Emoji, Sentiment and Emotion Aided Cyberbullying Detection in Hinglish

Krishanu Maity, Sriparna Saha, *Senior Member, IEEE*, and Pushpak Bhattacharyya, *Senior Member, IEEE*

***Abstract*— The advent of the Internet is a boon to society. However, many of its banes cannot be undermined, cyberbullying being one of them. The emotional state and sentiment of a person have a significant influence on the intended content. The current work is the first attempt in investigating the role of sentiment and emotion information for identifying cyberbullying in the Indian scenario. From Twitter, a benchmark Hind–English code-mixed corpus called BullySentEmo has been developed as there is no dataset available labeled with bully, sentiment, and emotion. Moreover, emoji information available with tweet texts can provide better understanding of user intention. The developed dataset consists of both modalities, tweet text, and emoji. In India, the majority of communication on different social media platforms is based on Hindi and English and language switching is a common practice in digital communication. An attention-based multimodal, adversarial multitasking framework is proposed for cyberbully detection (CBD) considering two auxiliary tasks: sentiment analysis (SA) and emotion recognition (ER). Experimental results indicate that compared to unimodal and single-task variants, the proposed framework improves the performance of the main task, i.e., CBD, by 3.59% and 2.56% in terms of accuracy and F1-Score, respectively. Furthermore, two different benchmark datasets (Twitter dataset and Aggression dataset) have been considered to show the robustness of our proposed model.**

***Index Terms*— Code mixed, cyberbullying, emotion, multitasking (MT), sentiment.**

I. Introduction

CYBERBULLYING [1] is described as the serious, intentional, and repetitive acts of a person's cruelty toward others using various digital technologies. It is mainly expressed through nasty tweets, texts, or other social media posts. This further instigates several adverse effects of victims, such as depression, hopelessness, psychosomatic problems, and loss of self-esteem; also, attempts or actual suicide are seen in users. According to data given by the National Crime Records Bureau, from 2017 to 2018, the number of cases of cyberbullying against women or children in India increased by 36%.[1] Young people who are subjected to cyberbullying are more likely to engage in self-harm and suicide conduct than those who are not. As a result, spotting cyberbullying early on is crucial for preventing its consequences.

The process of fluidly flipping between two or more languages in a discussion is known as code mixing (CM) [2]. India is a country of many languages and most of the times people use mixture of two languages for regular post and conversations on social media. About 691 million native speakers use Hindi as one of the official Indian languages.[2] 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.

The emotional state and sentiments of a person have a significant influence on the intended content [3]. Sentiment and emotion are inextricably linked, and one aids in the comprehension of the other. Emotions, such as happiness and joy, are intrinsically associated with positive sentiments. It is common knowledge that a tweet labeled as a bully usually reflects negative sentiment. Because of the strong correlation between emotion and sentiment, we should consider the sentiment of the tweeter as well as its emotion while predicting bullying tweets.

There are several works in the literature where the tasks of sentiment analysis (SA) and emotion recognition (ER) are treated as auxiliary tasks to boost the performance of primary task (such as sarcasm detection [4] and tweet act classification (TAC) [5]) in a multitasking (MT) framework. SA [6] aims to extract subjective information from different social media posts and categorize them into three predetermined labels: positive, negative, or neutral. Multitask learning is proven to be effective when working on related tasks [7]. The use of domain-specific information to related tasks enhances the overall learning process.

Furthermore, multimodal inputs, such as a combination of text and other nonverbal clues (emojis in tweets) [8], contribute in creating trustworthy classification models that aid in detecting the tweeter's emotional state and sentiment, which in turn assist the cyberbully detection (CBD) task. Emojis are becoming increasingly popular in social media posts because they take less typing effort, have a smaller vocabulary, and help to draw attention.

The aim of this article is to develop a multitask multimodal framework for cyberbullying detection in Hindi–English code-mixed data where SA and ER act as the secondary/auxiliary

Manuscript received 4 January 2022; revised 24 April 2022 and 30 May 2022; accepted 2 June 2022. Date of publication 7 July 2022; date of current version 2 October 2023. *(Corresponding author: Krishanu Maity.)*

Krishanu Maity and Sriparna Saha are with the Department of Computer Science and Engineering, IIT Patna, Patna 801103, India (e-mail: krishanu_2021cs19@iitp.ac.in; sriparna@iitp.ac.in).

Pushpak Bhattacharyya is with the Department of Computer Science and Engineering, IIT Bombay, Mumbai 400076, India (e-mail: pb@cse.iitb.ac.in).

Digital Object Identifier 10.1109/TCSS.2022.3183046

[1] https://ncrb.gov.in/en/crime-india-2018-0

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

tasks to increase the performance of primary task, i.e., CBD. Our developed model multitask multimodal cyberbully detection (MM-CBD) utilizes the BERT language model [9] for efficient representation of the code-mixed text data. Multilingual Representations for Indian Languages (MuRIL) BERT[3] was employed, which was pretrained on Indian languages and their transliterated equivalents.

The main contributions of this work are listed as follows.

1) We have created a new code-mixed corpus called BullySentEmo of tweets (text + emoji) annotated with bully, sentiment, and emotion labels. For further research work on sentiment and emotion-aware CBD, this dataset will certainly help a lot.[4]
2) We have addressed the need for considering the sentiment and emotion label information of the tweets while identifying cyberbully.
3) We have proposed an end-to-end dyadic attention mechanism (DAM)-based multitask multimodal framework called MM-CBD for sentiment and emotion-aided cyberbullying detection.
4) Experimental results illustrate the efficacy of solving the CBD, SA, and ER tasks together in a multitask framework. Multimodal (usage of emoji) and multitask CBD outperforms unimodal and single-task CBD by a substantial margin.

The organization of this article is as follows. A survey of previous works on cyberbullying detection has been exhibited in Section II. Section III describes the details of the creation and annotation of the proposed dataset. The proposed methodology has been explained in Section IV. Experimental evaluations along with the error analysis have been presented in Section V. The conclusion and future works are mentioned in Section VI.

II. Related Works

With the advancement of natural language processing (NLP), many studies on the identification of cyberbullying have been conducted in the English language rather than other languages [10]. Some of them are listed here. Dinakar *et al.* [11] 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 support vector machine (SVM) classifier and 63% with Naive Bayes classifier. Cheng *et al.* [12] introduced the generation of a temporal-based feature to detect cyberbullying by extracting interarrival time. They were able to achieve 72.60% accuracy with the XGBoost classifier. Reynolds *et al.* [13] worked on data collected from Formspring.me and labeled using web service to train their model, and they had used a Weka tool kit and were able to achieve 78.5% accuracy by using C4.5 decision tree learner. Pawar and Raje [14] and Haidar *et al.* [15] developed a multilingual cyberbullying detection system based on machine learning algorithms to detect English, Hindi, Marathi, and Arabic bullying messages.

