Multimodal Cyberbullying Detection Using Deep Learning Techniques

Shaik Rafi[1], Ranjita Das[2], Kalyanam Jahnavi Sai Priya[3], Bolla Lakshmi Varsha[3], Velchuri Bala Harshitha[3], Sunkari Kavya[3], T.G. Ramnadh babu[4], and Sireesha Moturi[5]

[1] Asst.Professor, Dept of Computer Science and Engineering
Narasaraopeta Engineering College (Autonomous),
Narasaraopet 522601, Palnadu District, Andhra Pradesh, India.
shaikrafinrt@gmail.com

[2] Asst.Professor, Dept of Computer Science and Engineering
National Institute of Technology Agartala,
Barjala-799046, Jirania, West Tripura, Agartala, Tripura, India.
ranjita.nitm@gmail.com

[3] Dept of Computer Science and Engineering
Narasaraopeta Engineering College (Autonomous),
Narasaraopet 522601, Palnadu District, Andhra Pradesh, India.
kjahnavisaipriya@gmail.com, varshibolla2507@gmail.com,
vharshitha738@gmail.com, kavyasunkari510@gmail.com

[4] Asst.Professor, Dept of Computer Science and Engineering
Narasaraopeta Engineering College (Autonomous),
Narasaraopet 522601, Palnadu District, Andhra Pradesh, India.
baburamnadh@gmail.com

[5] Assoc.Professor, Dept of Computer Science and Engineering
Narasaraopeta Engineering College (Autonomous),
Narasaraopet 522601, Palnadu District, Andhra Pradesh, India.
sireeshamoturi@gmail.com

Abstract. Cyberbullying detection is the process of classifying and recognizing cyberbullying activity, which includes using technology to harass or threaten people—usually via online platforms. In order to address this, we examined a publicly available dataset that was classified as bully or non-bully according to text, image and image-text. We next proposed applying a deep learning model to recognize cyberbullying based on multimodal data. The VGG16 pre-trained model detects bullying in photos, while the XLM-RoBERTa with BiGRU model detects bullying in text. By combining these models (VGG16 + XLM-RoBERTa and BiGRU) with attention processes, CLIP, feedback mechanisms, CentralNet and other tools, we created a model for detecting cyberbullying in image-text based memes. Our final model showed that the algorithm is able to identify most cyberbullying occurrences with a decent accuracy of 74%.

Keywords: Cyberbully · VGG16 · XLM RoBERTa · BiGRU.

1 Introduction

In the cutting-edge technology, social media platforms and virtual era are getting used by number of people. Numerous services, along with WhatsApp, Instagram, Facebook and Twitter, are liable to cyberbullying. Cyberbullying is the use of digital era to annoy, harass, or threaten someone. Recognizing and responding to cyberbullying is vital to strengthen the protection against online bullying.

Bullying through images and text [1] are two forms of cyberbullying. The act of sending threatening, intimidating messages to another person via social media, messaging, or email is known as text-based bullying. The consequences of such an act can result in serious psychological damage and emotional distress. Image bullying occurs when someone posts offensive images or improper images in online without permission. It can be very embarrassing and very damaging to a person's reputation and sense of worth. Both of these bullies can have negative consequences and that's the point. So we need to work together to prevent and end online bullying, especially when it comes to young people.We have to establish a welcoming, respectful and secure on-line network for everyone.

Since bulk of modern studies focused on text-based situations, image-based cyberbullying has not got attention as text-based cyberbullying. This is an excessive issue as several messages on social media contain both text and images. To address this, we are going to buid a multi-modal [2] that is capable of detecting bullying in both images and texts. Our main goal is to create a pleasant and secure online environment where each person may thrive with out fear of abuse or intimidation. Even though it is a challenging endeavor, our intention is to make internet users lives better.

With the use of many modalities, our research attempts to detect cyberbullying in image-text posts. To extract the necessary attributes from images, deep learning models use convolutional neural networks (CNNs) [3,4] and transfer learning techniques. The study additionally employs Ensemble XLM-RoBERTa with BiGRU and VGG16, to enhance the detection of text and image based cyberbullying on social media platforms.

