International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Jan-2024)

http://dx.doi.org/10.12785/ijcds/150130

Overcoming the Challenge of Cyberbullying Detection in Images: A Deep Learning Approach with Image Captioning and OCR Integration

Subbaraju Pericherla[1] **and E. Ilavarasan**[2]

[1,2]*Department of Computer Science Engineering, Puducherry Technological University, Puducherry, India*
Received 7 May. 2023, Revised 2 Jan. 2024, Accepted 6 Jan. 2024, Published 15 Jan. 2024

Abstract: Cyberbullying is a serious concern in today's digital age. The rapid increase in the use of social media platforms has made cyberbullying even more prevalent. The form of cyberbullying has also evolved with time. In the era of Web 1.0, cyberbullying was limited to text-based data, but with the advent of Web 2.0 and 3.0, it has expanded to images and multi-modal data. Detecting cyberbullying in text-based data is relatively easy as various natural language processing techniques (NLP) can be used to identify offensive language and sentiment. However, detecting cyberbullying in image-based data is a major challenge as images do not have a clear textual representation. Hence, bullies often try to bypass existing cyberbullying detection techniques by using images and multi-modal data. We proposed a deep learning technique named as CNBD-Combinational Network for Bullying Detection (CNBD), which is a combination of two networks: Binary Encoder Image Transformer (BEiT) and Multi-Layer Perceptron (MLP) network. To improve the performance of the CNBD, we supplied two additional input factors to the CNBD using concepts called Image Captioning(IC) and OCR (Optical Character Recognition) to extract text overlayed on the images. The experimental results proved the two additional factors gave an advantage to the CNBD technique in terms of accuracy, precision, and recall.

Keywords: Cyberbullying, Social media, Multi-Layer Perceptron, Deep learning, OCR, Image Captioning.

1. INTRODUCTION

Cyberbullying is a growing concern in many countries, including India. Social media platforms like Facebook and Twitter are increasingly being used as avenues for harassment, with a significant portion of such incidents involving young people. Cyberbullying can have serious mental health effects, with some cases even leading to suicide. Given the severity of the problem, there is a need to develop effective approaches to identify cyberbullying in social media messages. According to a study by the Cyber and Information Security Division of India's Ministry of Electronics and Information Technology (MEITY), 14% of all harassment incidents in India occur on Facebook and Twitter, with youngsters being particularly vulnerable. The study also found that women and children are the most frequent victims of cyberbullying in India.

The mental health effects of cyberbullying are well-documented in the literature. Several studies have found that cyberbullying is associated with anxiety, depression, and other mental health issues, with some victims even resorting to suicide. For instance, a study by Hinduja and Patchin et al.[1] found that cyberbullying victims are more likely to experience depression, anxiety, and suicidal ideation than non-victims. Similarly, a study by Kowalski et al. [2] found that cyberbullying is associated with increased suicidal ideation, attempts, and completions. Given the prevalence and severity of cyberbullying, there is a need for effective approaches to identify and intervene in such incidents. One promising approach is the use of natural language processing (NLP) techniques to automatically identify cyberbullying in social media messages. Several studies have explored the use of NLP techniques for this purpose, with promising results. Similarly, a study by Mishna et al.[3] emphasizes the importance of early detection and intervention, as well as providing emotional support and counseling to cyberbullying victims

In conclusion, cyberbullying is a serious problem in India and other countries, with social media platforms like Facebook and Twitter being significant sources of harassment. The mental health effects of cyberbullying can be severe, with some victims even resorting to suicide. Effective approaches are needed to identify and intervene in cyberbullying incidents, and NLP techniques offer promising possibilities for automating this process.

To the best of our knowledge, cyberbullying detection in images with vision transformer networks has not been applied so far. To achieve this task, the following contribu-

E-mail address: : raju.pericherla74@gmail.com, eilavarasan@pec.edu

tions were made to this work.

Proposed a CNBD technique for Cyberbullying (CB) detection in images.

Fine-tuned the transformer-based network, BEiT to our downstream task using Cyberbullying image dataset.

Fine-tuned VGG16 and LSTM (Long Short Term Memory) architectures with MS-COCO dataset to generate image captions.

The Multi-Layer Perceptron network is built to improve the performance of the model.