In 2020, Paul and Saha [16] developed a Bidirectional Encoder Representations from Transformers (BERT)-based framework, namely, cyberBERT, for cyberbully identification. They have evaluated cyberBERT on three benchmark datasets, i.e., Formspring (12k posts), Twitter (16 k posts), and Wikipedia (100 k 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. Djuric *et al.* [17] 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 dataset of user comments gathered from the Yahoo Finance website and found it 80.01% accurate. Ghosh *et al.* [18] collected tweets from various domains, such as religion, ethnicity, nationality, terrorism, and politics, to form a multidomain hate speech corpus (MHC). To identify hate speech from Twitter data, they developed a stacked ensemble-based hate speech classifier (SEHC) based on different well-known deep learning models, such as convolutional neural network (CNN), gated recurrent unit (GRU), and long short-term memory (LSTM). Pretrained Glove embeddings have been employed to get the vector representations of input words.

Also, a few works are present for the code-mixed dataset with supervised approach. Bohra *et al.* [19] developed a code-mixed dataset of 4575 tweets and annotated with hate speech and normal speech. The SVM classier achieved a 71.7% accuracy score when word n-grams, punctuations, character n-grams, hate lexicon, and negation words are considered as feature vectors. Ghosh *et al.* [20] developed a multilayer perceptron model as the classifier. They also created a code-mixed corpus collected from Facebook and labeled it manually. They extracted various features from them and obtained an accuracy of 68.50%. Maity and Saha [21] developed a Hindi–English code-mixed text corpus from Twitter for cyberbullying detection. They developed a model based on deep learning architectures that include GRU, CNN, BERT, and capsule networks and attained 79.28% accuracy.

Saha *et al.* [5] proposed a multitask ensemble adversarial learning framework for multimodal TAC and claimed that TAC performs significantly better than its unimodal and single-task TAC variants. Ghosh *et al.* [22] suggested a multitask learning architecture that uses external knowledge information to improve the overall performance on the emotion classification task on suicide notes. To analyze the effects of sentiment and emotion on the sarcasm detection task, Chauhan *et al.* [4] presented a multitask framework based on intersegment and intrasegment attention mechanisms in 2020.

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

III. Code-Mixed BullySentEmo-Annotated Corpora Development

First, we combed the literature for the code-mixed CBD dataset annotated with another two labels, i.e., sentiment

[3] https://tfhub.dev/google/MuRIL/1

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

TABLE I
COMPARISON BETWEEN BULLYSENTEMO AND OTHER EXISTING WIDELY USED CYBERBULLY DATASETS

Dataset	Language	Instance	Balancing	Modality
Formspring1 [13]	English	3915	14.2%	Text
YouTube2 [23]	English	4626	9.7%	Text
MySpace2 [24]	English	2200	–	Text
AskFM [25]	Dutch	85485	6.7%	Text
Schoolboard Bulletins (BBS) [26]	Japanese	2222	12.8%	Text
Twitter3 [27]	English	10007	6%	Text
Instagram [28]	English	1954	29%	Text
Formspring4 [29]	English	13160	19.4%	Text
Aggression Annotated Dataset[30]	Hindi+English	39000	57.4%	Text
Code-mixed-bully [21]	Hindi+English	5062	51.48%	Text
BullySentEmo (Our Dataset)	Hindi+English	5062	51.48%	Text+Emoji

and emotion. To the best of our knowledge, there is only one publicly available Hindi–English code-mixed corpus for cyberbullying detection. Thus, to achieve our goal (solving the MT model), we have further annotated the CBD dataset proposed in [21] with three sentiment classes and seven emotion classes.

A. Data Collection

To create a BullySentEmo dataset, we have considered standalone (only bully or nonbully label) Hindi–English code-mixed cyberbully dataset [21], which we have considered for further annotation with three sentiment classes and seven emotion classes, containing a total of 5062 tweets where 2456 tweets were marked as nonbully and the remaining 2606 tweets were marked as bully. The ratio between the bully and nonbully tweets in this corpus is 48.52:51.48. One of the key objectives of this work is to develop a corpus that will help further research on sentiment and emotion-aware CBD.

B. Data Annotation

Three annotators having proficient linguistic background in both Hindi and English were involved in data annotation. We have given them detailed instructions of annotations with examples and closely monitor the annotation process. They have worked in isolation to avoid any biasness. For each tweet, annotators had tagged two labels: one for the sentiment class (positive/neutral/negative) and another for the emotion class based on six Eckman's [31] emotion categories (happiness, sadness, fear, surprise, anger, disgust, and others). For finalizing the annotation label of each tweet, we use a majority vote procedure to resolve conflicts between annotators. We calculated the interannotator agreement (IAA) using Fleiss' [32] Kappa score to verify the quality of annotation. We attained the agreement scores of 0.81, 0.82, and 0.72 on the CBD, SA, and ER tasks, respectively.

Some samples from our annotated dataset are shown in Table II. These examples show that emoji information helps in identification of bullying comments. The text and usage of emoji together are used for annotation. For example, in the third instance (T3), no such vulgar words are present in the sentence. However, when we look into the emoji part, there are three emojis, i.e., winking face, face savoring food, and squinting face with tongue. This emoji part makes the entire sentence into a negative and sexist one. The combination of text and emoji helps in understanding the inherent intention of the user.

C. Dataset Statistics

Fig. 1(a)–(c) shows dataset statistics based on bully, sentiment, and emotion classes, respectively. Negative sentiment labels appear in 2744 data points in our corpus, whereas positive and neutral sentiment labels appear in 1252 and 1066 tweets, respectively. Fig. 2(a) shows substantial correlations between the bully and negative sentiment classes, indicating that most of the bully labeled tweets have a negative sentiment. On the other hand, nonbully tweets have a sentiment of neutral and positive. On the other hand, Fig. 2(b) shows a high correlation between the bully versus anger and disgust emotion classes, indicating that a tweet labeled as the bully is more probably to have anger or disgust emotion.

If we see classwise statistics in Fig. 1, the bully class is almost balanced, but the sentiment and emotion classes are imbalanced. These imbalanced statistics of sentiment and emotion classes occur due to the nature of our problem statement. We have noticed that a tweet labeled as a bully has negative sentiment most of the time. Thus, if we try to balance the bullying class, almost 50% of samples belong to the bully class, and then, almost 50% of samples will have negative sentiments. The remaining 50% nonbully tweets have either positive or neutral sentiments. Hence, if we try to balance the bullying class, then sentiment class would not be balanced and vice versa. As sentiment and emotion are inextricably linked, one aids in comprehending the other. Emotions, such as happiness and joy, are intrinsically associated with positive sentiments. On the other hand, emotions, such as sadness, disgust, and anger, are linked with negative sentiments. Thus, it is obvious that in our problem statement, if the bully class is balanced, then the sentiment and emotion classes will be imbalanced.

