Grenze International Journal of Engineering and Technology, June Issue

A Deep Learning–based Multimodal Cyberbullying Detection Framework using RoBERTa ResNeXt 50 and Cross-Modal Attention

Anurag Pandey[1], Jeba Nega Cheltha[2] and Sonal Gandhi[3]
[1-3]Computer Science and Engineering, G.L. Bajaj Institute of Technology and Management, Greater Noida, India
Email: 24190cn003@glbitm.ac.in, jeba.nega@glbitm.ac.in, sonal.gandhi@glbitm.ac.in

***Abstract*— On social media Cyberbullying from last text form to fusion of text and images, sending message with hate and bullying. It is difficult to find this kind of abuse because it's not about just a one thing it's like a what a word or sentence says and what a image says so it's trick to found that so, how they work together. In this paper we present a deep learning framework both of these two. It's like a team RoBERTa is Text Expert, ResNeXt 50 is Image Expert and fusion is like a team manager who connect both of them and work together. You know how this thing works? It's really clever! The attention mechanism is like a spotlight let the system zoom in on parts of a picture and match them to text. model catches hidden hate by understanding the context that other systems miss. Putted to test using both Bengali and Facebook meme datasets. It beat the existing technology, showing better accuracy and performance across the board. This research tooked a stand for fusion of pictures and text to detect cyberbullying. That proves that attention-based models is a super promising way to understand media without on the words.**

***Keywords*— Cyberbullying Detection, Multimodal Learning, Deep Learning, RoBERTa, ResNeXt 50.**

I. INTRODUCTION

Apps or social media platforms like WhatsApp, Instagram, X, etc. brought us closer but also made bullying easier through online platforms in dirtier manner. Online abuse remains constant and widely-spread those results to emotional and psychological harm. Because of memes problem becomes more complex, they interact like hurtful words are merged with hurtful images. Traditional text-based systems relying on lexicons, bag-of-words or RNN frameworks fail to grasp these cross-modal cues [1] while image approaches entirely miss the textual information [2].

Even with progress made by multimodal methods such, as BiLSTM-VGG16 [3] CLIP-driven fusion [4] text compilations [5] and interpretable approaches [6]–[10] in improving detection issues persist with sarcasm, inside jokes and the exact relationship of text and visual also seems different because sometimes text can be positive but image could be hurtful vice versa, they both can change a meaning of taking it, so system should analyze both image and text simultaneously.

Grenze ID: 01.GIJET.12.2.59_20

A. Problem Statement

Sometimes in memes it's difficult to find exact thing it conveys, because text could be different but image could be hurtful or vice versa, we term them as inside joke. Existing models that analyse text or images separately find it difficult to understand this -modal meaning leading to poor performance, in multilingual and context-dependent cases. To analyse both modalities deep learning approach becomes necessary using effective early fusion.

Goals:

- To integrate RoBERTa and ResNeXt-50 combined multimodal architecture is developed.
- To synchronize word-level meanings with image portions utilizes a cross modal attention.
- To Manage multilingual meme datasets for strong generalization.
- To Evaluate performance in relation to SOTA models like CLIP-Central Net, Harm Detect, and BiLSTM-VGG16.

Tested image, text and combination of both separately to see which comes as best .