The rest of the paper has been organized as follows. The related works on cyberbullying in images have been discussed in section 2. The main contribution of this proposed research as the proposed CNBD technique is presented in Section 3. The results and discussion are illustrated in section 4. Finally, we have concluded with possible future enhancements in continuing this work are given in section 5.

2. Related Works

In this section, we presented past research works related to cyberbullying on image data. P.K Roy and Mali[4] proposed a transfer learning-based automated model for cyberbullying detection in images from social media networks. The proposed model extracts hidden features from cyberbullying. The experiments were carried out with two different datasets of 1000 images and 3000 images. They consider three deep learning models for cyberbullying detection in images: 2-dimensional CNN, Visual Geometry Group 16 (VGG16), and InceptionV3. Among the three models, the Inception V3 outperforms in terms of precision (87

Homa Hosseinmardi et al.[5] present a novel approach to detecting cyberbullying incidents on the Instagram social network. The authors propose a system that utilizes machine learning algorithms to automatically classify Instagram posts as either cyber bullying or non-cyberbullying. The system uses a combination of text and image features extracted from the posts to train a classification model. The authors evaluate the system's performance on a dataset of 10,000 Instagram posts manually labeled as cyberbullying or non-cyberbullying. The results show that the system achieves a high accuracy (92%) in detecting cyberbullying incidents on Instagram. The paper also provides a detailed analysis of the features that contribute most to the classification performance of the system. The authors believe that their approach can be used to develop effective tools for combating cyberbullying on social media platforms.

Haoti Zhong et al.[6] proposed a content-based approach for detecting cyberbullying on the Instagram social network. The authors developed a system that uses NLP and ML techniques to analyze the textual content of Instagram posts and comments. The system uses a set of features such as sentiment, emotion, and content similarity to identify posts and comments that contain cyberbullying. The authors evaluate the performance of the system on a dataset of 22,899 Instagram posts and comments, manually annotated as cyberbullying or non-cyberbullying. The obtained results believe the proposed model has a superior accuracy rate of 91.4% in detecting cyberbullying incidents on Instagram. Elmezain Mahmoud [7] proposed a hybrid classification model based on transformers and SVM to predict whether bullying takes place or not. Using the proposed combined models with the SVM classifier, the authors claim to have achieved an accuracy rate of 96.05%. Furthermore, the proposed model has a 99% classification accuracy for the bullying class and a 93% accuracy for the non-bullying class. The study highlights the negative impact of bullying on students' academic performance and the importance of taking appropriate action against anti-bullying and raising community awareness of the problem. The authors suggested that future works will focus on using Twitter texts with Google form questionnaires for classifying cyberbullying and how to stop it.

Rui Cao et al. [8] proposed a model, PromptHate, that uses pre-trained RoBERTa language models and constructs simple prompts to prompt the model for hateful meme classification. To make use of the latent knowledge in the Pre-trained Language Models, the authors present real-world examples. The model's performance is measured against state-of-the-art baselines, and the findings demonstrate that it excels with an AUC of 90.96 on two publicly available datasets. Agarwal et al. [9] presented two approaches to identifying hate memes using deep learning techniques. The first method incorporates features from several modalities, while the second employs sentiment analysis based on image captioning and text placed on the meme itself. These methods use a trifecta of deep learning algorithms—GloVe, encoder-decoder, and OCR using the Adamax optimizer. We utilize the Facebook Challenge Hateful Meme Dataset, which includes over 8,500 meme images, to test the methods. Facebook uses both methods in the ongoing challenge competition, and they both show promise on the validation dataset. K R Prajwal et al.[10] proposed a novel method to capture the image content on social media images. They implemented a two-stage approach for image caption for images. In stage-1, emotional representations are captured using Transfer Learning and in stage-2, facial emotions are extracted using encoders which are extracted from stage1.