D. Dataset Comparison

Table I compares BullySentEmo to several frequently used cyberbully datasets. Table I shows that our corpus has some unique aspects/characteristics compared to other datasets: 1) BullySentEmo is the first multimodal cyberbully dataset having two modalities, emoji and text; 2) it is not only manually annotated with cyberbully labels but also with sentiment and emotion labels to solve the task of sentiment and emotion-aware CBD and this dataset can solve three tasks simultaneously, i.e., CBD, SA, and ER; and 3) the texts present in this dataset are in Hindi–English code-mixed form rather than English.

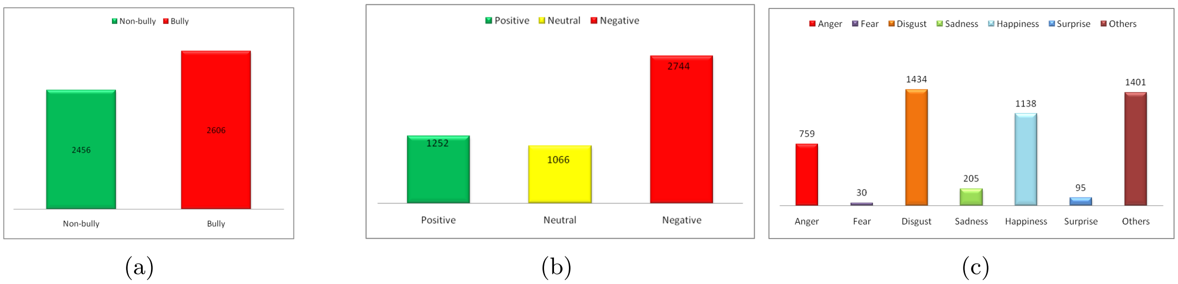
IV. METHODOLOGY

This section describes the MM-CBD framework that we have developed to identify cyberbullying. Figs. 3 and 4 show the overall architecture of MM-CBD model.

TABLE II

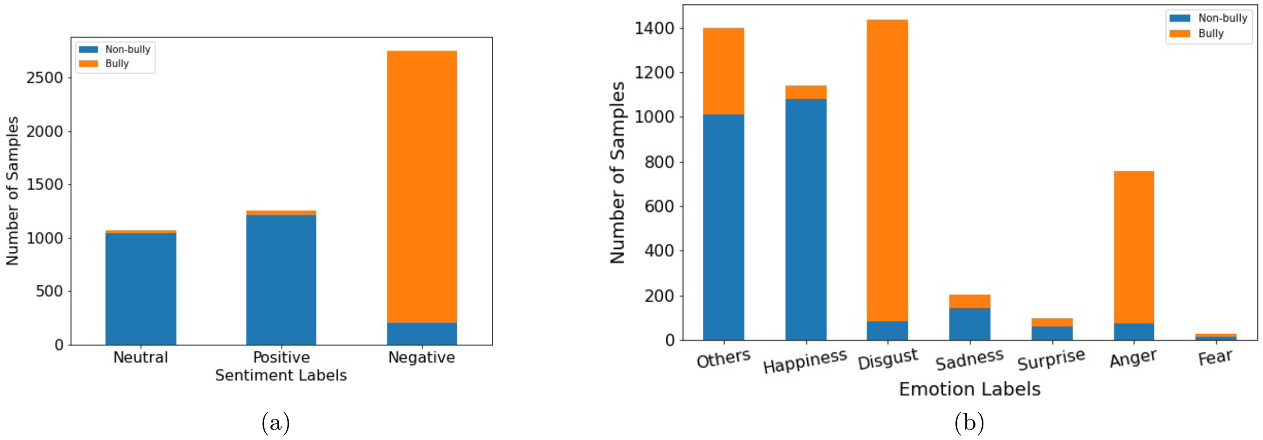
SOME SAMPLES FROM ANNOTATED DATASET WITH SENTIMENT, EMOTION, AND BULLY LABELS

Tweet	Sentiment Class	Emotion Class	Bully Class
T1: Did you just call yourself a liberal? Tu gadhi hai gadhi, not liberal 😠 **Translation**: Did you just call youself a liberal? You are sordid, not liberal	Negative	Anger	Bully
T2: Itihas gawa hai har khubsurat ladki Kisi chutiya se he pyaar karte hai 😞 😡 👿 **Translation**: History says, every beautiful women likes brawler man	Negative	Disgust	Bully
T3: Matlab Aunty bhi khubsurat aur unki beti bhi. 😉😋😝 **Translation**: It means aunty and her daughter both are lookinh hot	Negative	Surprise	Bully
T4: Bahut khubsurat lag rahe ho aap aur aap ki smile super hai ji 🥰 **Translation**: You are looking beautiful and your smile is superb	Positive	Happiness	Non-bully



(a) (b) (c)

Fig. 1. Classwise statistics. (a) Cyberbully statistics. (b) Sentiment statistics. (c) Emotion statistics.



(a) (b)

Fig. 2. Correlation between different classes. (a) Bully versus sentiment. (b) Bully versus emotion.

A. Feature Extraction

The procedure for extracting features from different modalities is described as follows.

1) Textual Features: The BERT language model has been employed to get the textual features from input tweets $W = \{w_1, w_2, \ldots, w_{n_x}\}$. Pooled and sequence of these two types of output have been generated by BERT. A $[batch\ size, 768]$ shaped pooled output represents the entire sentence using a special token called CLS. 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. We have considered sequence output for the MM-CBD model and pooled output has been considered for machine learning-based baselines.

2) Emoji Features: We leverage emoji, a python-based library for extracting the visual representation of an emoji (mainly that of a face, object, or symbol) from a tweet. There are 1816 distinct types of emojis in the emoji library. Following that, we used emoji2vec [33] to generate $d_y = 300$-D vector representation for each of the emojis retrieved from the tweet. Let us assume that a tweet contain n_y number of emojis, and then, the final emoji representation of a tweet is obtained as $Y \in \mathbb{R}^{n_y \times d_y}$.

