A Multitask Framework for Sentiment, Emotion and Sarcasm aware Cyberbullying Detection from Multi-modal Code-Mixed Memes

Krishanu Maity*
Indian Institute of Technology Patna
Patna, India
krishanu_2021cs19@iitp.ac.in

Prince Jha*
Indian Institute of Technology Patna
Patna, India
princekumar_1901cs42@iitp.ac.in

Sriparna Saha
Indian Institute of Technology Patna
Patna, India
sriparna@iitp.ac.in

Pushpak Bhattacharyya
Indian Institute of Technology Bombay
Bombay, India
pb@cse.iitb.ac.in

ABSTRACT

Detecting cyberbullying from memes is highly challenging, because of the presence of the implicit affective content which is also often sarcastic, and multi-modality (image + text). The current work is the first attempt, to the best of our knowledge, in investigating the role of sentiment, emotion and sarcasm in identifying cyberbullying from multi-modal memes in a code-mixed language setting. As a contribution, we have created a benchmark multi-modal meme dataset called *MultiBully* annotated with bully, sentiment, emotion and sarcasm labels collected from open-source Twitter and Reddit platforms. Moreover, the severity of the cyberbullying posts is also investigated by adding a *harmfulness* score to each of the memes. The created dataset consists of two modalities, text and image. Most of the texts in our dataset are in code-mixed form, which captures the seamless transitions between languages for multilingual users. Two different multimodal multitask frameworks (*BERT+ResNET-Feedback* and *CLIP-CentralNet*) have been proposed for cyberbullying detection (CD), the three auxiliary tasks being sentiment analysis (SA), emotion recognition (ER) and sarcasm detection (SAR). Experimental results indicate that compared to unimodal and single-task variants, the proposed frameworks improve the performance of the main task, i.e., CD, by 3.18% and 3.10% in terms of accuracy and F1 score, respectively. [1]

KEYWORDS

Cyberbullying, Sentiment, Emotion, Sarcasm, Multitask, Memes

ACM Reference Format:
Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. A Multitask Framework for Sentiment, Emotion and Sarcasm aware Cyberbullying Detection from Multi-modal Code-Mixed Memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531925

*Both authors contributed equally to this research.
[1]Code available at https://github.com/Jhaprince/MultiBully

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531925

1 INTRODUCTION

Cyberbullying [34] is described as the serious, intentional, and repetitive act of a person's cruelty towards others using various digital technologies. It is mainly expressed through nasty tweets, texts, memes or other social media posts. One such example of the written text in image is internet memes which are achieving very high popularity, resulting in over 180 million posts across all social media platforms until 2018 [40]. Memes are combinations of images and texts which are superimposed on the image providing a message. Most of the times, an image or text solely is not sufficient to understand the intended message.

Different studies have reported that cyberbullying affects between 10 to 40 percentage of internet users [39]. Cyberbullying results might vary from anxiety, sadness, transient fear to suicidal thinking. As a result, spotting cyberbullying early on is crucial for preventing its consequences.

The process of seamlessly flipping between two or more languages in a discussion is known as code-mixing (CM) [25]. Code-mixing is common in multilingual countries where people often use mixture of two languages for regular post and conversations on social media. In the current work, we aim to investigate the identification of cyberbullying from code-mixed memes. The emotional state and sentiments of a person have significant influence on the intended content [21]. Sentiment and emotion are inextricably linked, and one aids in the comprehension of the other. Emotions like happiness and joy, for example, are intrinsically associated with positive sentiments. A well-known observation is that a meme labeled as bully usually conveys negative sentiment. Because of the strong correlation between emotion and sentiment, we should consider the sentiment expressed in the post as well as its emotion information while predicting whether the meme is of bullying type or not. Identifying offensive memes is more challenging as they express sarcasm in an implicit way [35], which motivates us to consider sarcasm information while identifying memes containing bully information. For example, if we see first sample in Figure 1, there is no indication of bullying based on text, image, or both. But

if we consider the implicit sarcasm present in this meme, we can easily mark it as a bully.

Existing literature reports several works where the tasks of sentiment analysis (SA) and emotion recognition (ER) are treated as auxiliary tasks to boost the performance of primary task (like sarcasm detection (SAR) [5], tweet act classification (TAC) [33]) in a multitask (MT) framework. The use of domain-specific information to related tasks enhances the overall learning process. Multi-task learning is proven to be effective when working on related tasks [4].

Furthermore, multi-modal inputs, such as a combination of text and image [12], contribute in creating trustworthy classification models that aid in detecting the user's emotional state and sentiment, which in turn assist the CD task. Sometimes it is not enough to rely on either the text or the visual modality to correctly comprehend information; rather, both modalities should be processed together to infer the meme's accurate meaning.

This paper aims to design a multitask multimodal framework for cyberbullying detection from memes with Hindi-English code-mixed text where SA, ER and SAR act as the secondary/auxiliary tasks to increase the performance of the primary task, i.e., CD. Our developed model utilizes different multimodal feature extractor modules (BERT+ResNET, CLIP) for efficient representation of multimodal data. We have introduced two multitask frameworks (Feedback Multitask, Central-Net Multitask) for boosting the performance of main task with the help of secondary tasks.

The following are the primary contributions of this work:

(1) We are the first to introduce the task of sentiment-emotion-sarcasm aware cyberbully detection from multimodal memes in code-mixed language setting. For this purpose, a new code-mixed dataset called *MultiBully* of memes (image+text) annotated with bully, sentiment, emotion and sarcasm labels is introduced. We believe this dataset will help in future research on sentiment, emotion and sarcasm-aware cyberbully detection from memes [2].
(2) Further the severity of the cyberbullying post is also quantified by incorporating a *harmfulness* score into our dataset.
(3) We have proposed two multi-task multi-modal frameworks namely *BERT+ResNET-Feedback* and *CLIP-CentralNet* for sentiment, emotion and sarcasm aided cyberbullying detection. We find that CentralNet, well-known for multimodal data fusion, can be a suitable architecture for multitask learning.
(4) Experimental results illustrate the efficacy of solving the CD, SA, ER and SAR tasks together in a multi-task framework. Multi-modal and multi-task CD outperforms uni-modal and single-task CD by a substantial margin.

2 RELATED WORKS

With the advancement of natural language processing (NLP), much researches on the identification of cyberbullying have been conducted in the English language rather than other languages [32].