The rest of the paper is organized as follows: Literature Survey is discussed in Section 2 and Section 3 presents the Methodology. Whereas Section 4 highlights about Experimental Setup, Results are displayed in Section 5 and conclusion in Section 6.

2 Literature Survey

The growing number of social media platforms resulted to a corresponding rise in cyberbullying, prompting numerous studies to focus on efforts aimed at detecting cyberbullying. Some research on the detection of cyberbullying is discussed in this section.

2.1 Works on Text Data

A model that uses a machine learning algorithm to identify cyberbullying in literature was created by Varsha Reddy et al. [5]. They gathered postings and comments from other social media sites, including Facebook and Twitter. With an accuracy of 80.01%, the Logistic Regression model outperformed all other tested models. Using a dataset of 47,692 tweets, Mohammad Alshehri et al. [6] created a model utilizing CNN and LSTM. Their proposed model's accuracy of 96% demonstrates its effective ability to identify cyberbullying.

2.2 Works on Image Data

Using deep transfer learning models, Pradeep Kumar Roy et al. [7] created a model for the detection of cyberbullying in images. The 3000 image dataset was created by hand, with 1458 photos classified as bullies and 1542 images classified as non-bullies. They employed many deep transfer learning models, and InceptionV3 performed well, obtaining an accuracy rate of 89%. Using a combination of methods, Nishant Vishwamitra et al. [8] employed VGG16 and MLP to create a model that can identify cyberbullying in images. After 19,300 photos were used to train the model, its accuracy exceeded 93.36%.

2.3 Works on Multimodal Data

Using data from Facebook, Instagram and Twitter, Subbaraju Pericherla et al. [9] created a model that focuses on visuals and text features in social media images. They used OCR, VGG16 and LSTM to extract text and BEiT and MLP were used to create a CNBD model that had an accuracy of 98.23%. The CNN-based multi-model deep learning system developed by Kumari et al. [10] achieved 71% accuracy for bullying posts and 64% accuracy for non-bullying posts. They produced a dataset comprising 2100 posts, of which 1418 were deemed to be bullying and 619 to not be. While most studies focus on text and graphics, multimodal material also includes photos, videos and audio.

The Multi-Modal Cyberbullying Detection (MMCD) framework was introduced by Quingyu Xiong et al. [11]. It employs models such as TF-IDF, HAN, BiLSTM and Visual Embedding to detect cyberbullying content. Using Vine data, the machine obtained 83% accuracy. 86% accuracy rate on Instagram data shows that it can identify cyberbullying. The ability to identify cyberbullying content over a wide linguistic spectrum has improved with research. Using VGG16 and BiLSTM, Nahida Akter et al. [12] developed a hybrid deep neural network model to detect cyberbullying content in Bengali memes. With an accuracy of 87%, their investigation produced the Bengali Memes Dataset, which consists of 600 memes associated with bullying and 600 unrelated memes.

3 Methodology

This section covers preprocessing methods and models for text and image data. Additionally, we demonstrated our approach for detecting cyberbullying in multimodal data and the hyperparameters that the multimodel uses.

3.1 Data Pre-processing

Image Pre-processing The pre-processing techniques applied to the image data are as follows. These pre-processing methods helps in enhancing the data, which raises the model's performance.

- **PIL Library :**
  For opening, editing, and storing images as well as for image enhancements, the Python Imaging Library (PIL) is used for a variety of image-related tasks. It assists in locating and removing improperly formatted images.
- **Data Augmentation and Normalization (Using Image- DataGenerator) :**
  Data augmentation can improve model performance by expanding the size of the data. This is accomplished by using ImageDataGenerator class methods like brightness range, rotation range, zoom range, horizontal flip and width shift range etc. Pixel values must be normalized for deep learning models. This entails using a 1./255 rescaling factor to convert the pixel values from [0, 255] to [0, 1]. Specifically, this ensures data consistency with model predictions whether transfer learning or pre-trained models are used. The augmented image is shown in Fig. 1.

