Meme-ingful Analysis: Enhanced Understanding of Cyberbullying in Memes Through Multimodal Explanations

Prince Jha[1*], **Krishanu Maity**[1*], **Raghav Jain**[1*], **Apoorv Verma**[1], **Sriparna Saha**[1] and **Pushpak Bhattacharyya**[2]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna
[2] Department of Computer Science and Engineering, Indian Institute of Technology Bombay
princekumar_1901cs42@iitp.ac.in *

Abstract

Internet memes have gained significant influence in communicating political, psychological, and sociocultural ideas. While memes are often humorous, there has been a rise in the use of memes for trolling and cyberbullying. Although a wide variety of effective deep learning-based models have been developed for detecting offensive multimodal memes, only a few works have been done on explainability aspect. Recent laws like "right to explanations" of General Data Protection Regulation, have spurred research in developing interpretable models rather than only focusing on performance. Motivated by this, we introduce *MultiBully-Ex*, the first benchmark dataset for multimodal explanation from code-mixed cyberbullying memes. Here, both visual and textual modalities are highlighted to explain why a given meme is cyberbullying. A Contrastive Language-Image Pretraining (CLIP) projection-based multimodal shared-private multitask approach has been proposed for visual and textual explanation of a meme. Experimental results demonstrate that training with multimodal explanations improves performance in generating textual justifications and more accurately identifying the visual evidence supporting a decision with reliable performance improvements.[1]

Disclaimer: The article contains profanity, necessary for the nature of the work, but not reflecting the authors' opinions.

1 Introduction

The tremendous increase in multimodal content due to the widespread use of social media platforms renders human moderation of such information untenable (Cao et al., 2020). Memes, which are images with tiny text descriptions embedded in them, have become a popular kind of multimodal content on social media in recent years. Though memes are typically humorous, it also stimulates the propagation of online abuse and harassment, including cyberbullying. Cyberbullying (Smith et al., 2008) is any communication that disparages an individual on the basis of a characteristic such as color, gender, race, sexual orientation, ethnicity, nationality, or other features. The Pew Research Center estimates that 40% of social media users have encountered online harassment or bullying[2] (Chan et al., 2019). Cyberbullying victims may experience despair, worry, low self-esteem, and even suicidal thoughts (Sticca et al., 2013). Automatic cyberbullying detection techniques with the model's explainability are highly required to minimize those unpleasant consequences.

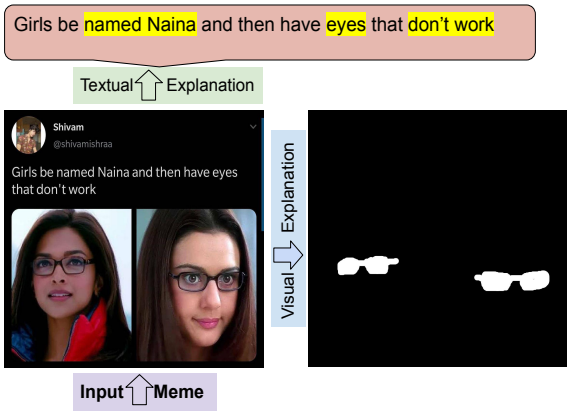

Figure 1: Cyberbullying Explanation in memes. Here the aim is to highlight both the image and text as an explanation of why the given meme is a bully.

Motivation and Evidence: Over the last decade, studies on cyberbullying detection have focused primarily on textual content (Agrawal and Awekar, 2018; Dadvar et al., 2014; Paul and Saha, 2020) and, recently on memes (Kiela et al., 2020; Pramanick et al., 2021; Maity et al., 2022a) in monolingual

* The first three authors contributed equally to this work and are jointly the first authors.

[1]https://github.com/Jhaprince/MemeExplanation

[2]https://www.pewresearch.org/internet/2017/07/11/online-harassment-2017/

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 930–943

setting, with limited research focusing on code-mixed language. The use of code-mixed languages in different social media and message-sharing apps proliferates rapidly in multilingual countries (Rijhwani et al., 2017). Code Mixing is a linguistic phenomenon where words or phrases from one language are inserted into an utterance from another language (Myers-Scotton, 1997). However, those researchers mostly concentrated on improving the performance of detecting offensive posts using various deep learning models without giving any insight or analysis into the explainability. Consequently, we propose a novel problem called **Multimodal Explanation of Code-Mixed Cyberbullying Memes (MExCCM)**. This task involves processing multimodal inputs and aims to generate both textual and visual explanations for multimodal cyberbullying memes.

Research Gap: Till now, most of the works on offensive memes are limited to classification tasks. In the explainability aspect, there are some works on text data only (highlighting the words or phrases in a sentence) (Mathew et al., 2020; Karim et al., 2021) and only one work on multimodal memes (internal layers' attention weight visualization) (Hee et al., 2022). Still, there is no work where both text and images are highlighted to justify the offensiveness of cyberbullying content like a human does. Thus, to mitigate the above-mentioned research gap, we aim to build a deep learning-based model that can explain cyberbullying nature of memes in both visual and textual modalities. We seek this idea from semiotic textology linguistic theory (García-Valero, 2020), which includes three subcomponents in order to consider how each textual media derives meaning; *dictum* (aka denotation), *evocatum* (aka connotation), and *apperceptum* (mental images), the latter one embodying the vision-grounded analysis of textual content.

Contributions: Our contributions are threefold: (i) We present *MExCCM*, a novel task for generating multimodal explanations for code-mixed cyberbullying memes, a first in this field. (ii) We introduce *MultiBully-Ex*, the first multimodal explainable code-mixed cyberbullying dataset. It includes manual highlighting of both text and image modalities in a meme to demonstrate why it is considered bullying (iii) We propose an end-to-end Contrastive Language-Image Pretraining (CLIP) approach for visual and textual meme explanation, aiming to encourage more research on code-mixed data.

2 Related Works

Cyberbullying is very reliant on linguistic subtlety. Researchers have recently provided a lot of attention to automatically identifying cyberbullying in social media. In this section, we will review recent works on the detection and explainability aspects of cyberbullying.