2.1 Works on Monolingual Datasets

Dinakar et al.[10] proposed an experimental work by applying binary classifiers on a corpus of 4500 YouTube comments for cyberbullying detection. They obtained an overall accuracy of 66.70% with SVM classifier and 63% with Naive Bayes classifier. Reynolds et al.[30] worked on data collected from the Formspring.me and labeled using web service to train their model, they had used a Weka tool kit and were able to achieve 78.5% accuracy by using C4.5 decision tree learner. Djuric et al. [11] proposed a methodology for distributed low dimensional representations of comments using paragraph2vec and continuous BOW (CBOW) approach for hate speech detection. They tested their method on a vast data set of user comments gathered from the Yahoo Finance website and found it 80.01% accurate. Balakrishnan et al. [2] developed a strategy for detecting cyberbullying for Twitter users based on psychological characteristics and machine learning approaches in 2020. They examined that considering personalities and sentiment features with baseline features (text, user, network) improves the cyberbullying detection task and achieves a sound accuracy of 91.7%. In 2020, Paul et al. [26] developed a BERT-based framework, namely cyberBERT, for cyberbully identification. They have evaluated cyberBERT on three benchmark datasets, i.e., Formspring (12k posts), Twitter (16k posts), and Wikipedia (100k posts), and obtained state-of-the-art results. Here, BERT generated pooled output(CLS token) of dimension 768 is the final representation of an input sentence.

2.2 Works on Code-Mixed Datasets

Kumar et al. [20] developed aggression-annotated corpus containing 18k tweets and 21k facebook comments written in Hindi-English code-mixed form. Bohra et al. [3] developed a code-mixed dataset of 4575 tweets and annotated with hate speech and normal speech. SVM classier achieved 71.7% accuracy score when word n-grams, punctuations, character n-grams, hate lexicon and negation words are taken into account as feature vectors. The authors in [18] proposed a deep learning based approach to identify hate speech from Hindi-English code-mixed corpus. With the help of domain specific word embedding, they outperformed the base model by 12% F1 score. In [23], authors have introduced a code-mixed Indian language dataset for cyberbullying detection. They developed a model based on deep learning architectures that include BERT, CNN, GRU, and capsule networks and attained 79.28% accuracy.

2.3 Works on Sentiment, Emotion and Sarcasm aware Multitasking

There are some works in the literature where the tasks of SA and ER are treated as auxiliary tasks to boost the performance of primary task. Authors in [33], proposed a multi-task ensemble adversarial learning framework for multi-modal tweet act classification(TAC). Authors have claimed that TAC performs significantly better than its uni-modal and single task TAC variants. Authors in [14] have suggested a multi-task learning architecture that uses external knowledge information to improve overall performance of the emotion classification task on suicide notes. To analyze the effects of sentiment and emotion on the sarcasm detection task, [5] presented a multi-task framework based on Inter-segment and Intra-segment attention mechanisms in 2020. In [22, 24], authors developed attention-based multitask models to investigate how sentiment and emotion information helps to identify cyberbullying from Hinglish code-mixed text.

[2]The dataset will be made available: https://www.iitp.ac.in/~ai-nlp-ml/resources.html

2.4 Works on Meme Datasets

There is relatively little research on identifying offensive and hate content in memes. Kiela et al. [19] introduced a benchmark multimodal meme dataset for hate speech detection. Using pretrained Visual-BERT, they have achieved 69.47% testing accuracy. Gomez et al. [15] presented MMHS150K, a multimodal dataset of tweets with both image and text information that has been manually annotated for hate speech. Authors in [35] created a Multi-modal(Image+Text) Meme Dataset (MultiOFF) for identifying offensive contents from memes. They employed an early fusion approach to merge the image and text modalities and compared its performance with respect to a text-only and an image-only baselines. Authors in [27] developed a benchmark dataset consisting of 3,544 memes, namely HarMeme, to detect harmful memes (as very harmful, partially harmful, or harmless) and their target (organization, individual, community, or society/general public/other).

After a thorough literature survey, we observed that there is no work available utilizing sentiment, emotion and sarcasm information for cyberbullying detection from code-mixed meme. This motivates us to work in this specific domain. The current work is the first attempt to fill this research gap.

3 CODE-MIXED *MULTIBULLY*-MEME ANNOTATED DATASET DEVELOPMENT

Firstly, we combed the literature for the multimodal CD dataset annotated with another three labels, i.e., sentiment, emotion and sarcasm. To the best of our knowledge, there is no publicly available code-mixed multimodal and multitask corpus for cyberbullying detection.

India is a multilingual country having several official languages. Code-mixing is very common in social media posts by Indians. About 691 million native speakers use Hindi as one of the official Indian languages.[3] In India, Hindi, English, and Hinglish make up the majority of text interactions on social networking platforms. The depiction of Hindi language in Roman form is known as Hinglish. Hence we decided to use Hinglish for developing a code-mixed corpus for identification of cyberbullying from memes.

3.1 Data Collection

We have scanned the internet for various platforms where memes are shared on a regular basis. We found a few places from where we can gather a range of multimodal data, including Facebook, Twitter and Reddit. After a comprehensive search and analysis of all of these platforms, we decided to focus on Twitter and reddit for further investigation of our objective because facebook does not provide as much flexibility of scraping data as other platforms do. We then, obtained a number of hashtags for scrapping images from twitter such as MeToo, KathuaRapeCase, Nirbhya, Rendi, Chuthiya, Kamini etc. and from reddit, we used subreddits such as Desimemes, HindiMemes, bakchodi etc. and fetched approximately 25000 images or memes.

[3]https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India

3.2 Data Preprocessing

Raw scraped data is comprised of numerous memes that are irrelevant. For making the annotation task more convenient, we have removed irrelevant memes based on the following criteria:

(1) Removed images that contain textual information only, such as screenshots of other tweets or posts that are entirely textual and missing visual information.
(2) Some images that we scrapped from reddit were corrupted and not opening, so we eliminated those images too using python inbuilt function.
(3) The text of the meme is unreadable. (e.g., hazy text, missing text, etc.)

