A Multitask Multimodal Framework for Sentiment and Emotion-Aided Cyberbullying Detection

Krishanu Maity, Abhishek Kumar, and Sriparna Saha, *Indian Institute of Technology Patna, Patna, 801106, India*

Cyberbullying has become more widespread, especially among teens with the growth of the digital sphere and advancement of technology. This article 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 was no dataset available labeled with bully, sentiment, and emotion. The developed dataset consists of both the 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. A multitask multimodal framework called* MT-MM-Bert+VecMap *based on BERT and VecMap embedding schemes with emoji modality, has been developed. Our proposed multitask-multimodal framework outperforms all the single task and unimodal baselines with the highest accuracy values of 82.05(+/- 1.36)%, 77.87(+/- 1.93)%, and 58.05(+/-2.78)% for the cyberbully detection task, sentiment analysis task, and emotion recognition task, respectively.*

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. According to data given by the National Crime Records Bureau, instances of cyberbullying against women or children in India surged by 36% from 2017 to 2018.[a] 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.[2] India is a country of many languages and most of the times people uses 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.[b] In India, Hindi, English, and Hinglish make up the majority of text interactions on social networking platforms. Hinglish is the depiction of the Hindi language in Roman form.

The emotional state and sentiments of a person have 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, for example, are intrinsically associated with positive sentiments. A well-known observation is that a tweet labeled as bully usually conveys 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 bully tweets.

There are several works in the literature where the tasks of sentiment analysis (SA)[4] and emotion recognition

[a][Online]. Available: https://ncrb.gov.in/en/crime-india-2018-0

[b][Online]. Available: https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers_in_India

1089-7801 © 2022 IEEE
Digital Object Identifier 10.1109/MIC.2022.3158583
Date of publication 11 March 2022; date of current version 19 July 2022.

(ER) are treated as auxiliary tasks to boost the performance of primary task [such as sarcasm detection,[5] tweet act classification (TAC)[6]] in a multitask (MT) framework. Pang and Lee[7] developed a cyberbully detection (CBD) model based on sentiment features. MT learning is proven to be effective when working on related tasks.[8] The transfer 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),[9] contribute in creating trustworthy classification models that aid in detecting the tweeter's emotional state and sentiment, which in turn assists the 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 MT multimodal framework for cyberbullying detection in Hindi–English code-mixed data where SA and ER act as the secondary/auxiliary tasks to increase the performance of primary task, i.e., CBD. We have incorporated two different embedding schemes (MuRIL BERT and VecMap) in our proposed model for the efficient representation of code-mixed data. BERT[10] is a transformer-based[11] language model. Multilingual representations for Indian languages (MuRIL) BERT[c] was pretrained on 17 Indian languages and their transliterated equivalents. VecMap is a bilingual word embedding mapping technology developed by Artetxe *et al.*[12] The basic idea of VecMap is to train the embeddings of source and destination languages individually using monolingual corpora, then align them in a common vector space where comparable words are grouped together using a linear transformation matrix.

The following are the primary contributions of this work:

1) We have developed a benchmark 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.
2) We have addressed the need to consider the sentiment and emotion label information of the tweets while identifying cyberbully.
3) We have proposed a MT multimodal framework called *MT-MM-Bert+VecMap* for sentiment- and emotion-aided cyberbullying detection. Furthermore, two distinct embedding strategies for efficient representations of code-mixed data have been incorporated.
4) We have shown how we improve the performance by solving the CBD, SA, and ER tasks together in a MT framework. Multimodal and MT CBD outperforms unimodal and single-task (ST) CBD by a substantial margin.

[c][Online]. Available: https://tfhub.dev/google/MuRIL/1

RELATED WORK

With the advancement of natural language processing, many researchworks on the identification of cyberbullying have been conducted in the English language rather than other languages.[13] Some of them are listed in the following. Dinakar *et al.*[14] 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 the support vector machine (SVM) classifier and 63% with the Naive Bayes classifier. Pawar and Raje[15] 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.*[16] 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. Bohra *et al.*[17]and Gupta *et al.*[18] developed a multilingual cyberbullying detection system based on machine learning algorithms to detect English, Hindi, Marathi, and Arabic bullying messages.