Detection: Researchers have made significant strides in detecting meme-based cyberbullying and offensive content. Maity et al. (2022a) created MultiBully, a Twitter and Reddit dataset in code-mixed language, proposing two multitask platforms for detecting bullying memes, sentiment, and emotion. Pramanick et al. (2021) extended the HarMeme dataset and developed a deep multimodal network to detect harmful memes, focusing on COVID-19 and US politics. Other notable works include Kiela et al. (2020)'s hate speech detection with 69.47% accuracy using Visual-BERT, Gomez et al. (2020)'s MMHS150K dataset of 150K Tweets, and Suryawanshi et al. (2020)'s MultiOFF dataset for identifying offensive meme content, which showcased a fusion method combining text and image modalities.

Explainability: LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) have been used to advance both textual and visual explainability in machine learning models. Zaidan et al. (2007) improved sentiment classification by employing human-annotated "rationales." Mathew et al. (2020) introduced *HateXplain*, finding that training with human rationales reduced bias. Karim et al. (2021) developed an explainable hate speech detection in Bengali, highlighting crucial words. Hee et al. (2022) visualized how VilBERT and VisualBERT models captured slurs in hateful memes, discovering the image modality's significant contribution. While most studies used explainability to justify model outputs, our task uniquely focuses on explainability as the output itself, specifically designed to offer textual and visual explanations for cyberbullying memes. This represents the first effort to generate *MExCCM*.

3 Multimodal Bully Explanations Dataset *(MultiBully-Ex)*

To create *MultiBully-Ex*, we utilize *MultiBully* dataset[3] (Maity et al., 2022b), which includes 3222 bully and 2632 nonbully memes. We selected

[3] https://github.com/Jhaprince/MultiBully

this dataset because it is the only openly available meme dataset on cyberbullying in a code-mixed setting. Our work focuses on jointly extracting textual rationales (words or phrases) and visual masks (image segmentation) to localize salient regions to explain cyberbullying detection tasks. Hence we only considered the bully memes for further annotation.

3.1 Annotation training

The annotation was led by three Ph.D. scholars with adequate knowledge and expertise in detection and mitigation of cyberbullying, hate speech, and offensive content and performed by three undergraduate students with proficiency in both Hindi and English. First, ten undergraduate computer science students were voluntarily hired through the department email list and compensated through honorarium[4]. For annotation training, we required gold standard samples annotated with rationale labels. We aim to annotate the text explainability (rationales) part first, and then, based on those rationales, the visual annotation will be done. Our expert annotators randomly selected 150 memes and highlighted the words (rationales) for the textual explanation. Each word in a meme has been assigned a value of 0 or 1, where 1 represents that it is one of the rationales. Later expert annotators discussed each other and resolved the differences to create 150 gold standard samples with rationale annotations. We divide these 150 annotated examples into three sets, 50 rationale annotations each, to carry out three-phase training. After the completion of every phase, expert annotators met with novice annotators to correct the wrong annotations, and simultaneously annotation guidelines (refer Appendix C.1) were also renewed. After completing the third round of training, the top three annotators were selected to annotate the entire dataset.

3.2 Main Annotation

We used the open-source platform Docanno[5] deployed on a Heroku instance for main annotation where each qualified annotator was provided with a secure account to annotate and track their progress exclusively. We initiated our main annotation process with a small batch of 100 memes and later raised it to 500 memes as the annotators became well-experienced with the tasks. We tried to maintain the annotators' agreement by correcting some errors they made in the previous batch. On completion of each set of annotations, final rationale labels were decided by the majority voting method. If the selections of three annotators vary, we enlist the help of an expert annotator to break the tie. We also directed annotators to annotate the posts without regard for any particular demography, religion, or other factors. We use the Fleiss' Kappa score (Fleiss, 1971) to calculate the token level inter-annotator agreement (IAA) among multiple raters for the rationale detection task signifying the dataset being of acceptable quality. IAA obtained a score of 0.72 for the rationales detection task signifying the dataset being of acceptable quality.
Once annotators finished doing rationale annotations, they were further asked to highlight the visual regions that could justify the rationale annotations. Visual annotations were done using open source image segmentation UI interface label studio[6], where the annotator has to mark the regions of the image to generate a binary image where the highlighted portion having pixel value 1 and others are 0. Figure 1 shows an annotated sample from the *MultiBully-Ex* dataset. We assessed the inter-annotator agreement for visual annotation using the Dice coefficient, which is a measure of overlap between two annotations. To ensure the accuracy of the annotations, we first had a single annotator create them and then assigned the same annotation to another annotator. We then compared their annotations using the Dice coefficient. If the coefficient was greater than 0.5, we included the annotation from the first annotator. However, if the coefficient was less than or equal to 0.5, an expert annotator was consulted to make the annotation. It's noteworthy that the average number of tokens highlighted as 'bully' was 6.79. Conversely, the total average number of tokens for 'meme' amounted to 14.12. Additionally, we discovered that the total average percentage of the area covered by visual explanations within the meme was 35.18[7].

4 Methodology

Formulation of MExCM: Formally, given a meme (M) with textual modality $T = \{t_1, t_2, ..., t_n\}$ and visual modality $V \in R^{3\times W\times H}$, where W is the width and H is the height of an image, we intend