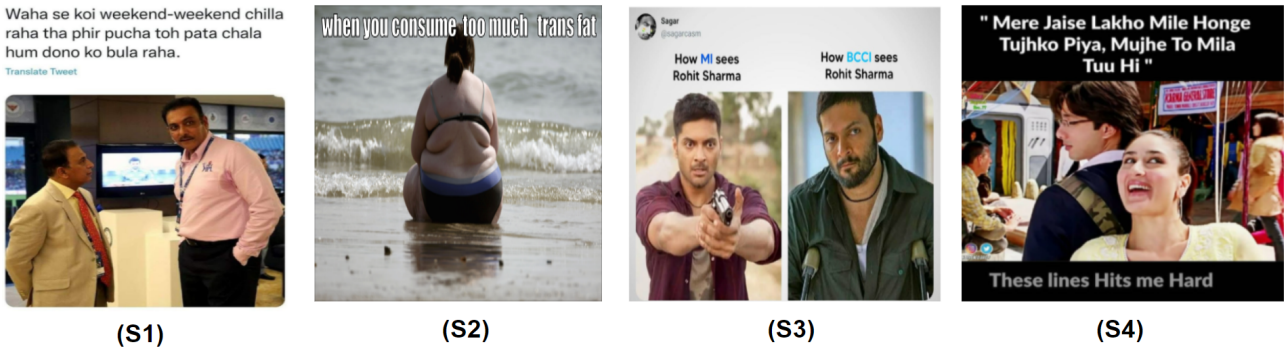
3.3 Data Annotation

Three annotators having proficient linguistic background in both Hindi and English were involved in data annotation. Based on the context of memes, annotators have assigned five labels for each meme: Bully class (Bully/Non-bully), sentiment class (Positive/Neutral/Negative), Emotion class, Sarcasm class (Yes/No) and Harmfulness class (Partially-Harmful/Very-Harmful/Harmless). For emotion class, we have considered eight Plutchik's emotion categories (Joy, Sadness, Fear, Surprise, Anger, Disgust, anticipation, and trust) as well as a new emotion class, namely *"Ridicule"*. First, we started emotion class annotation with eight Plutchik's emotion categories and "Other" class for handling samples that do not belong to any of the eight predefined emotions. After finishing the annotation, we saw that one-third of the samples (Approximately 2000) belong to "Other" category. Then we have manually checked those samples and found out that many samples are ridiculous in nature. Based on this observation, we have introduced this *"Ridicule"* emotion class in our multimodal meme dataset. The harmfulness class has three labels. *Harmless* signifies that there is no indication of cyberbullying. *Partially-Harmful* indicates that the post has cyberbullying content. However, these posts are not severe, and *Very-Harmful* indicates the post contains serious indications of cyberbullying (e.g., physical threats or excitements to commit suicide).

We have given some memes to annotators with gold labels and explanations for better comprehension. After manual annotation, the majority voting technique was used for selecting the final bully, emotion, sentiment and harmfulness labels . Annotators were also instructed to annotate the memes without being biased towards any specific demographic area, religion, etc. We calculated the inter-annotator

agreement (IAA) using Fleiss'[13] Kappa score to verify the quality of annotation. We attained the agreement scores of 0.78, 0.82, 0.69, 0.72 and 0.77 on the CD, SA, ER, Sarcasm and harmfulness task, respectively, indicating that those are of acceptable quality.

Importance of Multimodality, Sentiment, Emotion and Sarcasm Information: In Table 3, some examples of annotated samples are shown. These examples demonstrate the need for considering multimodal inputs as well as sentiment and emotion information in identification of cyberbullying from different social media platforms. The text and image together are used for annotation. For example, in the first sample of Figure 1, if we see only the image without considering the text, there is nothing abusive or bullying. Similarly, if we read the sentence without considering



Translation (s1): Someone was chanting with weekend-weekend but later it was realised that we two were being called.
Translation (s4): You may have met lakhs of people like me. I have met only you.

Figure 1: Some samples from annotated multimodal meme dataset. Different annotation labels corresponding to these samples are described in Table 3.

Table 1: Class wise data statistics of our developed meme dataset

Split	#Memes	Cyberbully		Harmfulness			Sentiment			Sarcasm	
		Non-Bully	Bully	Harmless	Partially Harmful	Very Harmful	Positive	Neutral	Negative	Yes	No
Train	4097	1854	2243	1936	2140	21	404	1796	1897	1545	2552
Validation	585	267	318	274	308	3	69	265	251	201	384
Test	1172	511	661	541	621	10	134	498	540	429	743
Total	5854	2632	3222	2751	3069	34	607	2559	2688	2175	3679

Table 2: Statistics of Different Emotion Classes

Split	#Memes	Emotion									
		Joy	Sadness	Fear	Anger	Anticipation	Surprise	Disgust	Trust	Ridicule	Other
Train	4097	427	404	100	472	295	593	646	38	490	632
Validation	585	70	60	16	65	41	95	88	4	56	90
Test	1172	137	121	21	119	78	166	187	15	149	179
Total	5854	634	585	137	656	414	854	921	57	695	901

the background image, then it is a non-bully sentence. But when we consider both text and image, then there is a clear indication of sarcasm. So this multimodal data with sarcasm information helps in identifying this sample as a bullying type. The gold label of sample-3 is bully as it tries to humiliate a cricketer based on his performance. There is nothing wrong with the text part, but if you see the left part of the image, there is a clear indication of negative sentiment with fear emotion. In contrast, the right part of the image indicates negative sentiment with sadness emotion. The overall sentiment (Negative) and emotion (Sadness) information from both the modalities help in identifying that sample-3 is a cyberbully.

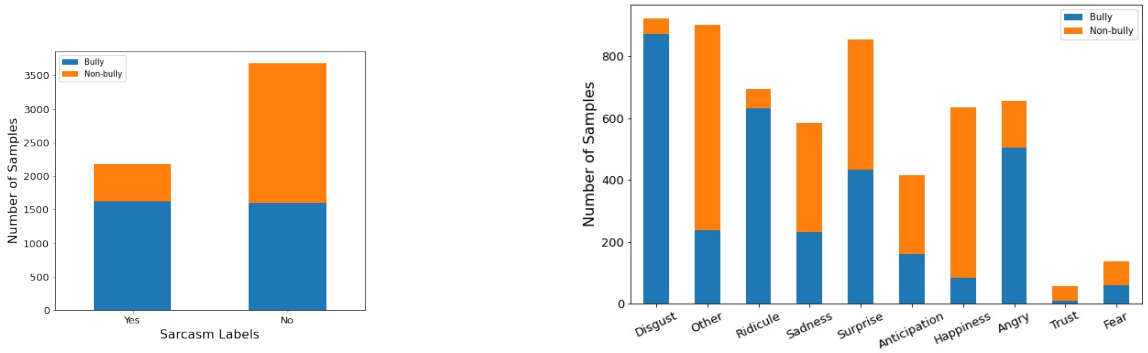
3.4 Dataset Statistics

Out of 5854 memes in our database, 2632 were labeled as nonbully, while 3222 were tagged as bullies. The percentages of non-bully and bully memes in our corpus are 44.96% and 55.04%, respectively. Dataset statistics based on Cyberbully, Harmfulness, Sentiment and Sarcasm classes are shown in Table 1 and Emotion statistics are shown in Table 2.