T Tiwary et al. [11] have introduced the Automatic Image Captioning (AIC) technique to help visually impaired consumers identify products in online grocery stores. To solve this problem, they proposed an ECANN (Extended Convolutional Atom Neural Network). For caption extraction from e-commerce image data, the ECANN model combines the LSTM architecture and CNN. On the Grocery Store Dataset, the proposed ECANN model achieved an accuracy of 99.46%, and on the Freiburg Groceries dataset,

it achieved an accuracy of 99.32. Al-Malla et al. [12] proposed an attention-based encoder and decoder for an image captioning model which is a combination of CNN and object detection module(YOLO4). The proposed model was evaluated on MS-COCO and Flickr30k datasets. Efrat Blaier et al. [13] proposed Caption Enriched Samples(CES) technique and applied it to BERT and RoBERTa models to image caption from the images. The authors noticed that CSE improves the performance by 8.6% and 10%, respectively on test data. Yi Zhou et al. [14] have propounded a novel method for hateful memes detection using image captioning, OCR, and object detection. They applied the Triplet relation network to the extracted features.

Pengyuan Lyu et al. [15] presented a technique called MaskOCR for text recognition in images. The architecture contains two parts: encoder and decoder transformers to recognize text representation on the images. They evaluated the proposed MaskOCR technique over benchmark datasets of different languages. J Chen et al. [16] have to deal with texts that can be presented in any direction, they devised a Transformer-Based Super Resolution Network (TBSRN) equipped with a Self-Attention Module for sequential information extraction. The authors introduced a Position-Aware Module to highlight the location of each character and a Content-Aware Module to highlight the content of each character to extract information down to the character level. Deli et al.[17] have created a unique framework for precise scene text recognition and named as semantic reasoning network (SRN). Within this framework, a global semantic reasoning module (GSRM) is provided to capture global semantic context through multi-way parallel transmission. Nishanth Viswamitra et al. [18] presented a comprehensive study on the nature of bullying images. They found that cyberbullying in images can be characterized by five visual factors: facial emotion, objects, gestures, body pose, and social factors. They presented a novel methodology to collect cyberbullying images. Initially, they crawled 117,112 images from different online sources using the keyword of cyberbullying. Finally, 19,3000 valid images were annotated for experimentation. They proposed four classifiers namely: Baseline model, Factors-only model, Fine-tuned pre-trained model, and Multi-modal. Among all the four classifiers, the Multimodal classifier achieves the best accuracy of 93.36%. The major contributions of cyberbullying detection in images from the above literature survey are summarized in Table 1.

We have noticed some of the limitations from past research work on image-based cyberbullying detection. Most of the deep learning models are trained with datasets of small sizes. The manual annotation of images is a challenging task. The majority of the works consider only image factors such as hand gestures, facial expressions, and objects in the images. We observed that two additional factors will improve the performance of the classification task. We introduced Image captioning to capture image content from the input images and OCR to extract text which is overlayed on the images.

3. Methedology

This section explains the working of proposed CNBD technique. The proposed Combinational Network for Bullying Detection(CNBD) technique consists of two phases : Phase-1:Feature extraction from the input images and Phase-2 Multi-Perceptron Layer network for classification of input image. Figure 1 shows the proposed CNBD technique. In the Phase-1, three categories of input features extracted from input images. 1) Image features extracted directly from input images 2) Text features using Image Captioning 3) Text features from Text embedded on the input images.

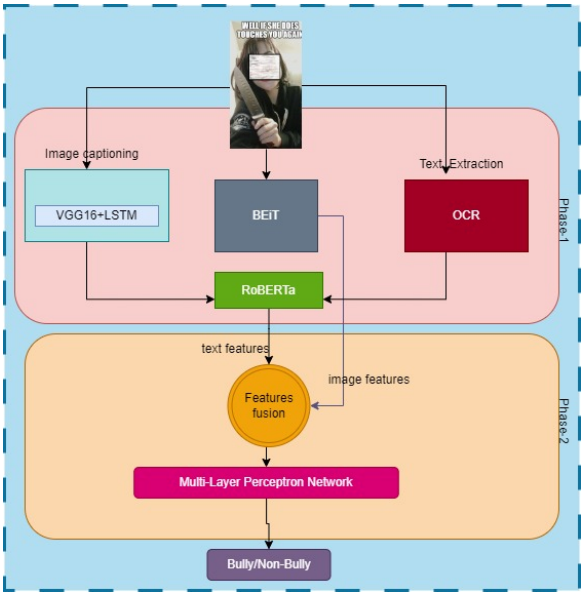

Figure 1. Architecture of CNBD technique.