B. Bi-GRU Layer

To preserve the contextual and semantic information between words in a tweet, the word vectors from both BERT and emoji2vec are passed through a Bi-GRU layer [34]. To capture long-term dependency of input word vectors, Bi-GRU encodes the input on both forward and backward direction as

$$\overrightarrow{h}_t^i = \overrightarrow{\text{GRU}}\left(w_t^i,\ h_{t-1}^i\right), \quad \overleftarrow{h}_t^i = \overleftarrow{\text{GRU}}\left(w_t^i,\ h_{t+1}^i\right) \qquad (1)$$

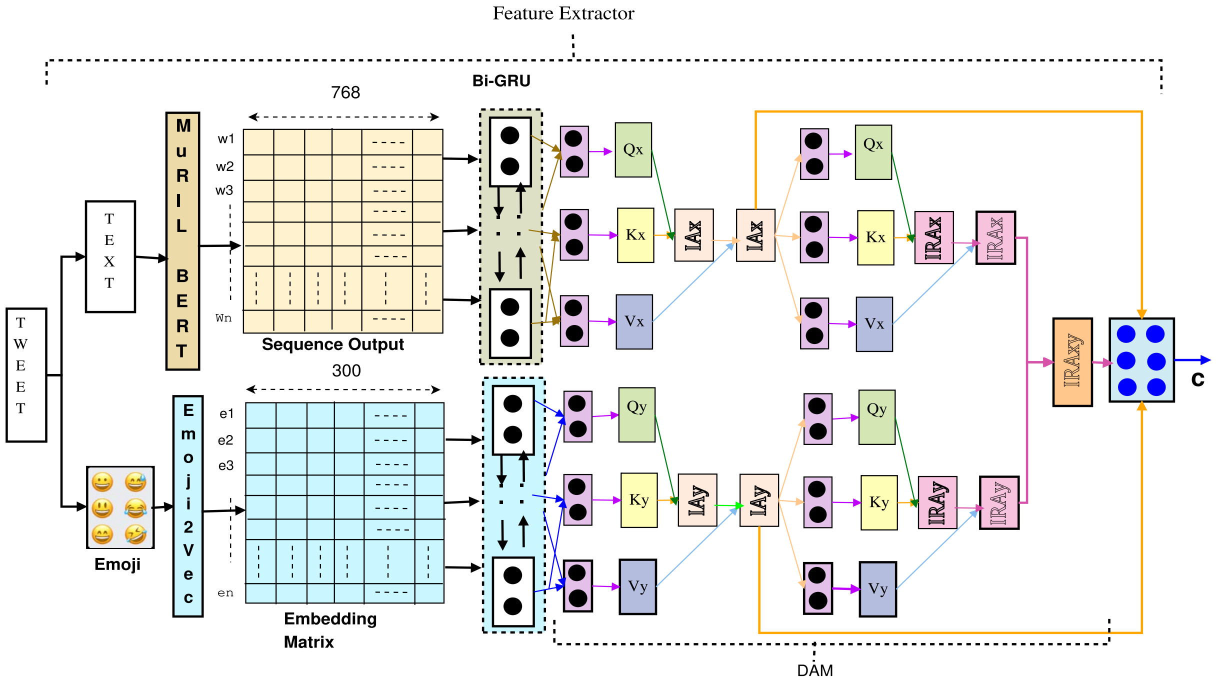


Fig. 3. Feature extraction module of the proposed MM-CBD architecture. IA: intramodal attention. IRA: intermodal attention. DAM: dyadic attention mechanism.

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 $\overrightarrow{\text{GRU}}$ and $\overleftarrow{\text{GRU}}$, respectively

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

Let $H_x \in \mathbb{R}^{n_x \times 2d_g}$ and $H_y \in \mathbb{R}^{n_y \times 2d_g}$ be the final hidden state matrices of tweet text and emoji modality, respectively. The number of hidden units in bi-GRU is represented by d_g.

C. Dyadic Attention Mechanism

We apply an approach similar to that described in [35], in which authors suggested that attention should be computed by mapping a query and a set of key–value pairs to an output. Bi-GRU generated outputs of both the modalities (H_x and H_y) are passed through three fully connected (FC) layers, namely, queries (Q), keys (K), and values (V) of dimension d_f. We have two triplets of (Q, K, V) as: (Q_x, K_x, V_x) and (Q_y, K_y, V_y) for the fully shared (FS) model. On the other hand, there are six such triplets for the shared-private (SP) model, as we have three tasks and two modalities for each task. These triplets are then used to calculate attention values by combining them in various ways, including intramodal attention (IA) and intermodal attention (IRA).

1) Intramodal Attention: To understand the interdependence between the current words and the former portion of the tweet, we compute IA for each of these distinct modalities. Individual modalities' IA scores are computed as follows:

$$\text{IA}_i = \text{softmax}\left(Q_i K_i^T\right) V_i \tag{3}$$

where $\text{IA}_x \in \mathbb{R}^{n_x \times d_f}$ and $\text{IA}_y \in \mathbb{R}^{n_y \times d_f}$.

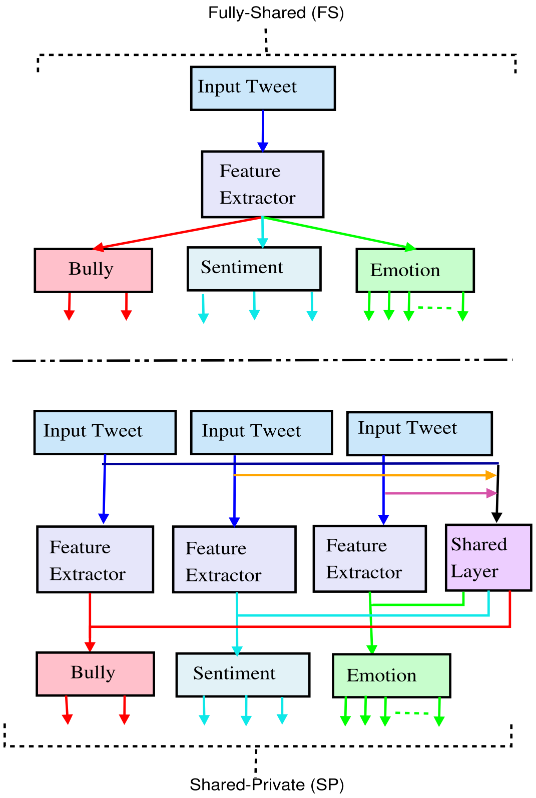


Fig. 4. FS and SP multitask modules of MM-CBD model.

2) Intermodal Attention: The IA scores mentioned above are then applied to calculate the IRA. We repeat the same method (described above) to generate (Q, K, V) triplets for these IA scores. Based on (3), by computing the matrix multiplication of a combination of queries and keys of distinct IA modality scores, we obtain IRA scores among triplets of all IA scores. In this way, for the FS model, we have one IRA score as $\mathrm{IRA}_{xy} \in \mathbb{R}^{n_x \times d_f}$ and three IRA scores for SP model as IRA_{xy1}, IRA_{xy2}, and IRA_{xy3}.[5] This is performed to differentiate significant contributions from multiple modalities to achieve a better representation of tweets.

3) Attention Fusion: All of these calculated IA and IRA vectors are concatenated as follows:

$$\begin{aligned} C &= \mathrm{concat}\big(\mathrm{IRA}_{xy}, \mathrm{IA}_x, \mathrm{IA}_y\big), \quad \text{for FS} \\ C_1 &= \mathrm{concat}\big(\mathrm{IRA}_{xy1}, \mathrm{IA}_{x1}, \mathrm{IA}_{y1}\big), \quad \text{for SP} \\ C_2 &= \mathrm{concat}\big(\mathrm{IRA}_{xy2}, \mathrm{IA}_{x2}, \mathrm{IA}_{y2}\big), \quad \text{for SP} \\ C_3 &= \mathrm{concat}\big(\mathrm{IRA}_{xy3}, \mathrm{IA}_{x3}, \mathrm{IA}_{y3}\big), \quad \text{for SP.} \end{aligned} \tag{4}$$

For SP variants, the final representation of a tweet is generated by taking the mean of these three separate concatenated attention vectors, C_1, C_2, and C_3. For FS variants, C attention vector has been considered as the final representation of input tweet.