Table 3: Samples with annotation labels

Sample	Sentiment	Emotion	Sarcasm	Bully	Harmfulness
S1	Negative	Ridicule	Yes	Bully	Partially Harmful
S2	Negative	Disgust	No	Bully	Very Harmful
S3	Negative	Sadness	No	Bully	Partially Harmful
S4	Positive	Joy	No	Non-bully	Harmless

In our corpus number of memes having negative sentiment is 2688, in which 2375 samples are marked as bully, and 313 samples are marked as non-bully. On the other hand, out of 607 positive sentiment memes, 29 were marked as bully, and the remaining 578 were marked as non-bully. Figure 3 reveals a correlation between the non-bully vs. positive sentiment and bully vs. negative sentiment. It indicates that a meme with a positive sentiment is probably a non-bully one, and a bully meme is more likely to have a negative



(a) Bully Vs. Sarcasm **(b) Bully Vs. Emotion**

Figure 2: Correlation between Bully Vs. Sarcasm and Bully Vs. Emotion

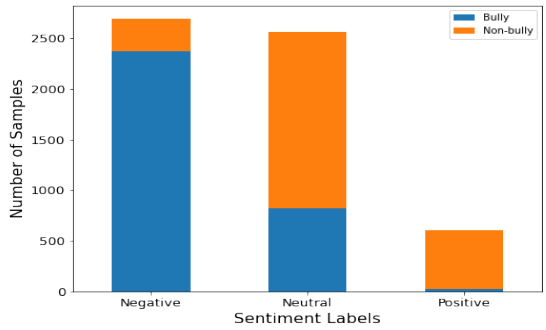


Figure 3: Correlation between Bully Vs. Sentiment

sentiment. Figure 2b shows a high correlation between the bully vs. anger and disgust emotion classes, indicating that a meme labeled as the bully is more likely to have anger or disgust emotion. Figure 2a shows the relationship between bully classes and sarcasm. We can see that a sarcastic meme has more chance of being bullied. But a non-sarcastic meme has an almost equal probability of being either bullied or non-bullied.

3.5 Dataset Comparison

Table 4 compares *MultiBully* to several frequently used cyberbully datasets. Table 4 illustrates that our corpus has some unique aspects/characteristics as compared to other datasets : (1) *MultiBully* is the first multi-modal cyberbully dataset having two modalities, image and text; (2) It is not only manually annotated with cyberbully labels but also with sentiment, emotion, sarcasm labels to solve the task of sentiment, emotion and sarcasm aware cyberbully detection. This dataset can solve four tasks simultaneously, i.e., CD, SA, ER, SAR; (3) The texts present in this dataset are in Hindi-English code-mixed form rather than English.

Table 4: A comparison between *MultiBully* and other existing widely-used cyberbully datasets.

Dataset	Language	Instance	Balancing	Modality
Formspring1 [30]	English	3915	14.2%	Text
YouTube2 [8]	English	4626	9.7%	Text
MySpace2 [7]	English	2200	–	Text
AskFM [36]	Dutch	85485	6.7%	Text
Schoolboard Bulletins (BBS) [28]	Japanese	2222	12.8%	Text
Twitter3 [1]	English	10007	6%	Text
Instagram [17]	English	1954	29%	Text
Formspring4 [31]	English	13160	19.4%	Text
Aggression Annotated Dataset[20]	Hindi+English	39000	57.4%	Text
Code-mixed-bully [23]	Hindi+English	5062	51.48%	Text
MultiBully (Our Dataset)	Hindi+English	5854	55.04%	Text+Image

4 METHODOLOGY

This section describes different deep multitask multimodal frameworks we have developed to identify cyberbullying from memes. Figure 4 and 5 depict different multimodal feature extraction models and multitask frameworks, respectively. We have developed two feature extraction modules (BERT-ResNet Feature Extractor and CLIP Feature Extractor) and two multitask frameworks (Feedback Multitask and CentralNet Multitask). Furthermore, we have conducted our experiments on different combinations of feature extraction and multitask frameworks to examine which combination works better for our multimodal multitask problem. We have four different combinations of the feature extraction+multitask models, i.e., BERT-Resnet+FeedBack, BERT-Resnet+CentralNet, CLIP+FeedBack, and CLIP+CentralNet.

4.1 BERT-ResNet Feature Extractor

Text Features BERT [9] is a transformer-based [37] language model. Most of the sentences in our meme dataset are written in Hindi-English code-mixed form, so we have utilized specific BERT variant, i.e., mBERT, which has been trained on 104 different languages, including Hindi and English. Google's optical

character recognition (OCR) Vision API5[4] has been employed to extract the text from the input image. The BERT language model has been employed to get the textual features from input text $W = \{w_1, w_2, ...w_{n_x}\}$. Let $X \in \mathbb{R}^{n_x \times d_x}$ be the sequence output obtained from the BERT model for input W, where n_x is the maximum sequence length and $d_x = 768$ is the dimension of each token. Outputs from BERT are passed through Bi-GRU [6] layer to learn the contextual information and capture long term dependency of input word vectors.

To capture long term dependency of input word vectors, Bi-GRU encodes the input on both forward and backward direction as

$$\overrightarrow{h}_t^i = \overrightarrow{GRU}(w_t^i,\ h_{t-1}^i)\ ,\ \overleftarrow{h}_t^i = \overleftarrow{GRU}(w_t^i,\ h_{t+1}^i) \tag{1}$$

where each word vector w_t^i of sentence i is mapped to a forward hidden state $\overrightarrow{h}_t^i$ and backward hidden state $\overleftarrow{h}_t^i$ by invoking $\overrightarrow{GRU}$ and $\overleftarrow{GRU}$, respectively.

$$\left[h_t^i = \overrightarrow{h}_t^i, \overleftarrow{h}_t^i\right] \tag{2}$$

Image Features ResNet-50 [16] has state-of-the-art performance in image classification tasks. Thus, we have used ResNet-50 as our base model for image feature extraction. To get a 2048 dimensional dense vector, we have passed the last convoluted features of ResNet of dimension $(7 \times 7 \times 2048)$ through a global average pooling layer. Outputs from the average pooling layer are passed through a fully connected layer (512 neurons), followed by a dropout layer to generate the final image feature vector (I). Finally, we have concatenated the text feature vector (T) from BERT+GRU with the image feature vector from ResNeT+FC to get the image text combined feature vector, F. Details of feature extraction modules can be found at Table 5