A. *Phase-1:(Feature Extraction)*

At this stage, the pre-processed image data is fed into the Bidirectional Encoder for Image Transformers (BEiT)[19] feature extractor. The BEiT is a self-supervised vision model which is based on masked image modeling (MIM) functionality (illustrated in Figure 2). Initially, an image is divided into grids called tokens. These blocks of tokens are masked randomly and flattened into a vector. These embeddings and positional embeddings are learned for patches. These embeddings are passed through the BEiT encoder. Finally, image data can be reconstructed using tokens. Once fine-tuned with cyberbullying image dataset, it can be used for bullying detection in images.

Masked Image Modeling (MIM) is a computer vision technique that involves predicting the values of missing pixels in an image based on the surrounding visible pixels. The approach is based on a supervised learning framework, where the deep neural network is trained to predict the masked pixel values given the visible ones. In the training phase, the masked regions are randomly selected from

TABLE I. Summary of major contribution of cyberbullying detection in Images.

Authors	Algorithms/ Techniques/ Models	Limitations
Vijaya kumar et al.	CNN, ReLU activation function	Custom based CNN used for feature extraction
Hoati Zhong et al.	Latent Dirichlet Allocation, pre-trained CNN,SVM classifier	feature extraction using image captioning
P.K Roy and F U Mali	2D-CNN,Transfer learning using VGG16, InceptionV3	Unable to predict textual bullying detection in images
Homa Hosseinmardi et al.	N-gram , SVM classifier	Confined to instagram images only
Mahmoud Elmezain et al.	Hybrid model, SVM classifier	Unable to predict textual bullying detection in images
Nishanth Viswamitra et al.	Multimodal classifier	low level image features, manual features selection

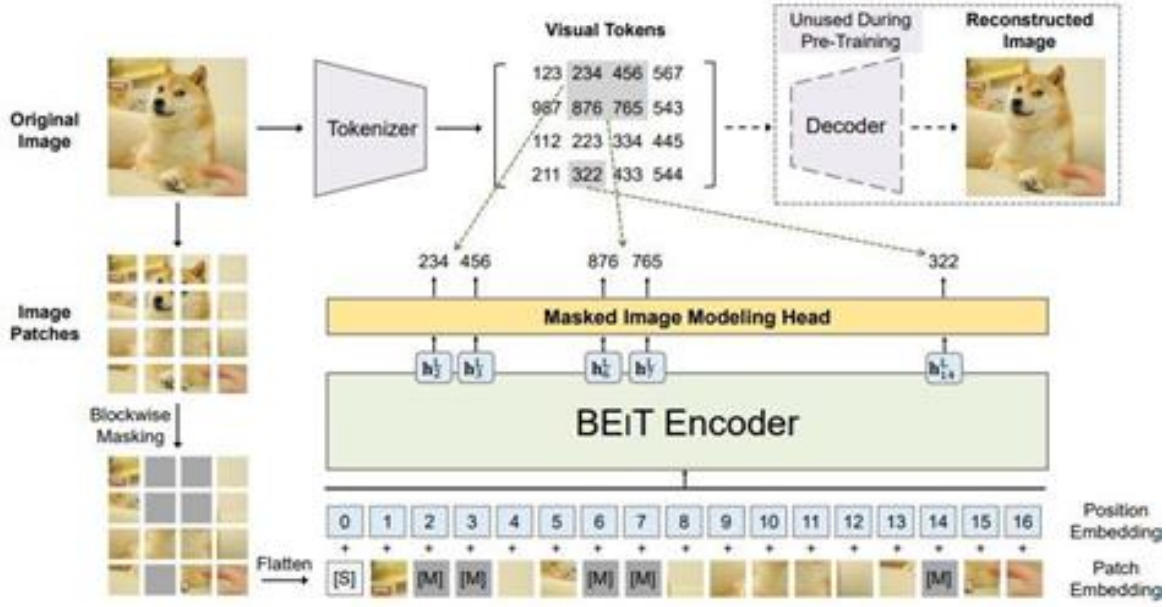

Figure 2. Overview of BEiT pre-tarining architecture[19]

the images (marked as [M] in Figure 2), and the corresponding visible regions are used as inputs to the network. The network is trained to minimize a reconstruction loss between the predicted and ground-truth-masked regions. Bidirectional Encoder for Image Transformers (BEiT) is a transformer-based architecture that can be thought of as a modified version of the transformer architecture used in natural language processing tasks, adapted for use in image recognition tasks.