Also, a few works are present for the code-mixed dataset with a supervised approach. Bohra *et al.*[19] 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. Maity *et al.*[20] developed a Hindi–English code-mixed text corpus from Twiter 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.

Pang and Lee's work[7] attempts were made to detect cyberbullying messages by using probabilistic latent semantic analysis (PLSA)[21] for feature selection. Here some sentiment features extracted using PLSA techniques are utilized for improving the CBD using LibSVM technique. Nahar *et al.*[6] proposed a MT ensemble adversarial learning framework for

TABLE 1. Some samples from annotated dataset with sentiment, emotion, and bully labels.

Tweet	Sentiment Class	Emotion Class	Bully Class
T1: smitaparikh2 Usse bhi sabse hard punishment milni chahiye kisne right diya hai usse kuch bhi likhane ka 😡 **Translation**: Smitaparikh2 should be punished heavily, who has given her the right to write anything.	Negative	Anger	Bully
T2: Pubg ban kue kiya bhosdike. Tera kya jata tha mad*rch*d Sale suor khud to maje leta hai par ham logo ko marta hain. 😞😡😈 **Translation**: Why have you banned Pubg. Whats was your problem mo*herf*cker pigs. You are all enjoying but we are suffering.	Negative	Disgust	Bully
T3: Mtlb mai dono condition mai shocked hi raha, chalo gi good ho gaya.🤔😮🙂 **Translation**: I mean, I was surprised by both the conditions; however, something good happens to me.	Positive	Surprise	Non-bully
T4: sir gi muzaffarpur Bihar se Kolkata Ka Train chalo karo. **Translation**: Sir, kindly operate trains from Muzaffarpur, Bihar to Kolkata.	Neutral	Others	Non-bully

multimodal TAC. They have claimed that TAC performs significantly better than its unimodal and ST TAC variants. Ekman[22] suggested a MT learning architecture that uses external knowledge information to improve 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.*[5] presented a MT 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 bridge this research gap.

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 and emotion. To the best of our knowledge, there is only one publicly available Hindi–English code-mixed corpus for cyberbullying detection. So to achieve our goal (solving the multitasking model), we have further annotated the CDC dataset proposed in Ghosh *et al.*'s work[20] with three sentiment classes and seven emotion classes.

Data Collection

To create a *BullySentEmo* dataset, we added additional 1022 tweets to the old standalone cyberbully dataset, bringing the total number of tweets to 6084. These additional tweets have been scraped from Twitter with the help of the Twitter Search API[d] based on some keywords related to cyberbullying, such as Chuthiya, Rendi, and Kamini. All these keywords are written in Hinglish and are frequently used in India for cyberbullying. The standalone Hindi–English code-mixed cyberbully dataset,[20] which we have considered for further annotation with three sentiment classes and seven emotion classes, contains 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 48.51:51.49. One of the key objective of this work is to develop a corpus that will help further research on sentiment and emotion-aware CBD. As we know, a dataset with more samples is better for training a deep learning model, so we have added 1022 more tweets to the previous dataset. We keep 578 nonbully tweets and 444 bully tweets out of the newly added 1022 tweets to maintain the balanced nature of the new dataset.

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[23] emotion categories (Happiness, Sadness, Fear, Surprise, Anger, Disgust, and others). In Table 1, there are some examples of annotated tweets. For finalizing the annotation label of each tweet, we use a majority vote procedure to resolve conflicts between annotators. Using Cohen's Kappa score the interannotator agreement has been calculated to verify the quality of annotation. We attained the agreement scores of 0.81 for CBD task, indicating that it is of acceptable quality.

Dataset Statistics

Figure 1(a)–(c) depicts dataset statistics based on bully, sentiment, and emotion classes, respectively. Negative sentiment labels appear in 3227 data points in our corpus, whereas positive and neutral sentiment

[d][Online]. Available: https://developer.twitter.com/en/docs/twitter-api

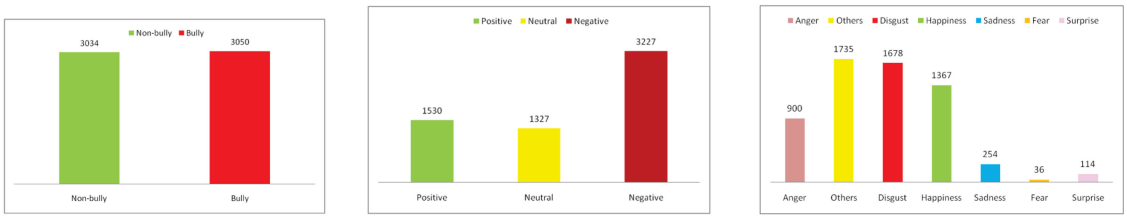

FIGURE 1. Classwise statistics. (a) Cyberbully statistics. (b) Sentiment statistics. (c) Emotion statistics.

labels appear in 1530 and 1327 tweets, respectively. Figure 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. Figure 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 likely to have anger or Disgust emotion.