[4] refer to Appendix C.2 and Appendix C.3 for more details on cost and timeline

[5] https://github.com/doccano/doccano

[6] https://labelstud.io/

[7] refer Appendix C.4 for more details on dataset statistics

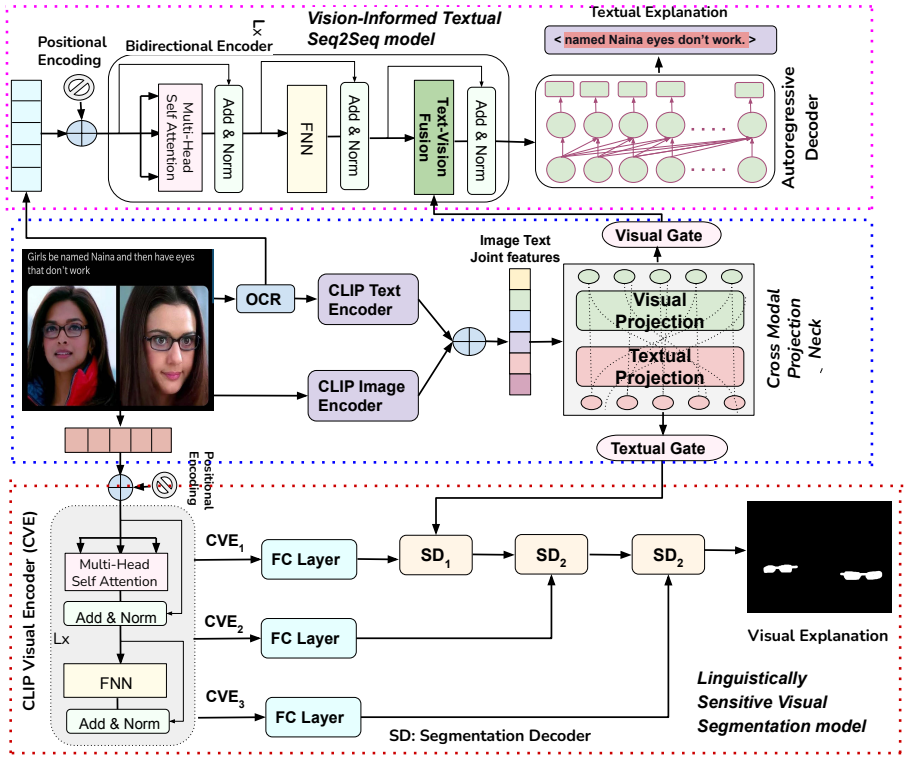

Figure 2: CLIP projection-based (CP) multimodal shared-private multitask architecture. The *Vision-Informed Textual Seq2Seq model* is represented by a pink dotted box. *The Cross Modal Projection Neck* is signified by a blue dotted box. The *Linguistically Sensitive Visual Segmentation model* is indicated by a red dotted box. Lx denotes number of transformer layers

to learn textual justification along with visual evidence which is defined as follow: **(1) Textual Explanation:** Textual explanation is the process of extracting pertinent rationales $R = \{r_1, r_2, .., r_k\}$ from the textual modality T of a meme M, which contributes to its classification as a cyberbullying instance. **(2) Visual Explanation:** Visual explanation involves a semantic segmentation task, the aim of which is to predict the segmented region $S \in R^{1\times W\times H}$ within the visual modality V. This segmented region is perceived as supporting evidence aligning with the textual justification.

Motivated from Liu et al. (2016), we propose a CLIP (Radford et al., 2021) projection-based (CP) multimodal shared-private multitask architecture. To enhance comprehension of our proposed method, we partition it into three distinct components: (1) CLIP Projection-Based Cross-Modal Neck, (2) Vision-Informed Textual Seq2Seq model, and (3) Linguistically-Sensitive Visual Segmentation model. In our design, the CLIP projection-based cross-modal neck acts as a shared layer, serving both the textual and visual explanation components. Meanwhile, we employ BART encoder and CVE (CLIP Visual Encoder) as private layers, enabling them to focus more effectively on their respective tasks. This decision to use separate task-specific encoders stems from our concern that a unified encoder's shared feature space might inadvertently contain task-specific features, potentially leading to unnecessary feature redundancy and a mixing of sharable features in the private space

4.1 CLIP Projection-Based Cross-Modal Neck

Our proposed CLIP projection-based Cross-Modal Neck acts as a common component bridging two task-specific networks: (1) the Vision-Informed Textual Seq2Seq model, and (2) the Linguistically-Sensitive Visual Segmentation model. We implement modality-specific gating mechanisms to manage the interplay of information between these textual and visual facets. The initial step in our process involves the acquisition of representations for each text-meme pair. This is facilitated by CLIP (Contrastive Language-Image Pre-training), a pre-trained model proficient in visual-linguistic tasks, which leverages its capabilities to encapsulate the holistic meaning of the meme. CLIP's effectiveness can be traced back to its extensive pre-training on 400 million image-text pairs harvested from

the Internet. This training process, driven by contrastive learning objectives, along with the breadth of imagery and natural language exposure, bestows upon CLIP's remarkable zero-shot performance. In this study, we use multilingual BERT for text encoding and the Vision Transformer for image encoding. We extract two core features from each meme: a CLIP visual feature, C_I, from the meme's image, M, and a CLIP textual feature, C_T, from its OCR-extracted text, T. Both these features, C_I and C_T, are represented as 512-dimensional vectors. After this, these two vectors (C_I and C_T) are concatenated to create a joint vector representation of both modalities which are fed into the following two gating mechanisms simultaneously:

4.1.1 Gated Visual Projection

Previous research (Zhang et al., 2018; Lu et al., 2018) highlights the infeasibility of correlating functional words, such as 'the,' 'of,' and 'well,' with any visual block. To address this, our approach includes a visual gate designed to dynamically calibrate the contribution of visual features. We also employ a cross-modal projection neck to transpose gated visual features into the space of a BART (or T5) encoder. The implementation of the cross-modal projection neck can be achieved via a transformer-based architecture, leveraging its capacity to enable global attention among input tokens. To facilitate this, we feed the visual encoding from CLIP into the transformer-based network, merging it with randomly initialized, learnable weights (RW). The integration of these learnable weights serves dual purposes. Firstly, it empowers the multi-head attention mechanism with access to valuable information from the CLIP embedding. Secondly, it enables the network parameters to adapt responsively to incoming information, thereby enhancing the system's ability to learn and evolve over time.

4.1.2 Gated Textual Projection

Recent literature (García-Valero, 2020; Jha et al., 2022) illustrates that several communicative aspects, including facial expressions, gestures, postures, spatial relationships, color schemes, and movement, are more accurately expressed via visual cues as compared to text-based communication. In response to these findings, our proposed model incorporates a textual gating mechanism that moderates the influence of textual features. Complementing this, we utilize a Feed Forward Network (FFN) to map these textual characteristics into the domain of the segmentation decoder. This integrated approach underscores the importance of both visual and textual elements, aligning with our overarching aim of developing a multimodal understanding of memes.

4.2 Vision-Informed Textual Seq2Seq Model

We introduce a module designed to generate explainable text, which harnesses visual understanding by employing a combination of CLIP-based gated visual projection and generative pre-trained language models (GPLMs), specifically BART and T5. The process begins with the tokenization of input text and its transformation into a sequence of embeddings, $X_t \in R^{N \times d_t}$, where N is the sequence length and d_t is the feature dimension. To preserve the positional information of these token embeddings, positional encodings, $E_{pos_t} \in R^{N \times d_t}$ are added elementwise. The resultant input Z_0, now encompassing the positional information, is channeled into our proposed vision-aware encoder.