Fig. 1: Image Augmentation

– **Image Resizing :**
Images are scaled to a specified objective dimension, such as (224, 224), before being fed into the neural network to ensure input dimension consistency for pre-trained models such as VGG16. This step is crucial for models like VGG16, which might not work as well with images of varied sizes.

Text Pre-processing Pre-processing techniques are necessary for text generation models to ensure that the data is suitable for deep learning applications. The steps in this process are loading, cleaning, entering the data into models and these are covered in more detail below.

– **Data Loading and Cleaning :**
We import the dataset using the pandas package's pd.read csv function. As part of data cleansing, duplicate values, null values, and superfluous columns must be eliminated. It is possible to re-move unwanted columns with df.drop(). df.dropna() can be used to remove rows that have missing values, preventing incorrect predictions from being made as a result of incomplete data. Duplicate rows can result in overfitting, hence df.drop duplicates() aims to remove them. After pre-processing, the sample size was reduced from 5865 texts to 5749 texts, of which 3149 texts were bullies and 2600 texts were not.
– **Label Encoding :**
Label encoding is crucial when handling classification problems involving categorical features. The deep learning method converts labels into a numeric format, such as 0 and 1, to make them simpler for the system to understand.This is done by using the LabelEncoder function of the scikit-learn toolkit.
– **Tokenization (Text Preprocessing) :**
Tokenization is a technique used by deep learning models to convert raw text into numerical representations. The XLMRober-taTokenizer splits input text into subword units to make code-switching between languages or managing complicated morphological languages easier. After tokenizing the text, it may now be used for training and testing models.
– **Train-Test Split :**
To assess a model's performance, the data is divided into train and test sets using a predetermined procedure. This separation helps in evaluating the performance of the modal accurately.
– **Label Conversion to Tensor :**
Labels are converted to PyTorch tensors before training the model, much like tokenized text conversion, to make sure they are in the correct format.

3.2 Image Based detection

In order to identify cyberbullying in photographs, we have experimented with a number of deep learning models on the image data. We trained the images using

ResNet50, EfficientNetB0, EfficientNetB4, and VGG16, using various hyperparameters. Fig. 2 shows comparisons and the accuracy of each model. Using our data, the VGG16 pre-trained model performed better than any other model. We

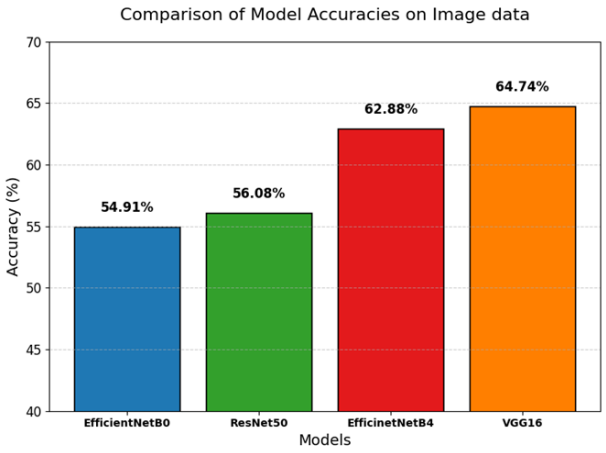

Fig. 2: Comparisons Of Image Models

employed a range of hyperparameters, such as the Adam optimizer, Dropout, Global Average Pooling, Dense layers etc., to optimize the training. Fig. 3 depicts the process clearly. After training the model across several epochs, we have also optimized it using a lower learning rate for better results.

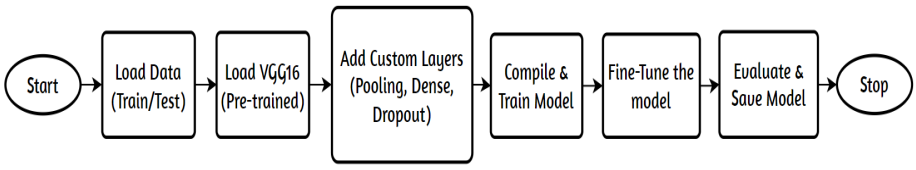

Fig. 3: Flow diagram of Image Modal

3.3 Text Based detection

Similar to the image training, we trained several models on the text data before deciding on the best-performing model. The models that we employed are mBERT, mBERT with BiGRU, XLM-RoBERTa, and XLM-RoBERTa paired

with BiGRU. When all of these models are compared, Fig. 4 demonstrates that the XLM-RoBERTa ensemble with BiGRU had the best performance.