METHODOLOGY

In this section, we have described the proposed MT multimodal methodology for cyberbullying detection in code-mixed languages. Figure 3 depicts the overall architecture of our proposed *MT-MM-Bert+VecMap* model.

VecMap

FastText[24] Hindi and English word embeddings were used as inputs to VecMap. The CBOW approach was used to train the FastText model, which returns a 300-dimensional dense vector for each token. We transliterated Hinglish to Devanagari–Hindi using the vocabulary given by Khapra *et al.*[25] We initially generated a Hindi–English aligned embedding vector space using the VecMap approach. Then, we utilized it to generate a $\text{vocab\_size}\times 300$-dimensional pretrained embedding matrix. vocab_size is the total number of unique words in the training data.

Bidirectional 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.[26] To capture long-term dependency of input word vectors, bi-GRU encodes the input on both forward and backward direction as follows:

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

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

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

Attention Layer

The attention mechanism's[27] core premise is to assign more weights to the words that contribute the most to

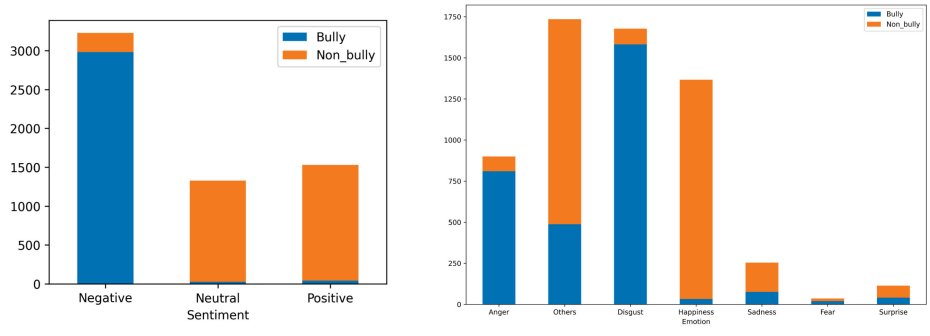

FIGURE 2. Correlation between different classes. (a) Bully versus sentiment. (b) Bully versus emotion.