This vision-aware encoder comprises three subcomponents: 1) Multi-head Self-Attention (MSA), 2) Feedforward Network (FNN), and 3) Text-Vision Fusion (TVF). Additionally, each sublayer is followed by a residual connection (He et al., 2016) and layer normalization (Ba et al., 2016). The MSA (Multi-head Self-Attention) and FNN (Feedforward Network) components of our model are standard transformer layers, designed to facilitate the processing of our input data.

CLIP visual features, C_I, and textual features, C_T, are processed through the Gated Visual Projection (GVP) (as defined in the previous section) to yield a controlled visual information $P_v \in R^{M \times d_t}$, where M is the projected sequence length with an embedding dimension of d_t.

$$P_v = GVP(C_I, C_T) \tag{1}$$

In the Text-Vision Fusion (TVF) component of our model, we employ two types of multimodal fusion mechanisms (refer Appendix A), namely dot product attention-based fusion and multi-head attention-based fusion as suggested in (Yu et al., 2021; Tsai et al., 2019). Formally, textual input $Z_t \in R^{N \times d_t}$ and gated visual input $P_v \in R^{M \times d_t}$ are fused to produce a vision-aware textual representation $F \in R^{N \times d_t}$ that has a same dimension as the textual input, which allows the continual stacking of layers.

4.3 Linguistically Sensitive Visual Segmentation Model

We introduce a transformer-based encoder-decoder model, inspired by the UNet architecture, that incorporates a novel gated textual projection mechanism (CP-UNet). This mechanism is designed to augment the representation capabilities of the encoder, thereby enhancing the overall efficacy of the model. Our encoder assembly includes a series of transformer layers based on the CLIP model, linked to the decoder via residual connections. The decoder is structured around a straightforward transformer-based architecture, leveraging the insights offered by the encoder to generate the final output. Formally, an input image $V \in R^{C \times W \times H}$ is processed through the CLIP visual encoder, resulting in a sequence of embeddings $X_v \in R^{P \times d_t}$, where P represents the projected sequence length with dimension d_t. To encapsulate spatial features from the visual information, we incorporate a positional embedding $E_{pos_v} \in R^{P \times d_t}$. The encoded representation is acquired by passing the input through a cascade of sub-layers, including MSA and FNN, succeeded by Layer Normalization. At each layer of the CLIP visual transformer, these encodings are captured and projected into the decoder's space. They are subsequently merged with the internal features of our decoder preceding each transformer block. The decoder is designed to match the number of transformer blocks extracted from the CLIP visual transformer. Importantly, the decoder inputs are modulated with a projected gated textual vector, facilitating a deeper comprehension of the linguistic context embedded in the input, thereby yielding more accurate and contextually aligned outputs.

4.4 Loss Prioritization

Inspired by Bengio et al. (2009), we introduce the concept of loss prioritization sequentially so that we can concentrate on a specific task on a priority basis. The basic hypothesis is that the cognitive process of *MExCCM* may not be entirely simultaneous. Both generation loss and segmentation loss must combine sequentially to achieve the desired output. We combine the loss function with a certain periodicity, i.e., after a given number of epochs $ep \in \{15, 20, 25\}$. The network initially learns its weight over a particular loss function (learning particular aspects of tasks), after which it self-tunes the weights over all loss functions combined sequentially (learning some other facets of the task). Mathematically, an overall global loss function, L^{ep}_{global} can be defined by the equation: $L^{ep}_{global} = L^{0.ep}_{i_0} + L^{1.ep}_{i_1}$ where L_{i_q}s are individual losses such that $i_q \in \{generation\_loss, segmentation\_loss\}$ and q can be non-negative integer, at a given periodicity of *ep* epochs. A regular cross-entropy loss is employed to calculate generation_loss and segmentation_loss.

5 Results and Discussion

For a fair comparison with proposed models, we have set up standard baselines such as BART (Lewis et al., 2019), T5 (Raffel et al., 2020), VG-BART, VG-T5 (Yu et al., 2021), and DeepLabv3 (Chen et al., 2017), MobileNetv3 (Howard et al., 2019), Fully Convolutional Networks (FCN) (Long et al., 2015), UNet (Ronneberger et al., 2015) for textual and visual explainability, respectively. Detailed explanations on baselines, evaluation metrics and training details are given in Appendix B). Our proposed model can be utilized in a single task (keeping one task-specific private layers) or multi-task (keeping both visual and textual private layers) settings. In single task setting, there is no gating mechanism.

5.1 Quantitative analysis

We have conducted a statistical t-test on the results of our proposed model and other baselines and obtained a p-value less than 0.05.

(i) Single Task: Unimodal models The performance of unimodal models is detailed in Table 2 (textual explanations) and Table 3 (visual explanations). T5-base and BART-base models outperform their larger counterparts, possibly due to overfitting from excessive parameters given the limited dataset size (3222 instances). For visual explanations, our CLIP-based UNet excels compared to baseline models using visual features from networks like ResNet, VGG19, AlexNet, etc., optimized for ImageNet, not memes. This superiority stems from CLIP's fine-tuning to better represent visual information through language supervision (Radford et al., 2021).

(ii) Single Task: Multimodal models Our proposed multimodal models use dot product attention-based fusion (A1) and multi-head attention-based fusion (A2) techniques, combined with gated visual projection. According to the results (see 2 and 3), our CLIP projection-based GPLMs outshine all

Table 1: Results of proposed multitask model for textual and Visual Explainability, A1: Dot-product attention, A2: Multi-head attention, CP-UNet: CLIP projection-based UNet, RW: Random weight, DC: Dice Coefficient, JS: Jaccard Similarity, mIOU: Mean Intersection over Union.