4) Shared Layer: In addition to task-specific layers, we include a shared layer to learn task-invariant features for the SP model. The shared layer is an FC layer of dimension d_f. The hidden representations of three IRA vectors, IRA_{xy1}, IRA_{xy2}, and IRA_{xy3}, are sent to the shared layer.

5) Adversarial Loss: The adversarial loss function's objective is to fine-tune the shared layer's weights to learn a representation that deceives the task discriminator. The adversarial loss, l_{adv}, aims to make shared and task-specific layers' feature spaces mutually exclusive [36]. We use a method similar to [36], in which a task discriminator (D) maps the shared feature to its original task. The adversarial loss function is computed as follows:

$$l_{\mathrm{adv}} = \min_F \left(\max_D \left(\sum_{m=1}^{M} \sum_{k=1}^{K} y_k^m \log\big[D\big(\mathrm{F}\big(x_k^m\big)\big)\big] \right) \right) \tag{5}$$

where M is the number of tasks, y_k^m represents true label among the type of the tasks, and x_k^m is the kth sample for task m. Finally, to optimize the min–max function, we use the gradient reversal layer [37].

D. Classification Layer

For the FS model, the final output from the DAM module is shared over three channels, one for each of the three tasks, i.e., CBD, SA, and ER. For the SP model, individual DAM representation from each of the three tasks is passed through the corresponding task-specific output layer. Shared loss (l_s), task-specific loss (l_t), and adversarial loss (l_{adv}) are the three types of losses that are employed in our model as follows:

$$l_f = \begin{cases} l_s + \gamma\, l_{\mathrm{adv}}, & \text{for FS model} \\ l_t + \alpha l_s + \gamma\, l_{\mathrm{adv}}, & \text{for SP model} \end{cases} \tag{6}$$

where $\gamma = 0.01$ and $\alpha = 0.5$ are the hyperparameters.

V. Experimental Results and Analysis

The findings of several baseline models and our proposed model are shown in this section, which was tested on our proposed Hindi–English code-mixed corpus. All our experiments were conducted on a hybrid cluster of multiple GPUs comprised of RTX 2080Ti. We have performed stratified fivefold cross validation on our dataset and reported the mean metric scores.

A. Hyperparameters

We use Tanh activation in bi-GRU (128 hidden cell) and ReLU activation in all FC layers ($d_f = 100$). With a batch size of 32, we train our models for ten epochs. We utilize the Adam optimizer and set the learning rate to 0.001 to backpropagate the loss across the network.

B. Baselines Setup

1) *MuRIL BERT+LR:* Input tweets are passed through MuRIL BERT followed by the LR classifier for unimodal framework. For multimodal, we concatenated BERT's text representation with emoji representation from emoji2vec, and then, the concatenated vectors were fed into the classifier. We have considered pooled output from BERT with dimension 768.
2) *MuRIL BERT+SVM:* Same as the previous one with one modification: LR is replaced by an SVM classifier.
3) *MuRIL BERTzMLP:* MuRIL BERT's pooled output is fed to a MLP, where we have kept two FC layers (100 neurons per layer) followed by an output layer.

As we have three tasks, including the main task (CBD), there are three multitask variants. The first one is MT(CBD + SA), where SA acts as a secondary task, and the second is MT(CBD + ER), where ER has been considered as a secondary task. The third one is MT(CBD + SA + ER), where SA and ER have been considered as secondary tasks. Unimodal experiments indicate that we have considered only the text part, and in case of a multimodal experiment, both text and emoji data have been considered as inputs. Furthermore, we have conducted our experiments on the different combinations of MT approaches (FS and SP), with and without adversarial (Adv) training. The combinations are given as follows.

1) *FS:* For multitask variants, tweets are passed through a feature extractor module followed by three task-specific output layers without adversarial loss. The numbers of neurons in bully, sentiment, and emotion output layers are 2, 3, and 7, respectively. In case of single-task setting, feature extractor module is followed by only one specific task.
2) *FS + Adv:* This is the same as previous one with adversarial training.
3) *SP:* We incorporate three task-specific feature extractors without DAM and Adv.
4) *SP+ + Adv:* This is basically SP with adversarial training.

[5] CBD, SA, and ER tasks are denoted by subscripts 1, 2, and 3, respectively.

TABLE III
EXPERIMENTAL RESULTS OF DIFFERENT BASELINES AND THE PROPOSED MODEL FOR CBD TASK

Model	Uni-modal (Text)		Multi-modal (Text + Emoji)	
	Accuracy	F1-Score	Accuracy	F1-Score
Single Task				
BERT+LR	72.17(+/-0.13)	72.22(+/-0.13)	73.34(+/-0.09)	73.51(+/-0.08)
BERT+SVM	74.15(+/-0.05)	74.12(+/-0.06)	74.98(+/-0.05)	74.75(+/-0.04)
BERT+MLP	74.88(+/-0.75)	74.76(+/-0.86)	75.78(+/-0.97)	75.88(+/-0.93)
FS	77.57(+/-0.42)	77.43(+/-0.12)	78.11(+/-0.59)	78.05(+/-0.23)
Multi-task (CBD+SA)				
FS	78.50(+/-0.14)	78.71(+/-0.15)	79.35(+/-0.13)	79.21(+/-0.12)
FS+Adv	79.99(+/-0.12)	79.66(+/-0.12)	80.35(+/-0.14)	80.29(+/-0.13)
SP	78.30(+/-0.14)	78.16(+/-0.08)	80.27(+/-0.07)	80.19(+/-0.05)
SP+Adv	79.52(+/-0.08)	79.46(+/-0.15)	80.57(+/-0.16)	80.49(+/-0.15)
SP+DAM+Adv	80.55(+/-0.13)	80.34(+/-0.12)	81.77(+/-0.12)	81.73(+/-0.12)
Multi-task (CBD+ER)				
FS	78.25(+/-0.12)	78.12(+/-0.13)	78.69(+/-0.4)	78.78(+/-0.15)
FS+Adv	79.31(+/-0.11)	79.28(+/-0.09)	79.88(+/-0.07)	79.96(+/-0.20)
SP	77.89(+/-0.11)	77.89(+/-0.21)	78.58(+/-0.45)	78.37(+/-0.87)
SP+Adv	79.23(+/-0.14)	79.12(+/-0.24)	80.12(+/-0.09)	80.15(+/-0.11)
SP+DAM+Adv	80.05(+/-0.12)	80.25(+/-0.13)	81.11(+/-0.16)	81.09(+/-0.19)
Multi-task (CBD + SA +ER)				
FS	79.20(+/-0.11)	79.23(+/-0.11)	80.18(+/-0.18)	80.14(+/-0.13)
FS+Adv	80.29(+/-0.13)	80.17(+/-0.12)	81.05(+/-0.15)	81.87(+/-0.12)
SP	80.13(+/-0.17)	80.09(+/-0.15)	80.89(+/-0.03)	80.77(+/-0.21)
SP+Adv	80.76(+/-0.11)	80.88(+/-0.14)	81.71(+/-0.13)	81.75(+/-0.09)
SP+DAM+Adv	81.23(+/-0.17)	81.26(+/-0.14)	**82.87(+/-0.15)**	**82.86(+/-0.11)**

5) *SP + DAM + Adv:* This is our main model (MM-CBD), which incorporates both DAM and Adv in the SP framework.