BEiT uses a bidirectional encoder that processes both the image patches and their contextual relationships, producing a sequence of hidden representations that capture both local and global features of the image.

B. *Image captioning*

To supply more additional features from the input image, we consider the image-captioning concept from the input images to the neural network. We employed VGG16 [20] and LSTM [21] networks for image captioning. Figure 3 block diagram for image captioning. The VGG16 model is

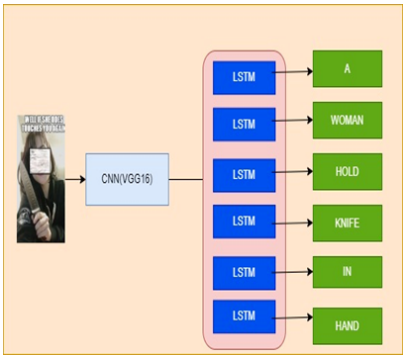

Figure 3. Block diagram VGG16+LSTM for Image Captioning.

a deep convolutional neural network that is highly effective

in identifying objects within images due to its ability to capture important image features through multiple layers of convolutions and pooling. The model consists of 16 layers, with 13 convolutional layers and 3 fully connected layers, and includes dropout layers to prevent overfitting during training. The model produces a feature vector representation of the image which is then fed into an LSTM layer for sequence modeling. This combination of VGG16 and LSTM layers is commonly used in image captioning tasks, where the model generates natural language descriptions of the contents of an image. A RoBERTa [22] architecture was employed to generate text features of the Image caption from the image.

C. Text extraction from Images

After that, we employed the Tesseract API[23] to extract text overlayed on images. The extracted text is passed to RoBERTa to generate text features of the extracted text.

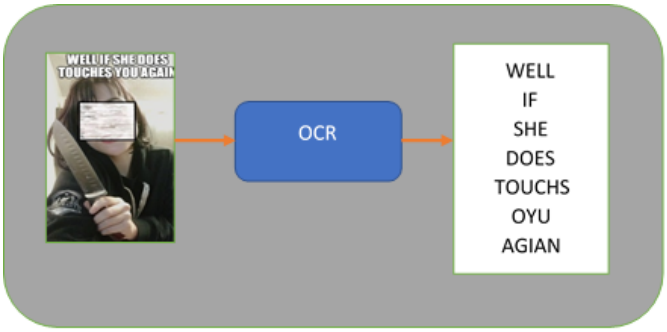

Figure 4. Text extraction using OCR

Figure 4 shows text retrieved from an image using OCR.

D. Phase-2 (Multi-Layer Perceptron)

In Phase-2, we combine all the features extracted from three architectures (BEiT, IC, and OCR) using late fusion [24] which is fed onto the MLP network.

Figure 5 shows the MLP network used in the proposed technique. The MLP network starts with 2034 neurons as the input layer which were features generated from BEiT, IC, and OCR. We have added 500, 100, 50, and 10 neurons as hidden dense layers for weight updating. The MLP network starts with random initialization of the process of the weight. We applied Xavier Glorot's [25] initialization for random initialization of weights for better convergence of weights.

Finally 2 neurons as the output layer for the classification of bullying and non-bullying images. We have used Leaky-Rectified Linear Unit (Leaky-ReLU) as an activation function, where 'L' is for weight updating . The Leaky-ReLU activation function avoids dying ReLU problems in the training process of the neural network. The formula for the Leaky-ReLU activation function is shown in Equation 1.

$$f(y) = max(0.01.y, y) \quad (1)$$

Here the function returns y if it receives a positive input value and it returns a really small value which is 0.01 times y when y is negative. At the last layer of the MLP network, we chose the Sigmoid activation function 'S' as it is a binary classification task. The Sigmoid activation function can be computed as shown in Equation 2.