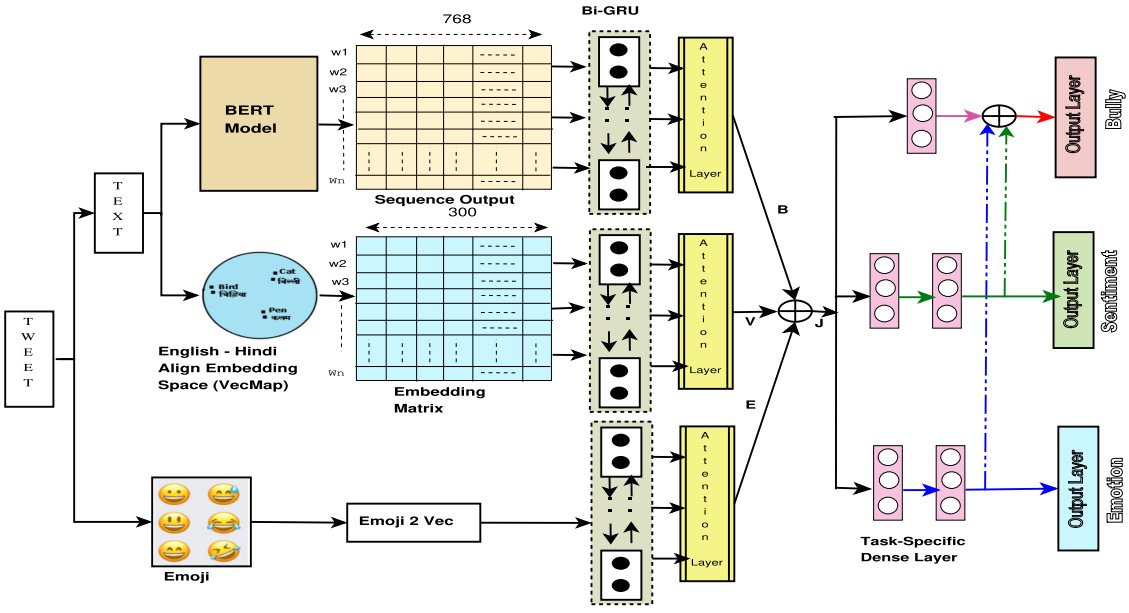

FIGURE 3. MT-MM-Bert+VecMap architecture.

the phrase's meaning. We use the attention mechanism on the bi-GRU layer's output as follows:

$$u_t^i = \tanh(W_w h_t^i + b_w) \tag{3}$$

$$\sigma_t^i = \frac{\exp({u_t^i}^T u_w)}{\sum_t \exp({u_t^i}^T u_w))} \tag{4}$$

$$S_i = \sum_t (\sigma_t^i * h_t^i) \tag{5}$$

where u_w is the context vector and u_t^i is the hidden representation of h_t^i generated by a feed forward neural network. σ_t^i is the attention weight for the ith word, and S_i is the final representation of a sentence after passing through the attention layer.

MT Multimodal Bert-VecMap Framework

Task Description

Let $(X_k, b_k, s_k, m_k)_{k=1}^N$ is a set of *N* tweets where $b_k, s_k,$ and m_k represent the bully, sentiment, and emotion labels corresponding to X_kth tweet, respectively. This MT-MM-Bert+VecMap framework aims to learn a function that maps an unknown tweet u_i to its appropriate sentiment label s_i, emotion label m_i, and bully label b_i.

Textual Features

The BERT language model and VecMap have been employed to extract the textual features from input tweets $X = \{x_1, x_2, \ldots ...x_n\}$. Pooled and sequenced these two types of output have been generated by BERT. A $[\text{batch} \quad \text{size}, 768]$ shaped pooled output represents the entire sentence using a special token called CLS. Let $W_B \in \mathbb{R}^{n \times D_B}$ be the sequence output obtained from the BERT model for input X, where n is the maximum sequence length and $D_B = 768$ is the dimension of each token. We have considered sequence output for the *MT-MM-Bert+VecMap* model and pooled output has been taken into account for machine learning-based baselines. On the other hand, let $W_V \in \mathbb{R}^{n \times D_V}$ be the bilingual embedding matrix obtained from the VecMap for input X, where $D_V = 300$. The BERT and VecMap outputs are processed via bi-GRU (128 hidden units), followed by an attention layer, which learns the context and assigns extra weights to the relevant phrases. Empirically, we have experimented with 64, 128, and 256 different hidden units of bi-GRU and achieved the best results with 128. This is the reason behind the selection of 128 hidden units.

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*[28] to generate $D_e = 300$-dimensional vector representation for each of the emojis retrieved from the tweet. Let us assume that a tweet contain n_e number of emojis, then the final emoji representation of a tweet is obtained as $I \in$

$\mathbb{R}^{n_e \times D_e}$. *emoji2Vec* generated emoji feature vector I is then passed through a bi-GRU (128 hidden units) followed by an attention layer.

To get text and emoji joint modality J, we have concatenated text features, B and V returned by BERT +GRU+attention layer and VecMap+GRU+attention layer with the emoji features E generated by the emoji modality attention layer. Up to this, three tasks have shared the layers, allowing them to share task-specific information. The concatenated feature vector J is passed through three separate task-specific fully connected (FC) layers (bully channel [$\mathrm{FC}_B^1(100 \text{ neurons})$], sentiment channel [$\mathrm{FC}_S^1(100 \text{ neurons}) + \mathrm{FC}_S^2(100 \text{ neurons})$]), and emotion channel [$\mathrm{FC}_M^1(100 \text{ neurons}) + \mathrm{FC}_M^2(100 \text{ neurons})$]) followed by their corresponding output layers. The outputs from FC_S^2 of the sentiment channel and FC_M^2 of the emotion channel are concatenated with the FC_B^1 of the bully channel for finding sentiment+emotion-aided bully features, which helps to enhance the performance of the primary task, i.e., CBD.