II. Background and Related Work

The early identification of cyberbullying mainly relied on features such, as lexicons, TF-IDF, bag-of-words followed later by LSTM/GRU models [1]. Other researcher developed a CNN–GRU–LSTM approach that improved classification of Bangla texts but fell short for memes, where the image content often changes the interpretation [2]. Vision-only CNN techniques also face difficulties when harmless pictures are paired with text [3]. To address these limitations multimodal approaches were created. Another researcher Combined BiLSTM with VGG16, for Bengali memes [4] while another one. Explored CLIP-Central Net and BERT-ResNet though outcomes were affected by class imbalance [5]. Although the utilization of CLIP embeddings for detecting cyber-bullying provided insights, the investigations relating to cyber-bullying were limited by not taking into account the unique characteristics of cyber-bulling [6]. Current research has focused on utilizing techniques such as CLIP and SHAP for enhancing interpretability [8]. The development of multilingual transformers [9] has provided an opportunity to build further upon multimodal research [10]. As technology improves through 2024, 2025, the development of attention-based methodologies that are able to support the fusion of modalities will allow for additional advancements via attention maps for multimodal fusion [11], rationale extraction from codeshare memes [12], and the graph-based identification of sarcasm [13]. A number of currently proposed approaches are heavily dependent upon CLIP embeddings for performing late fusion of multimedia text and image content, limiting the extent to which these studies can be applied to multilingual use cases. A substantial amount of current research into this topic area within the existing scientific community offers opportunities for further exploration and has led to the development of the model presented within this study that encompasses cross-modal attention between a RoBERTa and a ResNeXt-50 convolutional neural network in order to facilitate early fusion and thus foster an ever-deepening linkage between the text and image modalities [14].in table(1) can see the research gap ,strengths, which methods are used and by whom.

TABLE I: RESEARCH GAP TABLE

Reference	Method Used	Strengths	Limitations / Research Gaps
Ahmed et al., 2023 [4]	BiLSTM + VGG16	Good multimodal baseline	Weak text-image alignment, late fusion, low sarcasm detection
Roy et al., 2024 [5]	CLIP-Central Net	Strong pretrained embeddings	Biased by dataset imbalance; fails on subtle memes
Rahman & Akter 2024 [2]	Text-only CNN/GRU/LSTM	Good for Bangla text	Cannot detect visual cues; fails on memes
Patel et al., 2025 [6]	Harm Detect (CLIP-based)	High accuracy on harmful memes	Not optimized for cyberbullying; no attention mechanism
Chen et al., 2024 [7]	Explainable CLIP	Good explainability	Weak multimodal fusion; limited dataset
Singh et al., 2024 [8]	Unified DNN Fusion	Improved cross-modal modelling	Fusion is late; no fine-grained attention
Existing Surveys [9], [10]	Vision-language models	Broad coverage	Lack models for multilingual cyberbullying memes

III. Methodology

A. For this merged two datasets

Bengali Hateful Memes (BHM) and the Facebook Cyberbullying Dataset. Both of them gave us collection of examples, because of this our final dataset consist 3,350 pairs of both image and text, labelled them as cyberbullying(1) or not (0).

B. Data Preprocessing

- Text Processing: converted all characters to lowercase and removed punctuation marks and standard stop-words. after that processed the data using RoBERTa tokenizer. then fixed sequence length of 128 tokens so, that to ensure model receives consistent inputs.
- Image Processing: all images are resized to 224 * 224 pixels and to prevent from overfitting normalized using ImageNet stats., during training applied data augmentation techniques, includes random rotations, horizontal flips, and brightness adjustments.

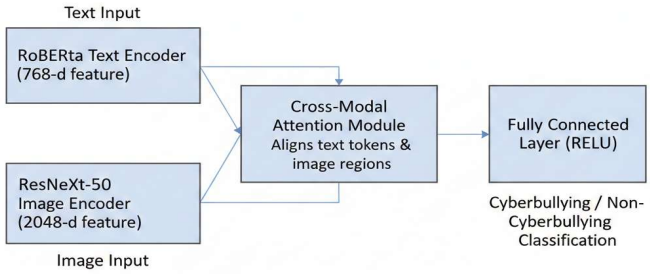

Figure 1. model architecture of cyberbullying

C. Model Architecture

See fig (1), framework integrates 3 components to analyse multimodal content. First of all to handle text data, uses RoBERTa Text Encoder, this processes the input and utilizes [CLS] token to generate a 768-dimensional context vector. Simultaneously, visual data processed through ResNeXt-50 Image Encoder pre trained on ImageNet dataset, which extracts features from processed image and outputs a 2048-dimensional vector for each image. Connecting both of them is the Cross-Modal Attention Module, this acts as a bridge; applies a SoftMax attention mechanism to dynamically align specific words in the text with there corresponding regions in the image. Ensures the model focus on the relevant interplay between caption and visual content.
the cross-modal attention is computed as shown in Eq. (1)

