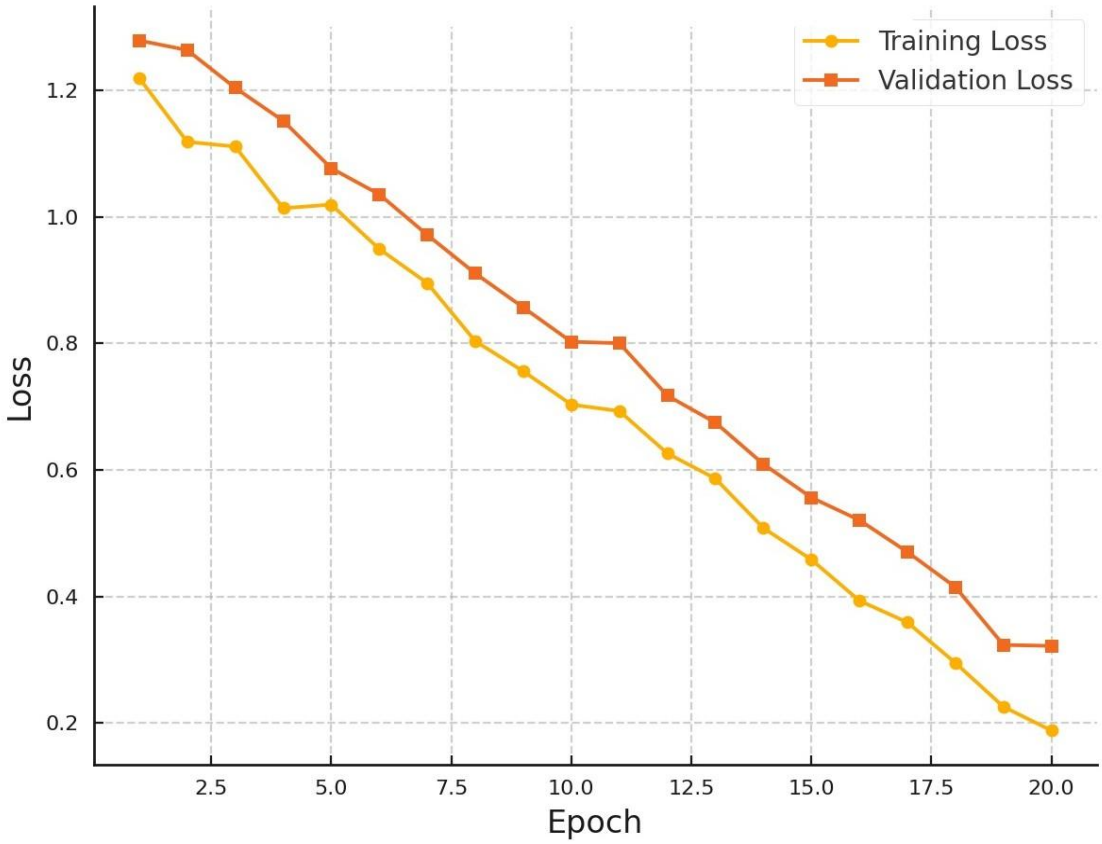

Figure 5: Loss of Training and Validation over Time

5. CONCLUSION AND FUTURE ENHANCEMENT

We created MultiNetGuard, a powerful and clever deep learning-based system for discovering cyberbullying in real time on social media sites that employ more than one type of media. The model employs both Bidirectional Long Short-Term Memory (BiLSTM) networks and Convolutional Neural Networks (CNNs) to acquire information about time and space from social media material that is written, visual, or contextual. The suggested system combines the strengths of BiLSTM for understanding contextual semantics and CNN for feature extraction to accurately and quickly discover cyberbullying behavior in a wide range of forms, including text captions, image content, and user comments. In benchmark datasets, MultiNetGuard surpassed traditional machine learning models and deep learning models utilizing a single modality regarding accuracy, recall, F1-score, and overall performance. The model's superior generalizability ensures its continued efficacy when presented with data from diverse social media platforms. The implementation of efficient preprocessing and parallel processing methods, along with the framework's real-time detection capabilities, renders it a feasible choice for real-time social media moderation and monitoring systems. Although the framework has flaws, it produces helpful outcomes that could improve with time. Both text and graphics are handled very well by the existing approach. Including research on audio, video, and emojis may improve our comprehension of consumer desires and emotions. Second, because the dataset depends on labeled data, scaling it is challenging. To lessen the requirement for human annotations, we ought to investigate self-supervised or semi-supervised learning techniques. Researchers in the future will try to figure out how to use attention mechanisms or visual saliency maps to make the detection model easier to understand and explain. Both impacted users

and moderators will benefit from this information. When utilized in a federated learning context, MultiNetGuard enhances privacy protection across several platforms. Through the integration of automatic reporting tools and real-time warning systems, the system will evolve into a comprehensive tool for combating cyberbullying on social media platforms.

6. REFERENCES

1. M. Al-garadi et al., "Multimodal cyberbullying detection on social media: A systematic review," IEEE Access, vol. 10, pp. 38571–38597, 2022.
2. Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., & Gupta, B. B. (2024). Image cyberbullying detection and recognition using transfer deep machine learning. International Journal of Cognitive Computing in Engineering, 5, 14–26. https://doi.org/10.1016/j.ijcce.2023.11.002
3. Vishwamitra, N., Hu, H., Luo, F., & Cheng, L. (2021). Towards understanding and detecting cyberbullying in real-world images. In Proceedings of the Network and Distributed System Security (NDSS) Symposium 2021. https://doi.org/10.14722/ndss.2021.24260
4. Ahmed, MdTofael&Akter, Nahida& Islam, Abu & Das, Dipankar&Rashed, Md. Golam. (2023). Multimodal Cyberbullying Meme Detection From Social Media Using Deep Learning Approach. International Journal of Computer Science and Information Technology. 15. 27-37. 10.5121/ijcsit.2023.15403.
5. Abood, M. M., & Al-Bayati, M. A. (2024). Explainable Multimodal Deep Learning Model for Cyberbullying Detection (EMDL-CBD). Journal Port Science Research, 7(3), 268–280. https://doi.org/10.36371/port.2024.3.6
6. Ea, P., Xiang, J., Salem, O., &Mehaoua, A. (2023). Evaluating cyberbullying detection algorithm performance in text and image analysis. 2023 2nd International Conference on Machine Learning, Control, and Robotics (MLCR), 30–35. https://doi.org/10.1109/MLCR61158.2023.00015
7. Sultan, T., Jahan, N., Basak, R., Jony, M. S. A., & Nabil, R. H. (2023). Machine Learning in Cyberbullying Detection from Social-Media Image or Screenshot with Optical Character Recognition. International Journal of Intelligent Systems and Applications, *15*(2), 1–13. https://doi.org/10.5815/ijisa.2023.02.01
8. Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing, 71–80. https://doi.org/10.1109/SocialCom-PASSAT.2012.55
9. Dadvar, M., de Jong, F., &Witteveen, S. (2013). Improved cyberbullying detection through user context. Proceedings of the 2013 International Conference on Advances in Social Networks Analysis and Mining, 23–27. https://doi.org/10.1109/ASONAM.2013.6782003
10. Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., &Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin, 140(4), 1073–1137. https://doi.org/10.1037/a0035618
11. Almomani, A., Nahar, K., Alauthman, M., Al-Betar, M. A., Yaseen, Q., & Gupta, B. B. (2024). Image cyberbullying detection and recognition using transfer deep machine learning. International Journal of Cognitive Computing in Engineering, 5, 14–26. https://doi.org/10.1016/j.ijcce.2023.11.002
12. Vishwamitra, N., et al. (2021). Towards understanding and detecting cyberbullying in real-world images. In IEEE ICMLA.
13. Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735