Loss Function

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

$$L_{\mathrm{CE}}(\hat{y}, y) = -\frac{1}{N}\sum_{j=1}^{C}\sum_{i=1}^{N} y_i^j \log(\hat{y}_i^j) \tag{6}$$

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 tweets, respectively.

The final loss function L is dependent on the three task-specific individual losses as follows:

$$L = p * (L_{\mathrm{CE}}^S) + q * (L_{\mathrm{CE}}^M) + r * (L_{\mathrm{CE}}^B \mid L_{\mathrm{CE}}^S, L_{\mathrm{CE}}^M). \tag{7}$$

Equation (7) implies that the loss for the CBD task L_{CE}^B is dependent on the sentiment classification loss L_{CE}^S and emotion classification loss L_{CE}^M. Here, *p*, *q*, and *r* are three hyperparameters responsible for giving task-specific loss weights ranges between 0 and 1.

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 10-fold cross-validation and reported the mean metric scores.

Baselines Setup

The following baselines are introduced for comparison with our proposed approach.

1) *ST+BERT:* The word vectors generated by BERT are sent to BiGRU, followed by the attention layer. The output of the attention layer is then passed to task-specific FC layers [FC1(100) + FC2 (100)] followed by the output layer.
2) *ST+VecMap:* Same as previous one, the only change is that this time the input is passed through VecMap rather than BERT.
3) *MT+BERT:* This is the same as our proposed model, except the tweet text is passed through only the BERT model.
4) *MT+VecMap:* This is the same as our proposed model, except the tweet text is passed through only the VecMap.
5) *ST-Bert+VecMap:* Here, the joint feature *J* is followed by only one task-specific layer (two FC layers and one output layer). The number of neurons in bully, sentiment, and emotion output layers are 2, 3, and 7, respectively.

Furthermore, each MT baseline has three 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 we have considered only the text part, and in the case of a multimodal experiment, both text and emoji data have been considered as inputs.

Table 2 shows the results for all of the baselines and the proposed model in terms of *F*1-score, precision, recall, and accuracy. The result table shows the MT variants outperform the ST classifiers significantly for CBD, SA, and ER tasks. Compared to ST-MM-Bert +VecMap, the proposed framework *MT-MM-Bert+VecMap* improves the performance in terms of accuracy by 1.94% for the CBD task, 4.24% for ER, and 1.88% for SA. From the results, it can be concluded that sentiment and emotion information significantly enhance the performance of the main task (CBD). The proposed model attained the highest accuracy of 77.87%, 58.05%, and 82.05% for three tasks, SA, ER, and CBD, respectively. This gain in performance indicates that our approach of combining BERT and VecMap is better than single text modality representation while handling code-mixed data. From the results table, we can conclude that MT (CBD+SA+ER) always performs better compared to the other two MT variants, i.e., MT (CBD+SA) and MT (CBD+ER). MT-BERT+VecMap (CBD

TABLE 2. Experimental results.