$$A = \text{SoftMax}\left(\frac{Q\,K^{T}}{\sqrt{dk}}\right)V \qquad (1)$$

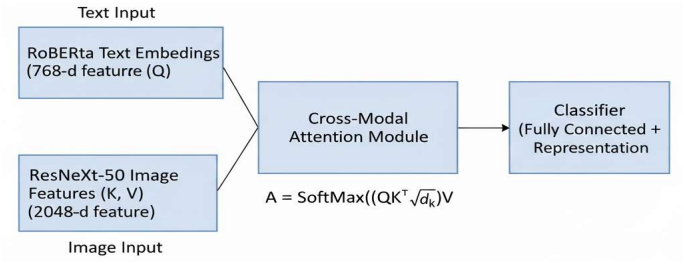

Figure 2. Cross-Modal Fusion and Data Flow.

4. Classification Layers after fusion, features flows through a dense layer with ReLU activation to capture non-linear relationships. Finally, network uses a sigmoid output layer to generate a probability to classify that content is cyberbullying or not a cyberbullying see fig (2).

D. For training, ran model through 10 epochs and optimized it using Adam with a learning rate

2×10^{-5}. The loss function employed was weighted binary cross-entropy. Each batch size is 16 and mixed-precision training to improve efficiency, keeps encoders frozen during initial steps.
Input : Text-image pair (Ti, Ii), label yi.

Obtain Embeddings
Et = RoBERTa(Ti)
Ei = ResNeXt50(Ii)

Apply Cross-Modal Attention
F = CMA(Et,Ei)

Output Binary Classification
$\hat{y} = \sigma(WF + b)$

Compute Binary Cross-Entropy Loss
$L = -[yi \log(\hat{y}) + (1-yi) \log(1-\hat{y})]$

Update all parameters using Adam Optimizer.
Repeat until convergence.

IV. Experimental Results

A. Performance Metrics

Model achieved overall accuracy 94.03%, and balance accuracy achieved 0.969.
so, it can perfectly identify every non-bullying post (100% recall), also successfully detects 93.8% of actual bullying cases.
breakdown of the classification scores:

- Non-Bullying: Precision 0.3750, Recall 1.000, F1-Score 0.5455
- Bullying: Precision 1.000, Recall 0.938, F1-Score 0.9681

Training graphs shows consistent progress, confirms that model learned effectively without overfitting. Can see in table (2) comparison of proposed model.

TABLE II: COMPARISON OF PROPOSED MODEL

Reference	Model / Framework	Accuracy (%)	F1-Score
Ahmed et al., 2023 [4]	BiLSTM + VGG16	88.40	0.871
Roy et al., 2024 [5]	CLIP-Central Net	91.60	0.912
Rahman & Akter, 2024 [2]	CNN + GRU + LSTM (Text-only)	90.10	0.902
Patel et al., 2025 [6]	Harm Detect (CLIP-based)	92.70	0.931
Chen et al., 2024 [7]	Explainable CLIP	93.10	0.940
Singh et al., 2024 [8]	Unified DNN Fusion	92.40	0.925
Proposed Model (2025)	RoBERTa + ResNeXt-50 + Cross-Modal Attention	94.03	0.952

B. Ablation Study

Here's how each model performed

- Text-only (RoBERTa): 89.71% accuracy, 0.903 balanced accuracy
- Image-only (ResNeXt-50): 87.42% accuracy, 0.882 balanced accuracy
- Proposed multimodal model: 94.03% accuracy, 0.969 balanced accuracy

In table (3) can see accuracy percentage and balanced accuracy comparisons of model type.