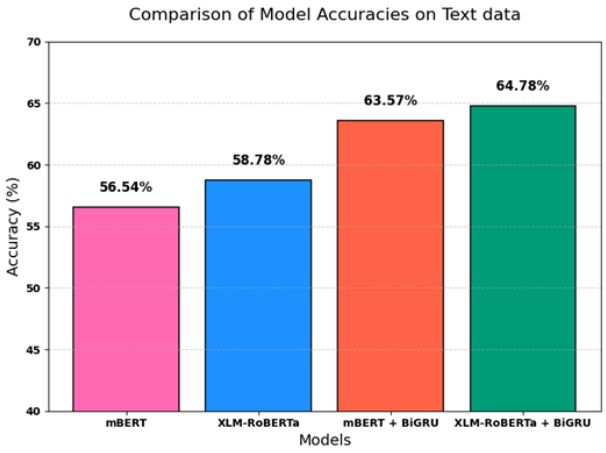

Fig. 4: Comparisons Of Text Models

The XLM-Roberta language model is integrated with a special design that includes a transformer layer to help with context understanding and a special kind of memory layer (bidirectional GRU), as shown in Fig. 5. We used a number of novel strategies, including batch normalization and dropout, to ensure that the model maintained its capacity to generalize and didn't become overly skilled at fitting the training set.In the end, we used AdamW, a smart optimizer that gradually changes the learning rate, to modify the training process and raise the model's learning efficiency. With these algorithms and architectural features combined, the model performs very well in its role, making it a dependable answer for text categorization problems.

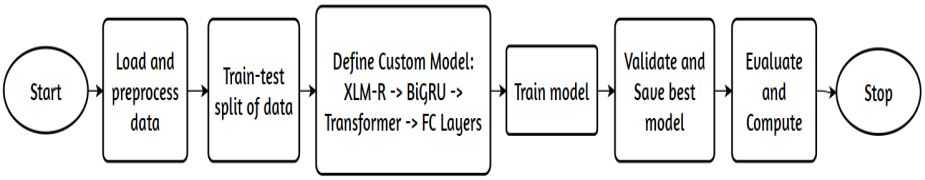

Fig. 5: Flow diagram of Text Modal

3.4 Multi-modal detection

After training the data independently, it's time to integrate the models that perform best with the text and image data. We created a new model by combining the pre-trained VGG16 model (used for images) with the XLM-RoBERTa and BiGRU model (used for text) using a number of strategies, including Intermodal Attention, Feedback Mechanism, CentralNet, CLIP Approach, and Fully Connected Layers. The entire flow of the model is displayed in Fig. 6. The model was trained using 80% of the available data (text and images), and its performance was verified using additional data. Hyperparameters, which are specified prior to the model's training phase, regulate the learning process and model architecture. The hyper parameters used in our proposed model to identify cyberbullying content in multimodal data is shown in Table 1.