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

C. Results and Discussion

As our proposed model is a multitask framework where three tasks, SA, ER as auxiliary tasks, and CBD as primary task, are simultaneously solved, it requires a dataset with three labels (sentiment, emotion, and cyberbully) annotated for training. First, we have experimented with our developed BullySentEmo dataset. Furthermore, to show the robustness of our proposed model, we have continued our experiments on two benchmark datasets; one is a code-mixed Aggression dataset, and another is an English Twitter dataset.

1) Result Analysis on BullySentEmo Dataset: Table III presents the results of CBD (main task) on our developed BullySentEmo dataset in terms of accuracy and F1-Score. From the table, we find that the following conditions hold.

1) All the multitask variants outperform the single-task classifiers for the CBD task.
2) In case of multitask (CBD + SA + ER), SP + DAM + Adv performs better than SP + Adv with the accuracy improvements of 0.47% and 1.16% for the unimodal and multimodal settings, respectively. We can also observe a similar scenario, i.e., SP with DAM outperforms SP without DAM for the multitask (CBD + SA) and multitask (CBD + ER) variants. This observation implies that SP variants with DAM perform better than the one without DAM.
3) Any multimodal (text + emoji) variant for both FS and SP models always performs better than the corresponding unimodal variants. This improvement highlights the significance of including multimodal features for various Twitter analysis tasks.
4) Table III also shows that MT-(CBD + SA + ER) consistently produced superior results than bitask variants, i.e., MT-(CBD + SA) and MT-(CBD + ER). This gain in performance of MT-(CBD + SA + ER) is intuitive as sentiment or emotion alone cannot always convey all of the information about a tweeter's mindset. We know that disgust, fear, sadness, and other unpleasant emotions can create a negative sentiment. Similarly, positive sentiment might arise due to emotions such as happiness and surprise. As a result, a person's actual state of mind cannot always be detected based on the sentiment.
5) When comparing several MT methods, such as FS, SP, and adversarial loss (Adv), it was examined that the SP model most of the times outperformed the FS model. Incorporating adversarial loss improved the performance of several multitask models with a significant margin.

Experimental results of this work demonstrate that concurrent execution of three tasks, CBD, SA, and ER, improves the performance of the main task (CBD). Furthermore, the multimodal framework outperforms the unimodal framework by a substantial margin since emoji data add some extra information to inputs, which helps the classifier for better prediction. All the reported results for the proposed model and baselines are statistically significant.

TABLE IV

DATASET STATISTICS OF AGGRESSION, TWITTER, AND TWITTER+ DATASETS

Dataset	#Posts			Total	# Class
Aggression	CAG 6072	CAG 6115	NAG 2813	15000	3
Twitter	Sexism 3117	Racism 1937	None 11036	16090	3
Twitter+	Sexism 9351	Racism 9685	None 11036	30072	3

TABLE V

RESULTS OF STATE-OF-THE-ART MODELS AND THE PROPOSED MODEL ON THE AGGRESSION DATASET

Model	F1-score
Dense Neural Network(Input-1024-512-256-Output) [42]	59.51
Ensemble method(CNN, LSTM, and Bi-LSTM) [43]	60.37
Single Task - FS (Our model)	**62.59**

2) Result Analysis on Twitter and Aggression Datasets: The Aggression dataset [38] has 15 000 Facebook posts/comments in Hindi and English, where each post is labeled into one of the three classes, namely, Nonaggressive (NAG), Overtly Aggressive (OAG), and Covertly Aggressive (CAG). The second is a publicly available Twitter [39] dataset consisting of 16 090 tweets and manually annotated with three classes (none, sexism, and racism). We randomly chose 80% of the data for training, 10% for validation, and the remaining 10% for testing for both datasets. We oversampled the minority class of the Twitter dataset (Twitter+) to produce an unbiased prediction and more fine-grained classification. Using the SMOTE technique [40], we have performed the oversampling. We keep the same oversampling statistics as mentioned in [16], i.e., racism class has been oversampled by 400% of its original size and sexism class by 200%. The detailed classwise distributions of Aggression dataset, Twitter dataset, and Oversampled Twiter dataset (Twitter+) are shown in Table IV.

We have generated the sentiment label of the Twitter dataset using VADER [41] sentiment analyzer. Negative sentiment labels appear in 6525 data points in the Twitter corpus, whereas positive and neutral sentiment labels appear in 4894 and 4671 samples, respectively. To the best of our knowledge, there is no such publicly available sentiment analyzer tool that can work on Hindi–English code-mixed text. Thus, we have compared the results of the aggression dataset with our proposed single-task FS model only.

Table V shows the results of all the state-of-the-art approaches on the Aggression dataset. As is evident from the table, our proposed model outperformed all other state-of-the-art techniques by a significant margin.

Table VI shows the results of all the state-of-the-art approaches on the Twitter+ dataset. This table shows that our proposed SP + DAM + Adv model outperformed the SOTA significantly. We have also noticed that our single-task FS model slightly underperforms the BERT$_{\text{Large}}$ model. Still, there is always an improvement in F1-Score over SOTA when compared with any multitask frameworks. These results established that concurrent execution of multiple related tasks certainly improves the performance of the main task.

TABLE VI

RESULTS OF STATE-OF-THE-ART MODELS AND THE PROPOSED MODEL ON THE TWITTER + DATASET

Model	F1-Score
SOTA	
CNN [16]	0.93
RNN+LSTM [16]	0.85
BiLSTM(attention) [16]	0.93
$BERT_{Large}$ [16]	**0.94**
Our Model (Single Task)	
FS	0.9369
Our Model (MT- CBD+SA)	
FS	0.9412
FS+Adv	0.9456
SP	0.9448
SP+Adv	0.9513
SP+DAM+Adv	**0.9586**

D. Complexity of the Proposed Model and Baselines

Here, we have compared the trainable parameters and training time of different baselines and the proposed model. The corresponding results are shown in Table VIII. From this table, we can examine that our proposed model SP + DAM + Adv takes 928.57 s more time compared to ST-SP. It is obvious that a combination of many methods will take more time compared to any single model. Our results also indicate the same. When comparing two multitask variants (MT-CBD + SA or MT-CBD + ER) versus three multitask variants (MT-CBD + SA + ER), two multitask variant-based models always take less time. However, the execution time difference is not too much (maximum 15 min). When we added more tasks into a multitask framework, the number of trainable parameters has increased. Another observation from Table VIII is that the SP multitask model always takes more time compared to any FS multitask model because the number of trainable parameters of the SP model is almost twice than the FS model.