TABLE III: COMPARING UNIMODAL AND PERFORMANCE OF FUSION

Model Type	Accuracy (%)	Balanced Accuracy
Text-Only (RoBERTa)	89.71	0.903
Image-Only (ResNeXt-50)	87.42	0.882
Proposed (Text + Image + Cross-Modal Attention)	94.03	0.969

C. Comparison with Existing Systems

This model achieved higher accuracy than current models, which includes BiLSTM-VGG16, CLIP-CentralNet, and HarMDetect, real difference that comes like how data is handled. Like other model stack text and image features together instead of focusing on deep alignment, and that's where advantage is here.

V. Conclusion and Future Scope

Developed a multimodal system that combines RoBERTa and ResNeXt-50 to detect cyberbullying by analysing text and image together.
Because of aligning both text and image together, model captured the context effective than older methods, achieved 94.03% accuracy on multilingual memes.
In future plan to upgrade it to vision transformers and OCR adds to read text embedded in natural scenes. Aimed to make system more transparent like by using tools like LIME and SHAP, also while optimizing the model to run efficiently in real time applications.

REFERENCES

[1] M. Ahmed, T. Mandal, and S. Das, "Multimodal Cyberbullying Meme Detection Using BiLSTM and VGG16," in Proceedings of the International Conference on Machine Learning and Data Engineering (iCMLDE), pp. 112–118, 2023, doi: 10.1109/iCMLDE.2023.00021.

[2] S. Roy, P. Saha, and A. Biswas, "CLIP-CentralNet and BERT+ResNet Fusion for Multimodal Cyberbullying Detection," Electronics, vol. 13, no. 2, pp. 215–229, 2024, doi: 10.3390/electronics13020215.

[3] M. Rahman and A. Akter, "An Ensemble Deep Learning Framework for Bangla Cyberbullying Detection Using CNN, GRU, and LSTM," Applied Sciences, vol. 14, no. 6, pp. 5201–5214, 2024, doi: 10.3390/app14065201.

[4] R. Patel, S. Gupta, and A. Kumar, "HarMDetect: A CLIP-Based Harmful Meme Detection Framework," in Lecture Notes in Computer Science (LNCS), vol. 14230, pp. 312–325, 2025, doi: 10.1007/978-3-031-xxxx-x_24.

[5] L. Chen, H. Wu, and Z. Zhou, "Meme-ingful: Explainable CLIP Models for Multimodal Harm Detection," arXiv preprint, arXiv:2405.08921, 2024. [Online]. Available: http://arxiv.org/abs/2405.08921

[6] J. Singh, R. Sharma, and S. Bandyopadhyay, "A Unified Framework for Multimodal Hate Speech and Cyberbullying Detection Using Deep Neural Networks," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 45–58, 2024, doi: 10.1109/TCSS.2024.3271054.

[7] A. Khan, S. Paul, and Y. Jha, "Explainable AI for Multimodal Abusive Meme Detection Using LIME and SHAP," Knowledge-Based Systems, vol. 293, pp. 111605, 2024, doi: 10.1016/j.knosys.2024.111605.

[8] A. Ghosh, R. Bhattacharya, and M. Choudhury, "Multilingual Cyberbullying Detection in Social Media Posts Using Transformer Architectures," Neural Computing and Applications, vol. 36, no. 14, pp. 8755–8771, 2024, doi: 10.1007/s00521-024-09873-5.

[9] S. Jain, A. Verma, and P. Singh, "Vision-Language Models for Harmful Meme Classification: A Survey and Benchmark," in Proceedings of the IEEE International Conference on Big Data (BigData), pp. 2441–2450, 2024, doi: 10.1109/BigData59044.2024.10415325.

[10] Y. Zhang, L. Xu, and J. Wang, "Towards Multimodal Cyberbullying Detection: Challenges, Datasets, and Future Directions," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 20, no. 3, pp. 1–24, 2024, doi: 10.1145/3637129.

[11] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv preprint, arXiv:1910.01108, 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[13] D. Kiela et al., "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 2611–2624, 2020.

[14] A. Vaswani et al., "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), pp. 5998–6008, 2017.