$$\sigma(y) = \frac{1}{1 + e^{\ y}} \quad (2)$$

Here y is the input to the sigmoid function and 'e' is Euler's

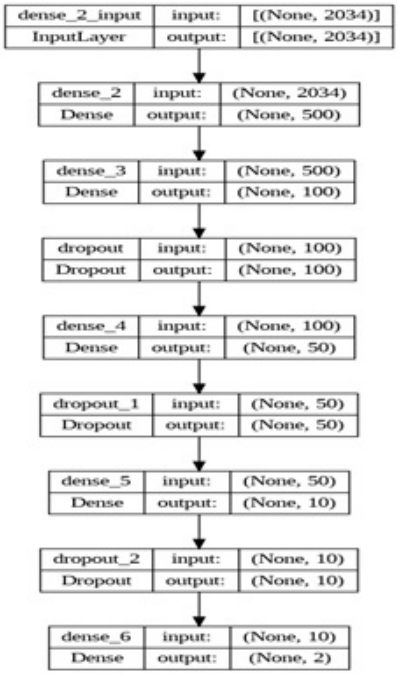

Figure 5. Multi-Layer Perceptron Network

constant (2.781).

4. Results and Discussion

A. Environment Specifications

Training and fine-tuning of deep learning models and pre-trained architectures require very high-end computing power for the parallel processing of tasks.

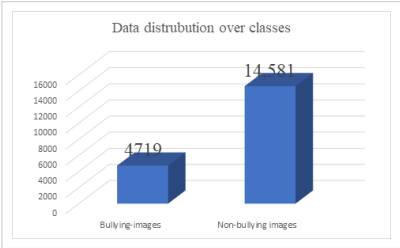

Figure 6. Class-wise data distribution

In this regard, we used Google Colab Pro to run the programs, which is a cloud-based platform with

Figure 7. Sample of Cyberbullying images

Figure 8. Sample of Non-cyberbullying images

Tensor Processing Units(TPUs) and Graphical Processing Units(GPUs). For the local environment, we used the Windows 11 operating system, 16GB of Random Access Memory. Tensorflow and PyTorch libraries were used in Python programming.

B. Dataset collection

We considered 19,300 images for experiments. The dataset was prepared by Nishant Viswamitra et al.[19] . These images were collected from various social media platforms like Facebook, Instagram and Twitter. Finally, 14,581 images were annotated as non-bullying images which are labelled as '0' and 4719 images were annotated as bullying images which are labelled as '1'. Figure 6shows class-wise data distribution. The images contains facial expressions of persons, hand gestures and objects to expressing bullying in the form of images. Figure 7 and 8 shows examples of cyberbullying images and non-cyberbullying images in the dataset.

C. Performance evaluation metrics

We considered accuracy, precision, recall, and f1-score to evaluate the proposed CNBD technique. Accuracy This has been the primary and most effective metric that has been utilized in classification algorithms. The ratio between the number of accurately predicted cyberbullying images and the total number of detected cyberbullying images. It is possible to refer to it as in Eq.3

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Where TP: (True Positive) means that the model predicts the image is bullying when the actual image is bullying. TN: (True Negative) means that the model predicts the image is non-bullying when the actual image is non-bullying. FP: (False Positive) means that the model predicts the image is bullying when the actual image is non-bullying. FN: (False Negative) means that the model predicts the image is non-bullying when the actual image is bullying.

Precision Precision is defined as the ratio of the number of accurately identified instances of bullying to the total number of identified instances of bullying. It is represented by Eq.4

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall The recall is a metric that counts the number of bullying images retrieved from the entire dataset of bullying images and is calculated using Eq.5

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

F1-Score The F1-score is the harmonic mean of both

recall and precision and is given as fowls in Eq.6:

$$\text{F1-score} = 2 \frac{precision \quad recall}{precision + recall} \tag{6}$$

First, all the input images are sent to the data preprocessing

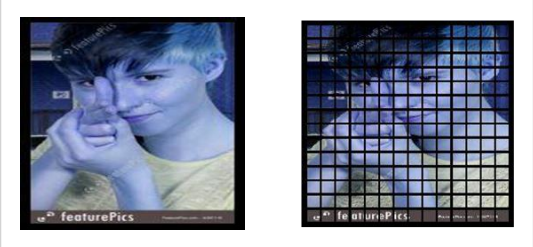

Figure 9. Original image is converted into Patches.

stage hence the images had different formats of sizes and structures. In the data preprocessing state we have applied data augmentation techniques of Normalization and Resizing and Rescaling. Finally, all the input images are set to a height and weight of 224*224 pixels size with RGB channels (Red, Green, and Blue).