Model	Textual Explinability								Visual Explinability			
	ROUGE			BLEU				HE	DC	JS	mIOU	HE
	R1	R2	R-L	B1	B2	B3	B4					
CP-UNet-T5_A1	60.94	45.58	60.43	60.16	53.32	49.73	46.93	3.91	**68.72**	54.72	60.93	**4.37**
CP-UNet-T5_A1+RW	61.06	46.33	60.59	60.63	54.44	51.05	48.15	4.07	68.7	**54.76**	61.29	4.36
CP-UNet-T5_A2	61.46	45.63	61.07	60.86	54.55	50.93	47.33	4.31	68.32	54.11	60.82	4.28
CP-UNet-T5_A2+RW	61.67	45.28	61.21	61.75	55.24	51.39	47.82	4.34	68.38	54.42	59.93	4.29
CP-UNet-BART_A1	61.76	45.68	61.54	61.68	56.96	52.26	49.57	4.38	67.95	53.67	61.58	4.25
CP-UNet-BART_A1+RW	63.06	46.63	62.57	62.86	56.55	52.92	49.33	4.57	67.32	53.95	61.13	4.24
CP-UNet-BART_A2	62.91	46.93	62.57	62.44	56.51	53.03	49.21	4.42	67.03	53.69	**62.53**	4.21
CP-UNet-BART_A2+RW	**63.54**	**47.36**	**63.07**	**62.75**	**57.13**	**53.39**	**50.81**	**4.59**	67.19	53.03	62.29	4.23

Table 2: Results of different baselines and proposed Single task model for textual explainability

Model	ROUGE			BLEU				HE
	R1	R2	R-L	B1	B2	B3	B4	
Unimodal Baselines								
T5-base	59.97	44.01	59.61	60.48	53.7	50.03	47.14	3.67
T5-large	59.57	43.47	59.07	58.87	52.43	48.83	45.86	3.62
Bart-base	60.05	46.35	59.86	60.55	56.46	50.52	49.98	3.81
Bart-large	58.64	43.17	58.15	58.4	51.62	47.95	45.03	3.24
Multimodal Baselines								
VG-T5 (Dot-product)	60.2	44.08	59.7	59.26	52.75	47.57	46.52	3.85
VG-T5 (Multi-head)	60.85	44.97	60.11	60.89	56.99	52.87	49.29	3.93
VG-BART (Dot-product)	60.84	45.76	60.25	61.2	54.54	50.78	47.81	3.91
VG-BART (Multi-head)	61.17	45.37	60.8	60.37	53.99	50.52	47.57	4.26
Proposed models								
CP-T5_A1	60.04	43.12	59.32	59.55	52.87	49.02	46.15	3.81
CP-T5_A1+RW	60.15	43.56	59.55	59.74	53.11	49.59	46.72	3.83
CP-T5_A2	61.16	44.69	60.72	60.1	54.76	50.16	47.31	4.21
CP-T5_A2+RW	61.36	44.92	60.97	60.59	54.34	50.88	48.02	4.26
CP-BART_A1	61.71	45.98	61.27	62.17	55.55	51.73	48.88	4.27
CP-BART_A1+RW	62.37	46.51	62.06	62.53	57.09	53.55	50.85	4.32
CP-BART_A2	61.99	46.11	61.5	62.43	55.72	51.96	48.68	4.3
CP-BART_A2+RW	**62.33**	**46.49**	**61.85**	**62.44**	**55.9**	**52.14**	**48.69**	**4.39**

Table 3: Results of baselines and proposed Single task model for Visual explainability; V: Vision; L: Language; HE: Human Evaluation

Model	Visual Explainability			
	DC	JS	mIOU	HE
Unimodal Baselines				
DeepLabv3	38.85	24.92	32.25	1.79
MobileNetV3	39.49	25.49	32.16	2.07
FCN	39.21	25.29	31.97	2.12
UNet	41.89	27.35	31.79	2.41
Proposed Models				
(V) CP-UNet	65.71	51.86	**63.03**	3.83
(V+L) CP-UNet	**66.22**	**52.45**	62.95	**3.91**

other models. The top model, CP-BART_A2 + RW, notably improves over previous best models by up to 2.28 ROUGE-1, 0.14 ROUGE-2, and 3.7 ROUGE-L scores. Using visual embeddings from CLIP with randomly initialized learnable weights (+RW) significantly enhances performance in textual explainability tasks. The language-aware CLIP-UNet model outperforms its unimodal counterpart, with improvements up to 1.01 in DC and 0.59 in JS scores, and substantial margins over the previous best unimodal model. However, the enhancement by the language-aware variant is marginal, likely because CLIP embeddings are optimized for visual rather than textual information.

(iii) Multi-Task: Shared-Private Architecture As evidenced by the results presented in Table 1, 2, and 3, it can be observed that the CLIP projection-based multimodal shared-private multitask approach outperforms all single task baselines by a significant margin, thus supporting the notion that training with multimodal explanations leads to enhanced performance in the generation of textual justifications and more precise identification of visual evidence. Notably, our most effective multitask model, CP-UNet-BART_A2 + RW, which is optimized for text explanations, outperforms the best single-task textual explainability model (CP-BART) by 1.21 R1, 0.87 R2, and 1.22 R3. Additionally, the best multitasking model, CP-UNet-T5_A1 + RW, which is optimized for visual explanations, outperforms the single task visual explainability model (CP_UNet) by 2.48 DC, and 2.31 JS.

(iv) Human Evaluation (HE): We conducted a human evaluation to assess the quality of generated explanations from our proposed methods. *MEx-CCM* was evaluated based on the following criteria: **1 - Very Irrelevant:** The explanation does not address the topic or concept adequately. **5 - Very Relevant:** The explanation is highly relevant to the topic or concept. Our analysis revealed notable findings regarding the relevance of different models in various settings. Specifically, when considering unimodal approaches, our best language-based model, CP-BART_A2+RW, achieved an impressive average relevance score of 4.39 for textual explanations. On the other hand, our vision-based

model, CP-UNet, obtained an average relevance score of 3.91 for visual explanations. Moving on to the multitask setting, our model CP-UNet-BART_A1+RW demonstrated exceptional performance by achieving an average relevance score of 4.59 for textual explainability. Similarly, our model CP-UNet-T5_A1 excelled in providing relevant visual explanations, securing an average relevance score of 4.37 (refer to Table 1, 2 and 3 for more details).

5.2 Qualitative Analysis

Figure 3 presents results comparing visual and textual explainability of ground truth vs. model predictions.
(i) In the first meme, textual context ("boyfriend ke sath ka argument toh solve ho hi jayega...") is more vital than the visual. Both single-task and multi-task models identify the same number of correct rationales. The multi-task model better captures the visual aspect, representing the girl's face. (ii) In the second meme, visual cues surpass the textual message ("Happy Holi especially jo ghar pai hai"). The multi-task model more accurately identifies visual and textual cues than the single-task model. (iii) For the final meme, both modalities equally contribute to the meme's meaning. The multi-task model fares better, capturing most rationales with minor mistakes. Visually, the single-task model's prediction is less accurate than the multi-task model's.
From this qualitative analysis, we can conclude that (a) Multi-task-model is performing better than single task model, but visual explainability is still not convincing. More research is needed in this direction. (b) In cases where any one of the modalities dominates the others (example i and ii) single-task model performance is comparable to multi-tasking. (c) In cases where both modalities have an equal contribution, the multi-task model significantly performs better than the single-task model, which reveals that simultaneously learning both textual and visual explainability helps improve the performance of both tasks.