Table 5: Model parameters of different feature extractor modules

Features	Model	Type	Output Size
Text	MBERT+BiGRU	MBERT	50×768
		BiGRU	50×512
	CLIP-Text Encoder	BERT	512
Image	ResNet+Dense	ResNet	$7 \times 7 \times 2048$
		GlobalAvgPool	2048
		Dense	512
	CLIP - Image Encoder	Vision Transformer	512

4.2 CLIP Feature Extractor

We have used CLIP (Contrastive Language–Image Pre-training) [29], a pre-trained visual-linguistic model, to encode each text–image pair, leveraging its representation capability to capture the meme's overall semantic. CLIP was pre-trained on 400 million image–text pairs extracted from the Internet. Given a batch of N (image,text) pairs, it is trained to predict N correct match out of $N \times N$ possible pairings. CLIP maximizes the cosine similarity of N real pairs in batch by training image and text encoders together to create an efficient multimodal embedding space; hence minimizing the cosine similarity over $N^2 - N$ incorrect pairs. A symmetric crossentropy loss is used for optimization over cosine similarity scores. Due to the wide range of images it has seen and the natural language supervision, CLIP has outstanding zero-shot capabilities. We use Vision Transformer as an image encoder and BERT as a text encoder. We have extracted a CLIP image embedding, F_I, and a CLIP text embedding, F_T, from the meme's image, I and its OCR-extracted text T, respectively; both F_I and F_T are 512-dimensional vectors.

[4]https://cloud.google.com/vision/docs/ocr

4.3 Inter-modal Attention

The text modality is more significant for some memes, whereas the visual modality is more important for others. To merge the representations from the textual and visual modalities, Inter-modal Attention has been employed. We apply an approach similar to that described in [37], in which authors suggested that attention should be computed by mapping a query and a set of key-value pairs to an output. Outputs of both the modalities (T and I) are passed through three fully connected layers, namely queries(Q), keys(K) and values(V) of dimension d_f. Inter-modal attention (IA) scores are computed as follows:

$$IA_i = softmax(Q_i K_i^T) V_i \tag{3}$$

where $IA_i \in \mathbb{R}^{n_x \times d_f}$. Table 5 describes details about model parameters of different feature extractor modules.

4.4 Feedback Multitask Framework

In this framework, to learn n number of tasks simultaneously, multimodal features are passed through n number of different task-specific fully connected (FC) layers followed by an output layer(Softmax). There is a feedback path from the last FC layer of tasks $(T_1, T_2, ...T_n)$ to the main task T_n. This feedback path aims to examine how different task specific features help in boosting the main task's performance. Each task-specific layer excluding the main task consists of two FC layers (100 neurons) followed by an output layer (softmax). We have kept one FC layer for the main task because other task-specific features from the last FC layer are concatenated with the features of the main task.

Let us assume there are three tasks, with sentiment and emotion identification as secondary tasks and cyberbully identification as the primary task. The concatenated feature vector F is passed through three separate task-specific fully connected layers (bully channel $[FC_B^1(100\ neurons)]$, sentiment channel $[FC_S^1(100\ neurons)$ + $FC_S^2(100\ neurons)]$) and emotion channel $[FC_M^1(100\ neurons)$ + $FC_M^2(100\ neurons)]$) followed by their corresponding output layers. The outputs, FC_S^2 of the sentiment channel and FC_M^2 of the emotion channel are concatenated with FC_B^1 of the bully channel for generating sentiment+emotion-aided bully features, which help in enhancing the performance of the primary task, i.e., cyberbully detection.

4.5 CentralNet Multitask Framework

CentralNet [38] is a multimodal data fusion network. In our work, we have reformed the CentralNet as a multitask framework. CentralNet Multitask is a neural network architecture in which we have n independent networks for task-specific networks and one central network. The task-specific network consists of $n - 1$ secondary tasks (ST) and one main task (MT). The central network combines

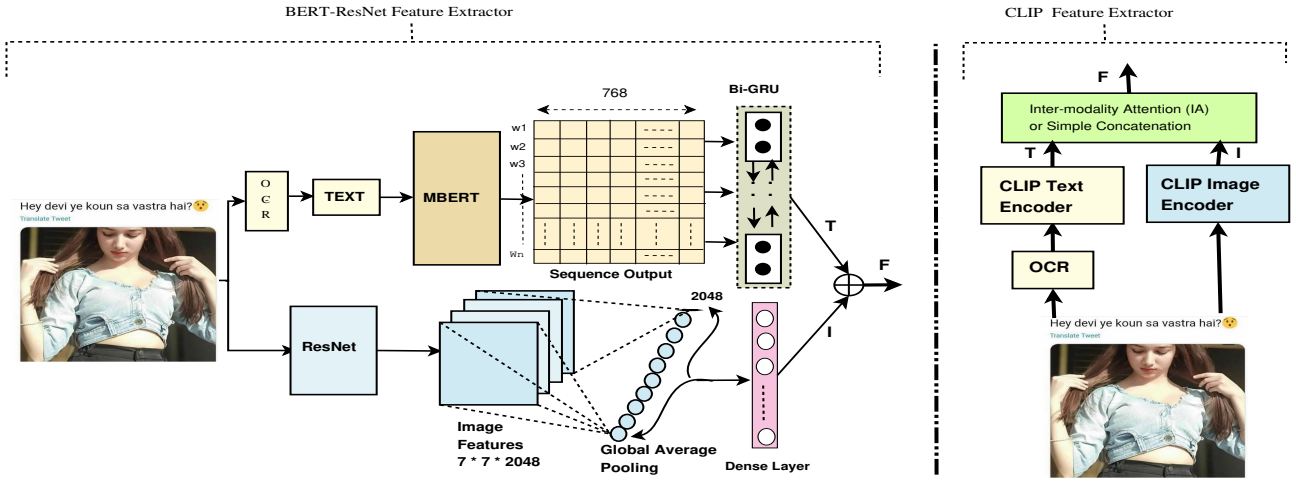


Figure 4: Different feature extraction architectures from multimodal (Image+Text) meme: The left part of the image represents the feature extraction using BERT+ResNet; CLIP based feature extraction module is shown on the right side of the image