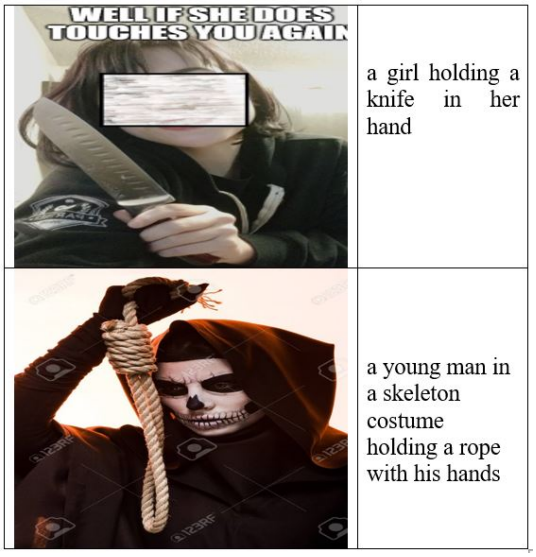

Figure 10. Example of .
caption generated from input images

A 224*224*3 is the input image for the BEiT feature extractor. The BEiT feature extractor works based on masked image modeling like the BERT model (masked language model). Each image is converted into 16*16 patches as shown in Figure 9. Some parts of the patches are masked randomly and flatten the image patch into a vector. Now these patches are passed to the BEiT encoder and finally reconstructed the image using tokens. BEiT was designed as 12 layers neural network to the reconstructed original image. The 12-layer generated 768 dimension feature vector which is input into the MLP network.

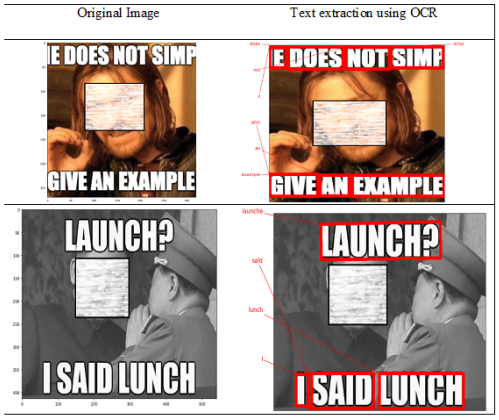

Figure 11. Example of Text extraction from Input Images using OCR.

In similar way, the input image passed to VGG16 and LSTM for image captioning. Figure 10 shows the caption captured by the network for given input images. Similarly, the input image is passed to OCR for text extraction on the images. Figure 11 shows the text extracted from input images using OCR.

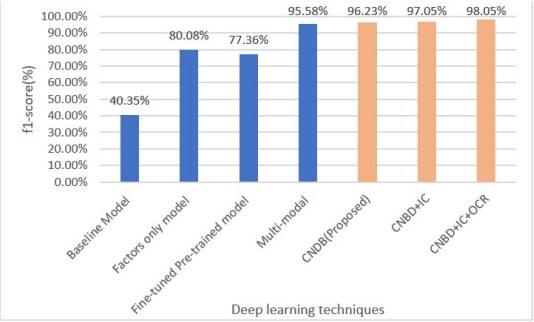

Figure 12. Comparison of f1-score of CNBD technique with existing methods

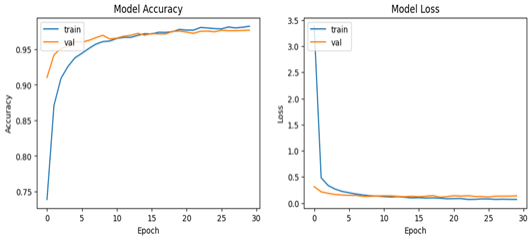

Figure 13. Accuracy and loss of proposed CNBD with IC+OCR.