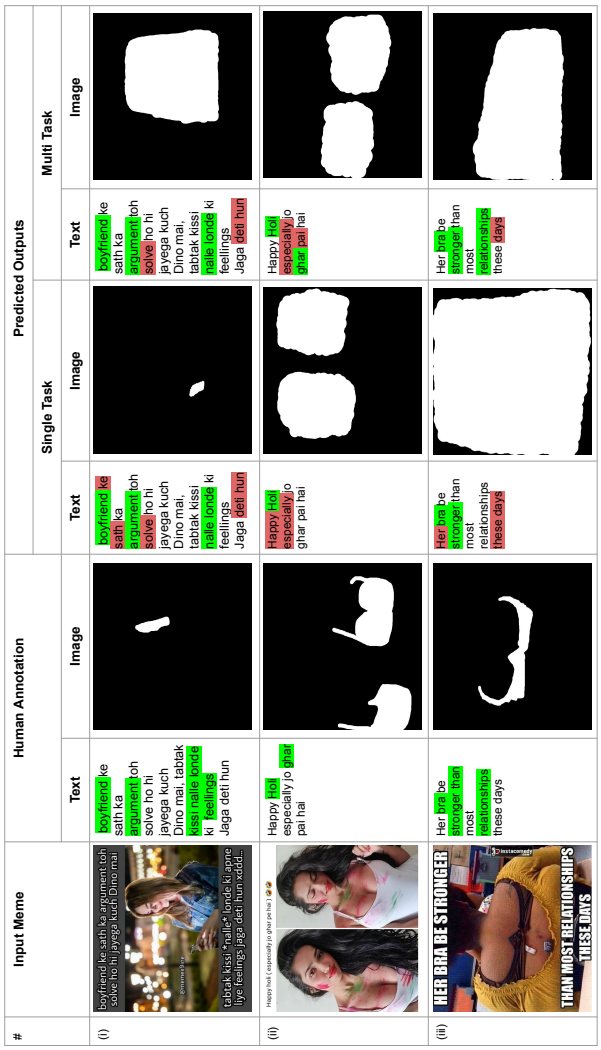

Figure 3: Human annotation vs. proposed model's visual and textual explanations; Green highlights indicate an agreement between the human annotator and the model. Red highlighted tokens are predicted by models, not by human annotators.

6 Conclusion and Future Work

To encourage more research on explainable meme cyberbullying detection in code-mixed language, we introduced *MultiBully-Ex*, manually annotated with textual and visual explanations. This work introduces a CLIP projection-based multimodal shared-private architecture to generate rationales (textual explainability) and binary segmented image maps (visual explainability). Experimental results demonstrate that multitask models outperform the single-task model by a significant margin. We showed examples where visual modality is more effective than textual ones and vice versa, supporting the idea that multimodal explainable models provide better insight than unimodal approaches. Future attempts will be made to develop methods for identifying stereotypes in cyberbullying memes to capture implicit content and training models on a diverse dataset to enhance the performance of visual explainability.

7 Limitation

We have proposed a shared-private multimodal multitask architecture and a new benchmark dataset, *MultiBully-Ex*, to improve the explainability of cyberbullying memes in code-mixed Indian languages. However, there are some limitations to this approach:
1) Specifically, the textual explainability of memes is limited to the lexical level, which precludes the detection of implicit cyberbullying or stereotypes.
2) One of the main limitations of our work is its lack of generalizability to other code-mixed languages such as English and Spanish. However, this limitation can be addressed by fine-tuning the model on other code-mixed languages, which will enable it to capture the cultural nuances of the language.
3) Additionally, the visual explainability aspect of our approach, which involves predicting binary segmentation maps, is susceptible to the center bias commonly observed in computer vision models. This can impede the correct identification of visual cues that support the textual explanations, particularly for objects or features located in the corners or edges of the image.
4) This study is specifically dedicated to the analysis and understanding of memes in this image and text-based format. It is essential to highlight that our research delves into the unique characteristics and communication potential of static memes, distinct from the analysis of dynamic video memes. The latter, involving audiovisual elements, falls beyond the scope of our investigation.

References

Sweta Agrawal and Amit Awekar. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *European conference on information retrieval*, pages 141–153. Springer.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *12th ACM conference on web science*, pages 11–20.

Tommy KH Chan, Christy MK Cheung, and Randy YM Wong. 2019. Cyberbullying on social networking sites: the crime opportunity and affordance perspectives. *Journal of Management Information Systems*, 36(2):574–609.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Maral Dadvar, Dolf Trieschnigg, and Franciska de Jong. 2014. Experts and machines against bullies: A hybrid approach to detect cyberbullies. In *Canadian conference on artificial intelligence*, pages 275–281. Springer.

Lee R Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Benito García-Valero. 2020. Borreguero zuloaga, m. and vitacolonna, l.(eds.), the legacy of jános s. petőfi. text linguistics, literary theory and semiotics. *Journal of Literary Semantics*, 49(1):61–64.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3651–3655. ACM.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. 2019. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.

Paul Jaccard. 1901. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat*, 37:241–272.