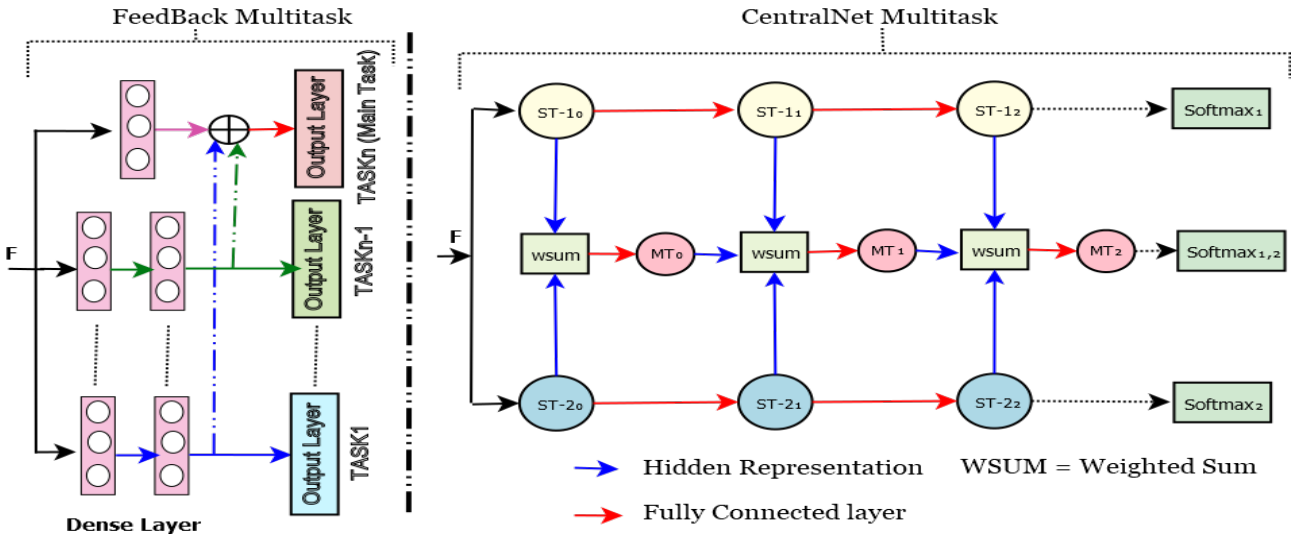


Figure 5: Proposed Multitasking Framework: FeedBack Multitask (left side), CentralNet Multitask (right side)

Table 6: Model parameters of CentralNet multitask framework

Multitask Framework	Task									
	Emotion		Sentiment		Central		Sarcasm		Bully	
	Type	Output Size	Type	Output Size	Type	Output Size	Type	Output Size	Type	Output Size
CentralNet	Dense	512	Dense	512			Dense	512	Dense	512
	Dense	256	Dense	256	Dense	256	Dense	256	Dense	256
	SoftMax	10	SoftMax	3	SoftMax	2	SoftMax	2	SoftMax	2

the features generated from different single tasks by considering a weighted summation of task-specific networks and its own previous layers. Such multitask layers can be defined by the following equation:

$$MT_{i+1} = \alpha m MT_i + \sum_{k=1}^{n} \alpha s_i^k ST_i^k \quad (4)$$

Where n is the number of task-specific networks, αs are scalar trainable weights, ST_i^k is the hidden representation of k^{th} task-specific network at i^{th} layer and MT_i is the central hidden representation of the main task. The resulting layer, MT_{i+1}, is fed to an operating layer (a dense layer followed by an activation layer). The inputs to the first layer of the Central network are only the weighted summation of other task-specific initial features as there is no previous central hidden representation. The central network's output is considered as the final prediction of the main task. More details about the model parameters are given in Table 6.

4.6 Loss Function

We employed categorical cross-entropy $L(\hat{y}, y)$ as a loss function to train the network's parameters.

$$L_{CE}(\hat{y}, y) = -\frac{1}{N}\sum_{j=1}^{C}\sum_{i=1}^{N} y_i^j log(\hat{y}_i^j) \tag{5}$$

Where $\hat{y}_i^j$ is the predicted label and y_i^j is the true label. C and N represent the number of classes, and the number of memes, respectively. The final loss function, $Loss$, is dependent of N task-specific individual losses as follows:

$$Loss = Loss_M + \sum_{k=1}^{n} \beta_k Loss_S^k \tag{6}$$

Where $Loss_M$ is the main-task loss and $Loss_S$ is ST loss. The variable β, which ranges from 0 to 1, defines the loss weights that characterise the per task loss-share to the total loss.

5 EXPERIMENTAL RESULTS AND ANALYSIS

The findings of several variants of our proposed model are shown in this section, which were tested on our proposed multimodal meme corpus. All our experiments were conducted on a hybrid cluster of multiple GPUs comprised of RTX 2080Ti. 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.

5.1 Hyperparameters

We use Tanh activation in bi-GRU (256 hidden cell) and ReLU activation in all fully connected layers (100 neurons). With a batch size of 32, we train our models for 15 epochs. We utilize Adam optimizer and set the learning rate to 0.001 to backpropagate the loss across the network. All the models are implemented using Scikit-Learn 0.22.2[5] and Keras 2.4.3[6] with TensorFlow2 2.4.1[7] as a backend.

5.2 Different Multitask Variants

As we have four tasks, including the main task (CD), there are different multitask variants. We keep CD as the main task for any multitask variants and add other secondary tasks with different combinations. Total three combinations are there for two task settings, i.e., CD+SA, CD+ER, CD+SAR. Similarly, we have three multitask variants with three tasks, i.e., CD+SA+ER, CD+SA+SAR, CD+ER+SAR.

It is worth noting that we aim to improve CD's performance with the aid of the other three auxiliary tasks, SA, ER and SAR. Following that, we provide our findings and analyses, with CD serving as the central task in all task combinations.

5.3 Findings from Experiments

Table 7 presents the results of CD (main task) in terms of accuracy and F1-score for all the uni-modal and multimodal variants of different multitask frameworks. Single task results with different models are shown in Table 8.

From the table, we find: (1) All the multitask variants outperform the single-task classifiers for the CD task. Moreover, our proposed *CLIP+CentralNet* framework with three auxiliary tasks (SA, ER and SER) performs better than the single task CD with the improvements in accuracy and F1 score of 3.18%, 3.1%, respectively. The results imply that sentiment, emotion and sarcasm knowledge enhances the performance of the cyberbully detection task.

(2) In multimodal (Image+Text) scenario, CLIP+CNT combination outperforms other combinations, i.e., BERT-ResNET+FB, BERT-ResNET+CNT and CLIP+FB. This finding indicates that CLIP+CNT pair is capable of extracting task specific important features from multimodal memes which ultimately helps in increasing the performance of the main task.

(3) Another important observation is that when combining two modalities using either simple concatenation or inter-model attention (IA), CentralNet multitask framework with IA consistently outperforms the one with simple concatenation. But it is not always true for the models with simple feedBack multitask framework.