Model	Task	Unimodal (Text)		Multimodal (Text + Emoji)	
		Accuracy	*F*1-score	Accuracy	*F*1-score
ST					
ST-BERT	SA	73.51(+/-1.17)	73.67(+/-1.09)	74.54(+/-1.91)	74.39(+/-1.99)
	ER	51.41(+/-2.03)	46.25(+/-2.24)	52.19(+/-2.06)	47.77(+/-2.03)
	CBD	76.22(+/-1.13)	76.08(+/-1.24)	77.07(+/-1.11)	76.89(+/-1.12)
ST-VecMap	SA	75.64(+/-1.50)	75.12(+/-1.65)	75.95(+/-2.97)	75.73(+/-2.47)
	ER	53.40(+/-1.83)	51.06(+/-1.56)	53.12(+/-1.45)	52.07(+/-1.35)
	CBD	78.52(+/-2.01)	78.43(+/-2.06)	79.07(+/-1.76)	79.14(+/-1.76)
ST-BERT+VecMap	SA	75.53(+/-1.76)	75.38(+/-1.72)	75.99(+/-1.60)	76.09(+/-1.31)
	ER	53.65(+/-1.69)	52.02(+/-1.35)	53.81(+/-2.13)	52.73(+/-1. 98)
	CBD	79.97(+/-1.23)	80.13(+/-1.59)	80.11(+/-1.40)	80.04(+/-1.42)
MT (CBD+SA)					
MT-BERT	SA	74.36(+/-2.30)	74.31(+/-1.89)	75.53(+/-2.13)	75.61(+/-2.24)
	CBD	78.43(+/-1.60)	78.19(+/-2.44)	78.78(+/-1.55)	79.05(+/-2.44)
MT-VecMap	SA	75.43(+/-1.54)	75.29(+/-1.68)	75.77(+/-1.54)	75.57(+/-1.61)
	CBD	79.68(+/-1.46)	79.77(+/-1.47)	79.86(+/-1.79)	80.21(+/-2.30)
MT-BERT+VecMap	SA	76.78(+/-1.91)	76.39(+/-1.70)	77.06(+/-1.81)	76.89(+/-1.85)
	CBD	80.77(+/-1.92)	80.98(+/-2.40)	81.23(+/-1.12)	81.36(+/-1.64)
MT (CBD+ER)					
MT-BERT	ER	54.40(+/-1.16)	50.49(+/-1.67)	55.16(+/-1.69)	52.55(+/-1.58)
	CBD	78.31(+/-1.42)	78.39(+/-1.61)	79.16(+/-1.62)	78.95(+/-1.82)
MT-VecMap	ER	53.14(+/-1.37)	47.45(+/-1.68)	54.13(+/-0.81)	49.19(+/-1.18)
	CBD	77.63(+/-2.31)	78.12(+/-3.45)	78.92(+/-1.17)	79.92(+/-1.12)
MT-BERT+VecMap	ER	54.24(+/-1.46)	51.57(+/-1.66)	55.03(+/-1.94)	51.88(+/-2.07)
	CBD	79.88(+/-1.35)	79.35(+/-2.23)	80.28(+/-1.35)	80.62(+/-1.49)
MT (CBD + SA +ER)					
MT-BERT	SA	74.28(+/-2.20)	74.25(+/-2.13)	75.03(+/-1.83)	74.67(+/-1.72)
	ER	54.57(+/-2.65)	51.29(+/-2.48)	55.10(+/-1.79)	52.94(+/-1.62)
	CBD	78.49(+/-2.24)	78.97(+/-2.14)	79.11(+/-1.43)	79.14(+/-1.64)
MT-VecMap	SA	75.68(+/-1.56)	75.48(+/-1.65)	76.22(+/-1.17)	75.81(+/-1.15)
	ER	55.19(+/-2.06)	51.71(+/-1.75)	56.07(+/-1.76)	52.21(+/-2.13)
	CBD	79.99(+/-1.73)	80.06(+/-1.75)	80.26(+/-1.09)	80.39(+/-1.32)
MT-BERT+VecMap	SA	76.81(+/-1.01)	76.41(+/-1.42)	**77.87(+/-1.93)**	**77.75(+/-1.92)**
	ER	56.50(+/-1.75)	53.41(+/-1.43)	**58.05(+/-2.78)**	**55.70(+/-2.93)**
	CBD	81.41(+/-1.52)	81.67(+/-1.73)	**82.05(+/-1.36)**	**82.27(+/-1.22)**

Note: ST: single task, MT: multitask, SA: sentiment analysis, ER: emotion recognition, CBD: cyberbully detection, MM: multimodal. The proposed Multimodal MT-BERT+VecMap outperforms all the baselines in both accuracy and F1 score. The numbers in bold firmly establish the obtained improvements.

TABLE 3. Results of state-of-the-art model and the proposed model for CBD.

Model	Accuracy	*F*1-score
BERT+CNN+Capsule[20]	77.70	76.58
BERT+LSTM+Capsule[20]	78.18	78.48
BERT+GRU+Capsule[20]	78.33	77.19
BERT+CNN+GRU+Capsule[20]	79.28	80.30
MT-MM-BERT+VecMap (our model)	82.05	82.27

TABLE 4. Training time of different baselines and proposed model.