The image dataset is divided into three parts: train set (70%), test set (20%), and validation set (10%). We trained the CNBD network with various hyperparameter of BEiT architecture such as learning rate 2*e-5, number of epochs 30, and weight decay 0.001. The proposed technique CNBD outperforms existing techniques in terms of accuracy, precision, and recall as shown in Table 2. Figure 12 shows the

TABLE II. Comparison of CNBD technique with existing methods

Classifier Model	Accuracy	Precision	Recall
Baseline Model [18]	77.25%	63.00%	29.68%
Factors only model [18]	82.96%	79.34%	80.84%
Fine-tuned Pre-trained model [18]	88.82%	81.40%	73.70%
Multi-modal [18]	93.96%	94.27%	96.93%
CNBD technique(Proposed)	**96.30%**	**96.16%**	**96.30%**
CNBD + IC	**97.50%**	**97.12%**	**97.05%**
CNBD+IC+OCR	**98.23%**	**98.05%**	**98.05%**

f1-score of the proposed technique with existing models.

The accuracy and loss functions concerning the epochs are shown in Figure 13. It can be observed that the loss associated with train data is lower compared to the loss of test data as expected since the test data is not seen during the training phase. It can also be observed that the accuracy showed a positive trend as we increase the training epochs. This demonstrates the adaptive nature of the network to our downstream task.

5. Conclusion and Future Enhancements

As the usage of social media platforms continues to grow, so too does the prevalence of negative online behaviors like cyberbullying, online hate speech, and trolling. Consequently, there is a growing need to explore effective ways to detect and address these harmful activities. One important aspect of this is the detection of cyberbullying on social media, which presents a particular challenge due to the diverse forms in which it can manifest, including text, images, and multimedia content. We proposed a technique named CNBD for cyberbullying detection in images. The proposed technique was evaluated using the metrics of accuracy, precision, and recall. The experimental results show that the proposed method with Image Caption features and OCR text features can improve results compared to the existing techniques with an accuracy of 98.23%, precision of 98.05%, and recall score of 98.05%. As part of the future, we consider cyberbullying detection for multi-media data such as text and images, videos, and regional language specific such as Telugu, Tamil, and Hindi.

A. Authors and A liations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

References

[1] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," *Cyberbullying Research Center*, vol. 14, pp. 206 – 221, 2010.

[2] S. A. L. M. Kowalski RM, Giumetti GW, "Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth," *Psychological Bulletin*, vol. 140, pp. 1073 –1137, 2014.

[3] G. Mishna, Mona Kassabri and Joanne, "Risk factors for involvement in cyberbullying: Victims, bullies, and bully-victims." *Children and Youth Services Review*, vol. 70, pp. 274–282, 2016.

[4] P. Roy and F. Mali, "Cyberbullying detection using deep transfer learning," *Complex Intelligent Systems*, vol. 8, pp. 5449–5467, 2022.

[5] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. O. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network." *Association for the Advancement of Artificial Intelligence*, 2015.

[6] H. Zhonga, H. Li, A. Squicciarini, R. majer, C. Griffin, Miller, and CorneliaCaragea, "Content-driven detection of cyberbullying on the instagram social network." *In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI'16).*, vol. 8, p. 3952–3958, 2016.

[7] E. Mahmoud, A. Malki, IbrahimGad, and E.-S. Atlam, "Hybrid deep learning model–based prediction of images related to cyberbullying." *International Journal of Applied Mathematics and Computer Science*, vol. 32, pp. 324–333, 2022.

[8] C. Rui, Lee, R. Ka-Wei, C. Wen-Haw, and J. Jiang, "Prompthate: Prompt-based hateful meme classification with pre-trained language models." *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 321–332, 2022.

[9] Aggarwal, T. Sharma, Yadav, Agrawal, Singh, Mishra, and Gritli, "Two-way feature extraction using sequential and multimodal approach for hateful meme classification," *IEEE Access*, vol. 9, pp. 121 962–121 973, 2021.

[10] K. R. Prajwal, C. V. Jawahar, and P. Kumaraguru, "Towards increased accessibility of meme images with the help of rich face emotion captions," *Proceedings of the 27th ACM International Conference on Multimedia*, p. 202–210, 2019.

[11] T. Tiwary and R. P. Mahapatra, "An accurate generation of image captions for blind people using extended convolutional atom neural network." *Multimedia Tools and Applications*, vol. 82, p. 3801–3830, 2022.

[12] Al-Malla, Jafar, and Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, vol. 9, 2022.

[13] Efrat, I. Malkiel, and L. Wolf, "Caption enriched samples for improving hateful memes detection." *Conference on Empirical Methods in Natural Language Processing*, 2021.

[14] Y. Zhou and Z. Chen, "Multimodal learning for hateful memes