Prince Jha, Gaël Dias, Alexis Lechervy, Jose G Moreno, Anubhav Jangra, Sebastião Pais, and Sriparna Saha. 2022. Combining vision and language representations for patch-based identification of lexico-semantic relations. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4406–4415.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022a. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1739–1749. ACM.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022b. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sayanta Paul and Sriparna Saha. 2020. Cyberbert: Bert for cyberbullying identification. *Multimedia Systems*, pages 1–8.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4439–4455. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1971–1982.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Fabio Sticca, Sabrina Ruggieri, Françoise Alsaker, and Sonja Perren. 2013. Longitudinal risk factors for cyberbullying in adolescence. *Journal of community & applied social psychology*, 23(1):52–67.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Michele L Ybarra, Kimberly J Mitchell, Janis Wolak, and David Finkelhor. 2006. Examining characteristics and associated distress related to internet harassment: findings from the second youth internet safety survey. *Pediatrics*, 118(4):e1169–e1177.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. *arXiv preprint arXiv:2109.02401*.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

A Attention Mechanism

A.1 Dot Product Attention Based Fusion

In this type of fusion mechanism, we begin by projecting the visual features to the same dimensional space as the textual features (Eq. 2). Then, the dot-product is calculated, and the softmax function is applied (Eq. 3). Finally, the input textual features are combined with the attention-weighted visual features and projected through a linear transformation to generate the vision-guided textual features (Eq. 4).

$$Z_v^{'} = Z_v W_1 \tag{2}$$

$$A = Softmax(Z_t Z_v^{'}) \tag{3}$$

$$Z_t^{'} = Concat(Z_t, AZ_v)W_2 \tag{4}$$

A.2 Multi-head Attention Based Fusion

In this type of fusion mechanism, a multi-head attention mechanism based on vision guidance is used for text-vision fusion. Query, Key, and Value are all projected linearly from the input text and visual components (Eq. 5 - Eq. 7). Cross-modal attention is utilized to gather the text-queried visual features (Eq. 8). Finally, the final output representation is created by combining input textual features and text-queried visual features (Eq. 9).

$$Q = Z_t W_q \tag{5}$$

$$K = Z_v W_k \tag{6}$$

$$V = Z_v W_v \tag{7}$$

$$O = CMA(Q, K, V) \tag{8}$$

$$Z_t^{'} = Concat(Z_t, O)W_3 \tag{9}$$

B Experimental Setups

B.1 Generation Baselines

BART (Lewis et al., 2019): BART is an encoder-decoder-based transformer model which is mainly pre- trained for text generation tasks such as summarization and translation. BART is pre- trained with various denoising pretraining objectives such as token masking, sentence permutation, sentence rotation etc.

T5 (Raffel et al., 2020): T5 is also an encoder-decoder-based transformer model which aims to solve all the text-to-text generation problems. The main difference between BART and T5 is the pre-training objective. In T5, the transformer is pre-trained with a denoising objective where 15these masked tokens whereas, during pre-training of BART, the decoder generates the complete input sequence

VG-BART (Yu et al., 2021): VG-BART is a multimodal variant of BART proposed by Yu et al. (2021) that uses a text-vision fusion mechanism inside BART encoder.

VG-T5 (Yu et al., 2021): The work of Yu et al. (2021) presents VG-T5, a multimodal version of T5 which incorporates a text-visual fusion technique within the T5 encoder.

B.2 Segmentation Baselines

Fully Convolutional Network (FCN): FCN (Long et al., 2015) is a type of CNN that can segment images of any size, it was one of the first models that can handle variable size inputs, now it is a standard in most segmentation models. The model upsamples the feature maps from lower layers and combine them with higher layer feature maps to produce the final segmentation mask.
DeepLabv3 DeepLabv3 (Chen et al., 2017), developed by Google in 2017, is a state-of-the-art semantic image segmentation model that utilizes an encoder-decoder architecture incorporating atrous convolution and skip connections to enhance segmentation accuracy.
MobileNetv3: MobileNetv3 (Howard et al., 2019) is a lightweight neural network architecture utilizes a combination of depthwise convolution and bottlenecks blocks to achieve high efficiency and accuracy. It also uses a new neural architecture search method to find the optimal combination of building blocks.
UNet: UNet (Ronneberger et al., 2015) is a convolutional neural network, utilizes a "U" shaped architecture that combines the feature information from a downsampling path with the upsampled output from an upsampling path. The architecture also uses skip connections to concatenate the feature maps from the downsampling path to the upsampling path, which helps to improve segmentation performance.

B.3 Evaluation Metrics

We present the scores of five automated evaluation metrics, including ROUGE (ROUGE, 2004) and BLEU (Papineni et al., 2002), which are used to measure the performance of the textual explainability, as well as Dice Coefficient (DC) (Dice, 1945), Jaccard Similarity (JS) (Jaccard, 1901), and mean Intersection over Union (mIOU), which are used to evaluate the visual explainability.

(i) **BLEU:** One of the earliest metrics to be used to measure the similarity between two phrases is BLEU. It was first proposed for machine translation and is described as the geometric mean of n-gram precision scores times a brevity penalty for short sentences. We apply the smoothed BLEU in our experiments as defined in (Lin and Och, 2004).

(ii) **ROUGE-L:** ROUGE was first presented for the assessment of summarization systems, and this evaluation is carried out by comparing overlapping n-grams, word sequences, and word pairs. In this work, we employ ROUGE-1 (unigram), ROUGE-2 (bigram) and ROUGE-L version, which measures the longest common subsequences between a pair of phrases.

(iii) **Dice Coefficient:** The Dice coefficient is a similarity metric used in image segmentation to measure the similarity between two sets. It ranges from 0 to 1, where 1 indicates perfect match and 0 indicates no match. The formula for Dice coefficient is $(2 * |A \cap B|)/(|A| + |B|)$, where A and B are the two sets being compared. It is particularly useful when working with imbalanced datasets.

(iv) **Jaccard Similarity:** Jaccard similarity is a similarity metric used to measure the similarity between two sets, it is often used in natural language processing, information retrieval and image segmentation. It ranges from 0 to 1, where 1 indicates perfect match and 0 indicates no match. The formula for Jaccard similarity is $|A \cap B|/|A \cup B|$, where A and B are the two sets being compared.

(v) **mIOU:** Mean Intersection over Union (mIOU) is an evaluation metric used in image segmentation tasks, it is the mean of the Intersection over Union (IoU) scores for all the classes. It is used to measure the similarity of predicted segmentation maps with ground truth segmentation maps, unlike Jaccard similarity which is used to measure the similarity between two sets.

B.4 Training Details

In this section, we detail various hyperparameters and experimental settings used in our work. We have performed all the experiments on Tyrone machine with Intel's Xeon W-2155 Processor having 196 Gb DDR4 RAM and 11 Gb Nvidia 1080Ti GPU. We have randomly chosen 70% of the data for training, 10% for validation, and the remaining 20% for testing. We have executed all of the models five times, and the average results have been reported. We have used BART (Lewis et al., 2019), T5 (Raffel et al., 2020) as the base model for our proposed model. All the models are trained for a maximum of 40 epochs and a batch size of 32. Adam optimizer is used to train the model with an epsilon value of 0.00000001. All the models are implemented using Scikit-Learn[8] and pytorch[9] as a backend.