(4) We can observe that (CD+SA+ER) most of the times performs better compared to other three task variants, i.e., CD+SA+SAR and CD+ER+SAR for multi-modal inputs and achieved second highest f1 score of 73.73 for CD task. With CLIP+CNT combination, CD+SA+ER achieves 0.51% and 1.96%, improvements in F1 values for CD task over CD+SA+SAR and CD+ER+SAR, respectively. This gain in performance of CD+SA+ER is intuitive as sentiment or emotion alone can't always convey all the information about a user's mindset. We know, disgust, fear, sadness, and other unpleasant emotions can create a negative sentiment. Similarly, positive sentiment might arise due to emotions such as happiness, surprise, and so on. As a result, a person's actual state of mind cannot always be detected based on only sentiment.

(5) When we keep the same feature extractor module, a model with the CNT framework consistently outperforms a model with FB multitask framework. BERT-RN+CNT achieves on average 5% improvements in F1 values for CD task over BERT-RN+FB for both multimodal settings, i.e., with concatenation and with IA. This improvement indicates that CentralNet, which is well-known for multimodal data fusion, can be a suitable architecture for multitask learning.

(6) The result table illustrates that any multi-modal(text and Image) variant for both single-task and multi-task based models always performs better than the corresponding uni-modal variants. This improvement highlights the significance of including multimodal features for various memes analysis tasks. We have also

[5]https://scikit-learn.org/stable/
[6]ttps://keras.io/
[7]https://www.tensorflow.org/overview/

Table 7: Experimental results of different multitask variants with unimodal and multimodal settings. CD: Cyberbully Detection, SA: Sentiment Analysis, ER: Emotion Recognition, SAR: Sarcasm, FB: FeedBack, CNT: CentralNet, RN: ResNet, BT-RN: BERT+ResNet, IA: Inter-modal Attention.

Modality	Model	2-Task Variants						3-Task Variants						4-Task	
		CD+SA		CD+ER		CD+SAR		CD+SA+ER		CD+SA+SAR		CD+ER+SAR		CD+SA+ER+SAR	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Text (T)	BERT+ FB	62.03	61.53	60.40	60.01	61.09	60.82	62.43	62.53	62.96	62.47	61.58	61.38	62.14	62.13
	BERT+CNT	65.94	65.94	65.14	65.29	66.69	65.98	65.34	65.47	65.89	65.92	66.15	66.17	66.61	66.07
Image (I)	RN+FB	64.81	64.51	65.84	65.60	64.13	63.67	64.67	64.31	65.18	64.63	65.18	65.25	64.84	64.92
	RN+CNT	66.89	66.79	66.39	66.45	65.79	67.86	66.42	66.32	66.33	66.27	66.08	66.01	66.43	66.37
T+I with Concat	BT-RN+FB	65.86	65.74	66.52	66.54	65.87	65.82	67.21	67.13	65.23	65.11	65.52	64.96	66.87	66.76
	BT-RN+CNT	69.64	69.36	70.08	69.77	70.20	70.05	69.89	69.64	69.13	68.83	68.46	68.18	69.72	69.44
	CLIP+FB	72.24	72.28	72.16	72.23	72.66	72.68	71.06	71.07	71.85	71.93	71.32	71.35	71.21	71.31
	CLIP+CNT	72.88	72.82	73.03	72.95	72.07	71.96	73.05	72.98	73.05	72.97	73.11	73.02	73.16	73.06
T+I with IA	BT-RN+FB	65.36	65.12	66.82	66.76	66.52	66.41	67.35	67.42	66.15	66.08	65.93	65.12	66.74	66.79
	BT-RN+CNT	73.02	73.05	73.54	73.02	72.22	72.13	73.15	73.07	72.96	72.82	73.28	72.59	73.68	73.53
	CLIP+FB	71.99	72.01	72.75	72.79	71.18	71.18	71.06	71.07	72.33	72.33	71.32	71.35	72.44	72.47
	CLIP+CNT	73.28	73.17	72.66	**72.63**	71.12	71.00	73.79	**73.73**	73.31	73.22	71.85	71.77	74.17	**74.11**

Table 8: Single task results in terms of Accuracy (Acc) and F1 score. FC: Fully connected layer.

Modality	Model	CD		SA		ER		SAR		Harmfulness	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Text (T)	BERT+GRU+FC	61.14	60.73	56.91	54.24	31.23	23.22	59.72	59.12	60.86	60.33
Image (I)	RN+FC	63.36	62.37	58.39	55.61	30.83	23.19	59.39	57.79	62.51	62.14
(T+I) with	BERT+RN+FC	65.04	65.03	60.22	58.82	30.08	26.26	62.20	61.47	65.21	64.66
Concat	CLIP +FC	70.91	70.89	59.8	**59.16**	29.6	**27.96**	63.59	61.24	66.71	65.89
(T+I) with	BERT+RN+FC	65.63	65.41	61.02	59.11	30.12	25.39	62.12	62.75	65.28	65.14
IA	CLIP +FC	70.99	**71.01**	58.96	57.83	26.58	23.32	62.99	**63.80**	66.91	**66.28**

examined that in any uni-modal setting, image modality performs better than text modality.

(7) We are getting inferior results in case of emotion recognition. The possible reason would be the highly imbalanced nature of emotion classes in our dataset. However, we have already mentioned that our prime focus was boosting the performance of the main task (CD) with the help of other auxiliary tasks. That's why during training, we have given more weight-age to the loss of the main task (1.0), while weight-ages of other secondary tasks vary between 0.3 to 0.5.

All the reported results for the proposed model and baselines are statistically significant as we have performed statistical t-test at 5% significance level.

5.4 Error Analysis:

We have manually checked those data instances for error analysis which were misclassified by the proposed model, *CLIP+CNT*. Below in Figure 6, some examples are shown which are misclassified by our proposed model: (i) The true label of meme-1 is bully but our model has predicted it as non-bully. In the text portion, there is no offensive word or foul language. But after examining the text properly with background image as context, we can infer that some sarcasm is present there. Thus the underlying sarcasm present in this example was not understood by the proposed system. (ii) The proposed model has predicted meme-2 as bully though its true label

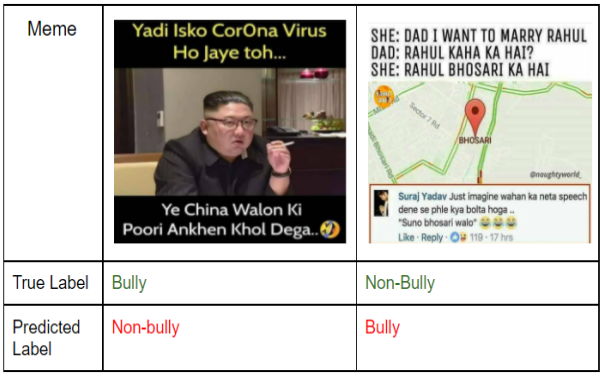


Figure 6: Misclassified Examples; Translation: Meme-1: He will open all the Chinese eyes if he is affected by the Coronavirus.; Meme-2: She: Dad, I want to marry Rahul; Dad: Where is Rahul from?; She: Rahul is from Bhosari.; Suraj Yadav: Just imagine how the politicians from Bhosari start their speeches, "listen bhosrian."

is non-bully. Actually in Hindi the word *"bhosari"* is used in the context of English word *"cunt"*. But here in this image *"bhosari"* is a place name. Hence the model is not able to identify whether it is abusive word or the name of a place. Also if you see a guy in

comment saying *"Suno bhosari walo"* (listen bhosrian) with *haha* emoji, he is actually not making any personal attack. As our model is not capable of processing any emoji so it is not able to identify the notion. Hence there is a misclassification.