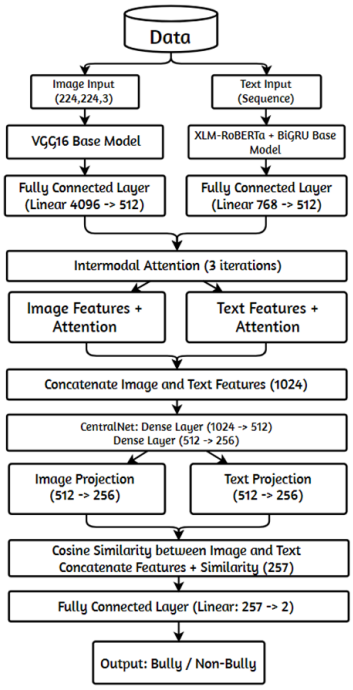

Fig. 6: Multimodal data detection

Table 1: Hyperparameters used in the model

Component	Hyperparameter	Value
Image Model (VGG16)	Dropout Rate	0.5
	Fully Connected Layer	512 (from 25088)
Text Model (XLM-RoBERTa)	Fully Connected Layer Output Size	512 (from 768)
	Image Projector	256 (from 512)
CLIP projectors	Image Projector	256 (from 512)
	Text Projector	256 (from 512)
Intermodal Attention	Embedding Dimension	512
	Number of Attention Heads	8
Final Classifier	Output Size	257 (256 + 1)
	Number of Classes	2
Training Settings	Batch Size	32
	Learning Rate	1e-5
	Weight Decay	0.01
	Early Stopping Patience	3
	Loss Function	Cross Entropy
	Optimizer	AdamW

4 Experimental Setup

4.1 Dataset Description

The code-mixed text and image posts from Sriparna Saha et al.'s [13] 2022 dataset served as the basis for our investigation. The dataset used in this study is

publicly available and can be accessed here for the images -https://drive.google.com/drive/folders/1vCgjNvgVFDl3SVsGmyxx3urJOgYeLIjs and here for the text data -https://docs.google.com/spreadsheets/d/1tD5yqGZ3TlDjeUFThautfZGegHrRz7FW/edit?gid=1650123160#gid=1650123160. A wide range of topics are covered by the dataset, such as harm-fulness, emotion, sarcasm, and cyberbullying. Since we are dealing with the detection of cyberbullying, we only choose the cyberbullying section. Among those the crucial columns are image name, text, image label, image-text label, and image-text label. More complicated models can be developed by combining textual and visual data to improve categorization accuracy and robustness. The dataset contains 6,006 posts in total, which are separated into bully and non-bully categories. Because there are so many images, a detailed analysis of bullying and non-bullying is possible. The dataset has been reduced to 5798 by pre-processing, with 3,193 bully images and 2,605 non-bully images. This made it possible to distribute data fairly and train and evaluate models in an effective manner.

4.2 Data Preparation

Images Data To achieve strong performance metrics and avoid over-fitting, the image-based model has an 80-20 split between the train and test parts. Of all the images, 20% are in the test directory (1160) and 80% are in the train directory (4638). The train directory further splits images into bully and non-bully classes to provide clear class identification. Out of the 4638 photos, 2,554 are associated with the bully class and 2,084 with the non-bully class. The test directory does not need to be further divided because we just utilize it for testing and validation. To get information, however, we counted the number of images that included and excluded bullies.The test directory has 1160 images total—639 images of bullies and 521 images of non-bully. Table 2 displays the data's overall distributions.

Table 2: Distribution of images

Split	**Images**	**Bully**	**Non Bully**
Train	4638	2554	2084
Test	1160	639	521
Total	**5798**	**3193**	**2605**

Text Data The dataset we examined had annotations like sentiment and sarcasm. In order to make cyberbullying the main topic, we removed unneeded

columns and selected text-related columns. The Img_Text and Img_Text_Label columns were chosen in the dataset. Img_Text is the text that was used to train the model, and Img_Text_Label is the class of the text that has been classified as either a bully or a nonbully. Furthermore, nearly equal percentages of texts in the textual data are texts that are bully and texts that are non-bully, which helps the model fit both kinds of data.

5 Results

We evaluated many models on text and image data and selected models with the best performance individually. We picked these best models and coupled them with a range of techniques, including feedback mechanisms, fully linked layers, CLIP projectors, Intermodal Attention, and CentralNet, to produce a modal that fits on multimodal data. More than a thousand instances are used to train this modal. After evaluating our model using test data, we obtained an overall accuracy of 74%. The model accuracy for each set of data is shown in Table 3.