[8] https://scikit-learn.org/stable/
[9] https://pytorch.org/

C Annotations

C.1 Annotation Guidelines

We follow cyberbullying definition by (Smith et al., 2008) for our annotation process. In order to help and guide our annotators, we provide them with several examples of memes with textual and visual explanations marked. We first read the entire text present inside the memes for rationale annotations and looked at the depicted visual clues. Each lexicon was marked either Bully or Non-bully based on the visual and textual context. Additionally, visual regions were segmented that prominently justified the rationale annotations for visual explanations.

C.2 Daywise Schedule

- **Day 1 and Day 4:** Each annotator was assigned to annotate rationales for 150 memes. They were instructed to annotate 30 memes per batch within one hour, followed by a mandatory break of 10 minutes (cf. Section C.3).
- **Day 2 and Day 5:** Each annotator was assigned to highlight the visual regions that could justify the rationale annotations.
- **Day 3:** We arrange meetings with the annotators to ensure that their mental well-being is not adversely affected during the annotation process (cf. Section C.3).

C.3 Annotation cost

The process of annotating multimodal explanation is time-consuming and expensive, with each meme sample requiring 2-3 minutes for textual and visual explanation each. We initially hired 10 annotators and selected 3 best annotators among them. An honorarium of 5 INR was offered per sample due to the inherent complexity, which was ensured to be appropriate considering the 160-750 INR minimum wage/day based on the Minimum Wages Act, 1948[10] in India (where the annotations were done) based on the average number of annotations across all annotators per day. The entire annotation process took approximately 10 weeks to complete following daywise schedule.

Ethics note: Repetitive consumption of online abuse could distress mental health conditions (Ybarra et al., 2006). Therefore, we advised annotators to take periodic breaks and not do the annotations in one sitting. Besides, we had weekly meetings with them to ensure the annotations did not have any adverse effect on their mental health.

C.4 Statistics of Annotated Multimodal Explanations

Figure 4 illustrates the distribution of meme text. The figure showcases that the length of meme text typically falls within the range of 0 to 80 characters. Upon conducting calculations, the average length of meme text was determined to be approximately 14.12 characters. In a similar vein, the length of rationales ranges from 0 to 40, as depicted in Figure 5. The average token length of annotated rationales was observed to be around 6.79. Furthermore, we observed that, on average, 35.18% of the image area is dedicated to visual explanations for cyberbullying memes. The distribution for the percentage of area selected for annotated visual explanations can be found in Figure 6.

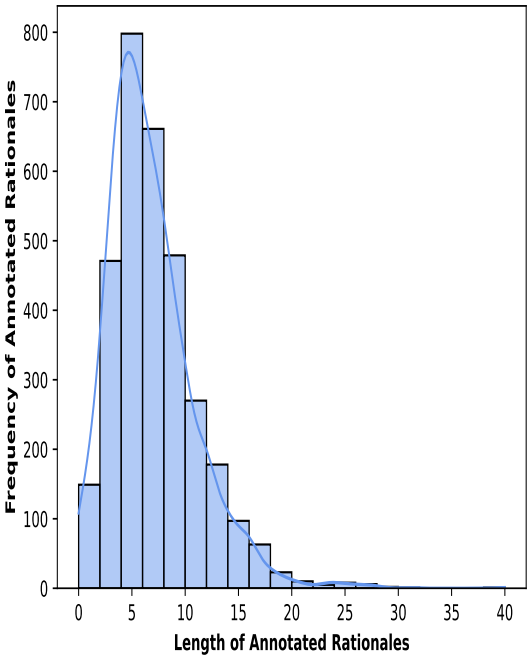

Figure 4: Distribution for Length of Meme Text

[10] https://en.wikipedia.org/wiki/List_of_countries_by_minimum_wage

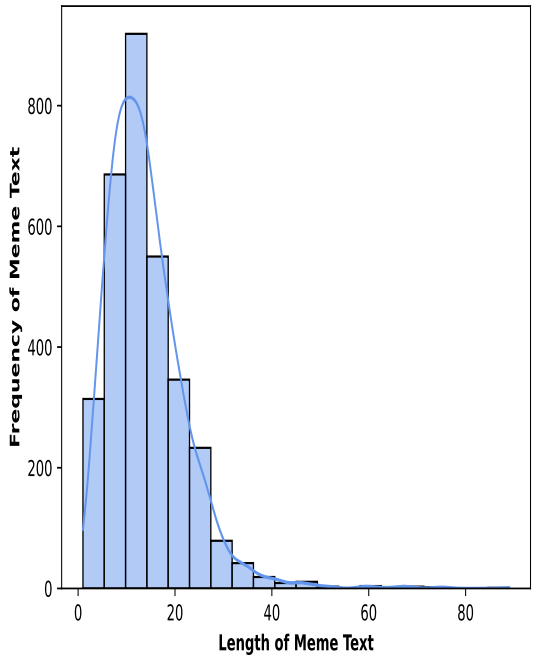

Figure 5: Distribution for Length of Annotated Rationales

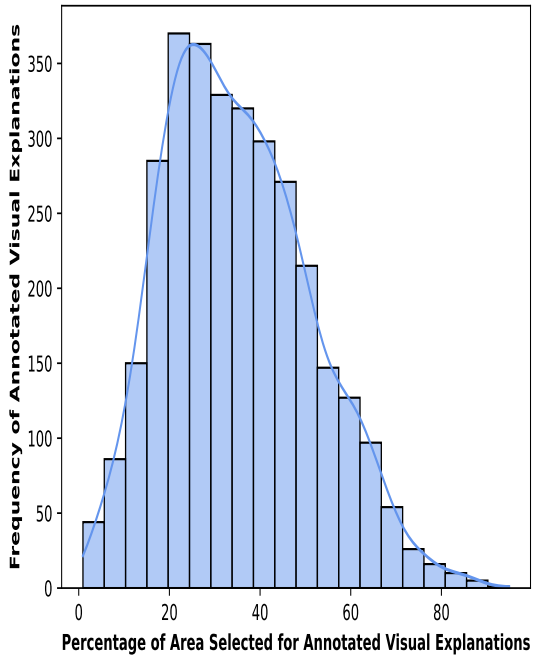

Figure 6: Distribution for Annotated Visual Explanations