6 CONCLUSION AND FUTURE WORK

In this paper, we are the first to introduce the task of sentiment-emotion-sarcasm aware multimodal cyberbully detection in code-mixed setting. In order to solve this task, we have created a novel multimodal memes dataset, *MultiBully*, annotated with bully, sentiment, emotion and sarcasm labels to determine if sentiment, emotion and sarcasm label information can assist in identifying cyberbully more accurately. We have introduced a new architecture, *CLIP-CentralNet*, an attention based multi-task multimodal framework for sentiment, emotion and sarcasm-aided cyberbullying detection. ResNet, mBERT and CLIP have been incorporated into our proposed model for efficient representations of different modalities and helping to learn generalized features over multiple tasks. Our developed *CLIP-CentralNet* framework outperforms all single task and uni-modal models, with a significant margin.

In the future, we would like to investigate the explainability of cyberbullying from memes and their targets.

ETHICS AND BROADER IMPACT

User Privacy.

Our dataset contains memes with annotation labels and no personal information about the users.

Biases.

Any biases detected in the dataset are inadvertent, and we have no intention of harming anyone or any group. We acknowledge that evaluating whether a meme is harmful can be subjective, so biases in our gold-labeled data or label distribution are unavoidable. Our high inter-annotator agreement gives us confidence that most of the time, the labels assigned to the data are correct.

Intended Use.

We share our data to promote more research on detecting cyberbullying from memes on the internet. We only release the dataset for research purposes and do not grant a license for commercial use.

ACKNOWLEDGMENTS

Dr. Sriparna Saha gratefully acknowledges the Young Faculty Research Fellowship (YFRF) Award, supported by Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia) for carrying out this research. The Authors would also like to acknowledge the support of Ministry of Home Affairs (MHA), India, for conducting this research.

REFERENCES

[1] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. 2016. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior* 63 (2016), 433–443.

[2] Vimala Balakrishnan, Shahzaib Khan, and Hamid R Arabnia. 2020. Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security* 90 (2020), 101710.

[3] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*. 36–41.

[4] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[5] Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4351–4360.

[6] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[7] Maral Dadvar, FMG de Jong, Roeland Ordelman, and Dolf Trieschnigg. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

[8] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*. Springer, 693–696.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International Conference on Weblog and Social Media 2011*. Citeseer.

[11] Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. 29–30.

[12] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524* (2017).

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

[14] Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2021. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cognitive Computation* (2021), 1–20.

[15] 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*. 1470–1478.

[16] 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*. 770–778.

[17] Homa Hosseinmardi, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2016. Prediction of cyberbullying incidents in a media-based social network. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 186–192.

[18] Satyajit Kamble and Aditya Joshi. 2018. Hate speech detection from code-mixed hindi-english tweets using deep learning models. *arXiv preprint arXiv:1811.05145* (2018).

[19] 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. *arXiv preprint arXiv:2005.04790* (2020).

[20] Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. *arXiv preprint arXiv:1803.09402* (2018).

[21] Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. 2010. *Handbook of emotions*. Guilford Press.

[22] Krishanu Maity, Abhishek Kumar, and Sriparna Saha. 2022. A Multi-task Multimodal Framework for Sentiment and Emotion aided Cyberbully Detection. *IEEE Internet Computing* (2022).

[23] Krishanu Maity and Sriparna Saha. 2021. BERT-Capsule Model for Cyberbullying Detection in Code-Mixed Indian Languages. In *International Conference on Applications of Natural Language to Information Systems*. Springer, 147–155.

[24] Krishanu Maity and Sriparna Saha. 2021. A Multi-task Model for Sentiment Aided Cyberbullying Detection in Code-Mixed Indian Languages. In *International Conference on Neural Information Processing*. Springer, 440–451.

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

[26] Sayanta Paul and Sriparna Saha. 2020. CyberBERT: BERT for cyberbullying identification. *Multimedia Systems* (2020), 1–8.

[27] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Akhtar, Preslav Nakov, Tanmoy Chakraborty, et al. 2021. Detecting harmful

memes and their targets. *arXiv preprint arXiv:2110.00413* (2021).

[28] Michal Ptaszynski, Fumito Masui, Taisei Nitta, Suzuha Hatakeyama, Yasutomo Kimura, Rafal Rzepka, and Kenji Araki. 2016. Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization. *International Journal of Child-Computer Interaction* 8 (2016), 15–30.

[29] 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. *arXiv preprint arXiv:2103.00020* (2021).

[30] Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, Vol. 2. IEEE, 241–244.

[31] Hugo Rosa, Joao P Carvalho, Pável Calado, Bruno Martins, Ricardo Ribeiro, and Luisa Coheur. 2018. Using fuzzy fingerprints for cyberbullying detection in social networks. In *2018 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 1–7.

[32] Hugo Rosa, Nádia Pereira, Ricardo Ribeiro, Paula Costa Ferreira, Joao Paulo Carvalho, Sofia Oliveira, Luísa Coheur, Paula Paulino, AM Veiga Simão, and Isabel Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (2019), 333–345.

[33] Tulika Saha, Apoorva Upadhyaya, Sriparna Saha, and Pushpak Bhattacharyya. 2021. A Multitask Multimodal Ensemble Model for Sentiment-and Emotion-Aided Tweet Act Classification. *IEEE Transactions on Computational Social Systems* (2021).

[34] 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 (2008), 376–385.

[35] 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*. 32–41.

[36] Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *International Conference Recent Advances in Natural Language Processing (RANLP)*. 672–680.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[38] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[39] Elizabeth Whittaker and Robin M Kowalski. 2015. Cyberbullying via social media. *Journal of school violence* 14, 1 (2015), 11–29.

[40] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. 188–202.