Table 3: Overall Results

Data	**Model**	**Accuracy**
Image	VGG16	64.74%
Text	XLM-Roberta + BiGRU	64.78%
Multimodal (Image + Text)	VGG16 + XLM-Roberta + BiGRU	74%

6 Conclusion

In this study, we introduced a model that can identify bullying in multimodal data—that is, both text and images. We used a publicly accessible multimodal memes dataset for this, which contains images with text and these are labeled as either bully or non-bully. After experimenting with many models, we discovered that the models that perform the best on images and text are VGG16 and XLM-RoBERTa + BiGRU. Finally, we combined these models with other deep learning techniques, such as CentralNet, CLIP, Intermodal Attention, Feedback mechanism, etc. for multimodal detection, which improved our model's ability to identify bullying in multimodal data.

References

1. Debnath, Dipanwita & Das, Ranjita & Rafi, Shaik. (2022). Sentiment-Based Abstractive Text Summarization Using Attention Oriented LSTM Model. 10.1007/978-981-16-6624-7_20.
2. Rafi, S., Das, R. Topic-guided abstractive multimodal summariza- tion with multimodal output. Neural Comput & Applic (2023). https://doi.org/10.1007/s00521-023-08821-5
3. Greeshma, B., Sireesha, M., Thirumala Rao, S.N. (2022). Detection of Arrhythmia Using Convolutional Neural Networks. In: Shakya, S., Du, KL., Haoxiang, W. (eds) Proceedings of Second International Conference on Sustainable Expert Systems . Lecture Notes in Networks and Systems, vol 351. Springer, Singapore. https://doi.org/10.1007/978-981-16-7657-4_3
4. S. L. Jagannadham, K. L. Nadh and M. Sireesha, "Brain Tumour Detection Using CNN," 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2021, pp. 734-739, doi: 10.1109/I-SMAC52330.2021.9640875
5. A. Varsha Reddy, Gugulothu Kalpana, N. Satish Kumar, and Dr. Sundaragiri Dheeraj, “Cyber Bullying Text Detection Using Machine Learning”, June 2022 , Available: https://doi.org/10.22214/ijraset.2022.44157
6. Alabdulwahab, Aljwharah & Haq, Mohd Anul & Alshehri, Mohammed. (2023). Cyberbullying Detection using Machine Learning and Deep Learning. International Journal of Advanced Computer Science and Applications. 14. 424-432. 10.14569/IJACSA.2023.0141045.
7. Pradeep, Kumar, Roy., Fenish, Umeshbhai, Mali. (2022). Cyberbullying detection using deep transfer learning. Complex & Intelligent Systems, 8(6):5449-5467. doi: 10.1007/s40747-022-00772-z
8. Vishwamitra, Nishant & Hu, Hongxin & Luo, Feng & Cheng, Long. (2021). Towards Understanding and Detecting Cyberbullying in Real-world Images. 10.14722/ndss.2021.24260.
9. Pericherla, Subbaraju & Ilavarasan, E.. (2024). Overcoming the Chal- lenge of Cyberbullying Detection in Images: A Deep Learning Approach with Image Captioning and OCR Integration. International Journal of Computing and Digital Systems. 15. 393-401. 10.12785/ijcds/150130.
10. Kumari, Kirti & Singh, Jyoti & Dwivedi, Yogesh & Rana, Nripendra. (2020). Towards Cyberbullying-free social media in smart cities: a unified multi-modal approach. Soft Computing. 24. 10.1007/s00500- 019-04550-x
11. Wang, Kaige & Xiong, Qingyu & Wu, Chao & Gao, Min & Yu, Yang. (2020). Multi-modal cyberbullying detection on social networks. 1-8. 10.1109/IJCNN48605.2020.9206663.
12. Ahmed, Md.Tofael & Akter, Nahida & Rahman, Maqsudur & Islam, Abu & Das, Dipankar & Rashed, Md. Golam. (2023). Multimodal Cy- berbullying 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.
13. Maity, Krishanu & Saha, Sriparna & Bhattacharyya, Pushpak. (2022). A Multitask Framework for Sentiment, Emotion and Sarcasm aware Cyberbullying Detection from Multi-modal Code-Mixed Memes. 1739- 1749. 10.1145/3477495.3531925.