E. Comparison With SOTA

In this section, we have compared the proposed model with state-of-the-art models in terms of hyperparameters, trainable parameters, training time, accuracy, and F1-Score. There is only one existing work on CBD in code-mixed Indian languages [21] where CBD is treated as a stand-alone task. Table VII shows the results of the existing state-of-the-art approaches on the Hindi–English code-mixed CBD task. Maity and Saha [21] developed BERT + CNN/RNN + Capsule-based approach where they experimented with both LSTM and GRU. As is evident from the table, our proposed model outperformed other state-of-the-art techniques by a significant margin, which illustrates the need to solve auxiliary tasks such as ER and SA in identifying CBD. It is also true that our proposed model takes more training time compared to SOTA. As our model is a multimodal multitask framework,

TABLE VII

COMPARISONS OF STATE-OF-THE-ART MODELS AND THE PROPOSED MODEL FOR CBD IN TERMS OF HYPERPARAMETERS, TRAINABLE PARAMETERS, TRAINING TIME, ACCURACY, AND F1-SCORE

	Parameters settings	**Model**	**#Trainable parameters**	**#Training Time**	**Accuracy**	**F1-Score**

SOTA	Bi-LSTM/GRU: # memory cells =128;
	1-D CNN: 64 filters of size 2;
	Capsule length: Primary capsule =8,
	Bully capsule =16; Routing Iteration = 5;
	Dropout = 0.10; Learning Rate = 0.0001;
	Batch Size = 32; Epoch = 10
	Loss = Marginal loss [44]; Optimizer = Adam [45]
		BERT+CNN +Capsule [21]	155,776	51.15	77.70	76.58

		BERT+LSTM +Capsule [21]	1,002,560	173.25	78.18	78.48
		BERT+GRU +Capsule [21]	772,928	147.93	78.33	77.19
		BERT+CNN +GRU+Capsule [21]	328,576	142.62	79.28	80.30

Our Model	Bi-GRU: # memory cells =128;
	Attention dimension = 100;
	Hidden Activation = ReLU (Dense layers);
	Output Activation = Softmax; Dropout = 0.25;
	Learning Rate = 0.001; Batch Size = 32; Epoch = 10;
	Loss = Categorical CrossEntropy;Optimizer = Adam
		SP+DAM+Adv	52,436,284	1233.08	82.87	**82.86**

TABLE VIII

TRAINABLE PARAMETERS AND TRAINING TIME OF DIFFERENT BASELINES AND THE PROPOSED MODEL

Model	**Uni-modal (Text)**		**Multi-modal (Text+Emoji)**	
	#Trainable Parameters	**Training Time (Sec)**	**#Trainable Parameters**	**Training Time (Sec)**
Single Task				
FS	16,693,210	298.78	18,302,610	305.23
Multitask (CBD+SA)				
FS	16,692,913	315.23	18,301,713	342.57
FS+Adv	16,692,944	361.42	18,301,744	374.32
SP	33,315,126	475.53	34,842,026	495.54
SP+Adv	33,315,157	491.15	34,842,057	510.21
SP+DAM+Adv	33,375,757	510.63	34,963,257	545.26
Multitask (CBD+ER)				
FS	16,693,317	366.18	18,302,117	398.29
FS+Adv	16,693,432	383.06	18,302,232	424.58
SP	33,315,934	504.27	34,843,234	530.33
SP+Adv	33,316,049	518.36	34,843,349	533.87
SP+DAM+Adv	33,376,649	528.41	34,964,549	588.51
Multitask (CBD+SA+ER)				
FS	16,693,620	398.22	18,302,420	425.33
FS+Adv	16,693,756	416.22	18,302,556	446.25
SP	49,968,548	755.28	52,254,348	883.24
SP+Adv	49,968,684	800.82	52,254,484	899.28
SP+DAM+Adv	50,059,584	1024.35	52,436,284	1233.08

it has more training parameters compared to single-task SOTA models. As we know, after training, we save the fine-tuned model for the prediction of unknown test samples. We have examined the prediction time of SOTA BERT + CNN + GRU + Capsule model (1.346×10^{-3} s) and our proposed SP + DAM + Adv (1.921×10^{-3} s) model and found that the difference in execution time is very minimum.

F. Error Analysis

We have manually checked those data instances for error analysis, which were misclassified by the proposed model SP + DAM + Adv. We have considered the following examples for error analysis.

1) Example 1: Tum to AliaBhatt k Tara Intelligent ho 😆😆*; Translation: You are as Intelligent as AliaBhatt:* Originally, this sentence was labeled as bully, but our proposed model has predicted it as nonbully. Though there are no such profane or vulgar words present in Example-1, still it indicates as bullying because it is trying to humiliate somebody based on intelligence. Thus, our proposed classifier is unable to identify the underlying sarcasm of a sentence. The possible reason for this misclassification is the lack of sarcasm data in our dataset.

2) Example 2: Reha ki Chut Jayegi Jail se?; Translation: Will Reha Get Released From Jail?: Example 2 is predicted as bully though it is a nonbully sentence. The reason for this misclassification could be that the model has been confused with the Hinglish word "chut." In this example, "chut" means released, but in general, "chut"(p*ssy) is mostly used as a vulgar word in the Hindi language. As "chut" is frequently present in bullying tweets, the model considers the sentence as bully.

3) Example 3: Kiso ko Paana Phir Khona Phir Uski Yaad me Randi Rona Agar yahi Love h toh Gaand Maraye Aisa Love ka😠*; Translation: If Love means Crying For Someone in Remembrance After Having and Loosing Somebody Then I Must Say, I Hate Love:* The proposed model had predicted Example 3 as bully though its gold label is nonbully. Here, we can notice that Example 3 has two vulgar words, *"Randi"* (whore) and *"gaand maraye"* (Ass Fucking) still, it is not bullying because sentence does not humiliate someone. The possible reason for the wrong prediction could be the presence of multiple abusive or profane words in a sentence.

VI. CONCLUSION AND FUTURE WORK

In this study, we have created BullySentEmo, a benchmark Hindi–English code-mixed corpus annotated with bully, sentiment, and emotion labels to determine whether sentiment and emotion label information can assist in identifying cyberbully more accurately. We have introduced the MM-CBD, an attention-based multitask multimodal framework for sentiment and emotion-aided cyberbullying detection. IA and IRA have been incorporated into our proposed model for the efficient representations of different modalities and helping to learn generalized features over multiple tasks. Our developed MM-CBD framework outperforms all single-task and unimodal models, with a significant margin.

In the future, we would like to develop an unsupervised approach for cyberbullying detection, as data annotation is time-consuming and error-prone work. Moreover, we would like to investigate the effect of detecting sarcasm in the CBD task.

Acknowledgment

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.

References

[1] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, "Cyberbullying: Its nature and impact in secondary school pupils," *J. Child Psychol. Psychiatry*, vol. 49, no. 4, pp. 376–385, Apr. 2008.

[2] C. Myers-Scotton, *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford, U.K.: Oxford Univ. Press, 1997.

[3] M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, *Handbook of Emotions*. New York, NY, USA: Guilford Press, 2010.

[4] D. S. Chauhan, S. R. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? A multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4351–4360.

[5] T. Saha, A. Upadhyaya, S. Saha, and P. Bhattacharyya, "A multitask multimodal ensemble model for sentiment-and emotion-aided tweet act classification," *IEEE Trans. Computat. Social Syst.*, vol. 9, no. 2, pp. 508–517, Apr. 2021.

[6] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," 2015, *arXiv:cs/0506075*.

[7] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, Dec. 1997.

[8] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," 2017, *arXiv:1708.00524*.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[10] H. Rosa *et al.*, "Automatic cyberbullying detection: A systematic review," *Comput. Hum. Behav.*, vol. 93, pp. 333–345, Apr. 2019.

[11] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Int. Conf. Weblog Social Media*, 2011, pp. 11–17.

[12] L. Cheng, R. Guo, Y. Silva, D. Hall, and H. Liu, "Hierarchical attention networks for cyberbullying detection on the Instagram social network," in *Proc. SIAM Int. Conf. Data Mining*, 2019, pp. 235–243.

[13] K. Reynolds, A. Edwards, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, vol. 2, Dec. 2011, pp. 241–244.

[14] R. Pawar and R. R. Raje, "Multilingual cyberbullying detection system," in *Proc. IEEE Int. Conf. Electro Inf. Technol. (EIT)*, May 2019, pp. 040–044.

[15] B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content," in *Proc. 1st Cyber Secur. Netw. Conf. (CSNet)*, Oct. 2017, pp. 1–8.

[16] S. Paul and S. Saha, "Cyberbert: BERT for cyberbullying identification," *Multimedia Syst.*, pp. 1–8, 2020, doi: 1007/s00530-020-00710-4.

[17] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, "Hate speech detection with comment embeddings," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 29–30.

[18] S. Ghosh, A. Ekbal, P. Bhattacharyya, T. Saha, A. Kumar, and S. Srivastava, "SEHC: A benchmark setup to identify online hate speech in English," *IEEE Trans. Computat. Social Syst.*, pp. 1–11, 2022.

[19] A. Bohra, D. Vijay, V. Singh, S. S. Akhtar, and M. Shrivastava, "A dataset of Hindi-English code-mixed social media text for hate speech detection," in *Proc. 2nd Workshop Comput. Model. People's Opinions, Personality, Emotions Social Media*, 2018, pp. 36–41.

[20] S. Ghosh, S. Ghosh, and D. Das, "Sentiment identification in code-mixed social media text," 2017, *arXiv:1707.01184*.

[21] K. Maity and S. Saha, "BERT-capsule model for cyberbullying detection in code-mixed Indian languages," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Cham, Switzerland: Springer, 2021, pp. 147–155.

[22] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes," *Cogn. Comput.*, vol. 14, pp. 110–129, Feb. 2021.

[23] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, "Improving cyberbullying detection with user context," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2013, pp. 693–696.

[24] M. Dadvar, F. D. Jong, R. Ordelman, and D. Trieschnigg, "Improved cyberbullying detection using gender information," in *Proc. 12th Dutch-Belgian Inf. Retr. Workshop (DIR)*. Ghent, Belgium: Univ. Ghent, 2012, pp. 1–3.

[25] C. Van Hee *et al.*, "Detection and fine-grained classification of cyberbullying events," in *Proc. Int. Conf. Recent Adv. Natural Lang. Process. (RANLP)*, 2015, pp. 672–680.

[26] M. Ptaszynski *et al.*, "Sustainable cyberbullying detection with category-maximized relevance of harmful phrases and double-filtered automatic optimization," *Int. J. Child-Comput. Interact.*, vol. 8, pp. 15–30, May 2016.

[27] M. A. Al-Garadi, K. D. Varathan, and S. D. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network," *Comput. Hum. Behav.*, vol. 63, pp. 433–443, Oct. 2016.

[28] H. Hosseinmardi, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Prediction of cyberbullying incidents in a media-based social network," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 186–192.

[29] H. Rosa, J. P. Carvalho, P. Calado, B. Martins, R. Ribeiro, and L. Coheur, "Using fuzzy fingerprints for cyberbullying detection in social networks," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2018, pp. 1–7.

[30] R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, "Aggression-annotated corpus of Hindi-English code-mixed data," 2018, *arXiv:1803.09402*.

[31] P. Ekman, "Basic emotions," in *Handbook cognition emotion*, vol. 98, nos. 45–60. 1999, p. 16, doi: 10.1002/0470013494.ch3.

[32] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, p. 378, 1971.

[33] B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel, "Emoji2vec: Learning emoji representations from their description," 2016, *arXiv:1609.08359*.

[34] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.

[35] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[36] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar, "Beyond binary labels: Political ideology prediction of Twitter users," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2017, pp. 729–740.

[37] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[38] R. Kumar, A. N. Reganti, A. Bhatia, and T. Maheshwari, "Aggression-annotated corpus of Hindi-English code-mixed data," 2018, *arXiv:1803.09402*.

[39] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter," in *Proc. NAACL Student Res. Workshop*, 2016, pp. 88–93.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 28, pp. 321–357, Jun. 2006.

[41] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Weblogs Social Media*, 2014, vol. 8, no. 1, pp. 216–225.

[42] K. Raiyani, T. Gonçalves, P. Quaresma, and V. B. Nogueira, "Fully connected neural network with advance preprocessor to identify aggression over Facebook and Twitter," in *Proc. 1st Workshop Trolling, Aggression Cyberbullying (TRAC)*, 2018, pp. 28–41.

[43] S. Madisetty and M. S. Desarkar, "Aggression detection in social media using deep neural networks," in *Proc. 1st Workshop Trolling, Aggression Cyberbullying (TRAC)*, 2018, pp. 120–127.

[44] L. Xiao, H. Zhang, W. Chen, Y. Wang, and Y. Jin, "MCapsNet: Capsule network for text with multi-task learning," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 4565–4574.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.