Model	Unimodal (Text)	Multimodal (Text + Emoji)
	Time (sec)	Time (sec)
ST		
ST-BERT	215.46	225.57
ST-VecMap	197.23	210.29
ST-BERT +VecMap	268.81	282.98
MT (CBD+SA)		
MT-BERT	302.12	328.23
MT-VecMap	289.12	310.86
MT-BERT +VecMap	325.23	352.41
MT (CBD+ER)		
MT-BERT	307.15	330.14
MT-VecMap	288.45	311.25
MT-BERT +VecMap	323.45	355.15
MT (CBD + SA +ER)		
MT-BERT	312.12	335.06
MT-VecMap	304.17	326.25
MT-BERT +VecMap	345.20	378.36

+SA+ER)(multimodal) achieves 1.77% and 0.82%, improvements in accuracy values for CBD task (main task) over MT-BERT+VecMap (CBD+ER) and MT-BERT +VecMap (CBD+SA), respectively. On the other hand, MT-BERT+VecMap (CBD+SA) attains the improvements in accuracy and *F*1 values over MT-BERT+VecMap (CBD+ER) as 0.95% and 0.74%, respectively. Multimodal MT-BERT+VecMap (CBD+SA+ER) outperforms unimodal MT-BERT+VecMap (CBD+SA+ER) with accuracy values of 1.06%, 1.55%, and 0.64%, for three tasks, SA, ER, and CBD, respectively. We have also examined that for all the baselines excluding MT-VecMap (CBD+ER) when embedded with VecMap performs better than the one embedded with BERT. The result table illustrates that any multimodal (text +emoji) variant for both BERT and VecMap-based models always performs better than the corresponding unimodal variants. This improvement highlights the significance of including multimodal features for various Twitter analysis tasks.

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 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, Surprise, and so on. As a result, a person's actual state of mind cannot always be detected based on only sentiment.

Comparison With SOTA

There is only one existing work on CBD in code-mixed Indian languages[20] where CBD is treated as a standalone task. Table 3 shows the results of the existing state-of-the-art approach on Hindi–English code-mixed CBD task. As is evident from the table, our proposed model outperformed other state-of-the-art techniques by a significant margin, which in turn illustrates the need of ER and SA in identifying CBD.

We have also computed the time taken by different baselines and the proposed model. From Table 4, we can examine that our proposed model MT-MM-BERT +VecMap takes 95.38 seconds more time compared to ST-BERT+VecMap. It is obvious that a combination of many methods will take more time compared to any single model. Our results also indicate the same. For both cases, i.e., when the comparison between ST versus MT and (BERT or VecMap) versus (BERT+VecMap), ST model and single embedding-based model take less time. But the time gap is not too much (maximum 2 min). We have also examined the prediction-time of ST $(1.733 \times 10^{-3}\text{sec})$ and MT $(1.812 \times 10^{-3}\text{sec})$ bert+VecMap model, and find that the time gap is very minimum.

All the reported results for the proposed model and baselines are statistically significant as we have performed statistical t-test at 5% significance level. Experimental results of this work demonstrate that concurrent execution of three tasks: CBD, SA, and ER

improves the performance of the main task. Furthermore, the multimodal framework outperforms the unimodal framework by a substantial margin since emoji data adds some extra information to inputs, which helps the classifier for better prediction.

Challenges and Novelties of Proposed Model

The first challenge was how to handle code-mixed data? In a social media text, the inclusion of out-of-vocabulary (OOV) words is a severe issue. The noisiness (short words, abbreviations, and misspelled words) in code-mixed data is generally higher than in monolingual data. Such words are not represented in the pretrained word embedding model, resulting in the loss of morphological information. FastText utilizes the character level in represented words into the vectors, unlike word2vec[29] and Glove,[30] which use word-level representations. But the limitation of FastText is that it cannot handle transliterated words, such as billi and khubsurat. So to overcome this issue, we have chosen MuRIL BERT. That's why we have considered both VecMap, which takes FastText Hindi and English word embeddings as inputs to solve OOV problems, and MuRIL BERT to handle transliterated (Hinglish) words. So the finding that a combination of BERT and VecMap representation can handle code-mixed data in a better way is one of the contributions of our proposed model.

The second challenge was designing shared layers and task-specific layers of a multimodal-MT framework in such a way that it can take advantage of multimodal data and valuable information from other secondary tasks, which ultimately helps in boosting the performance of the main task (CBD). Note that there is no standard architecture for MT framework. Depending on the tasks in hand and input feature vectors/modalities, we need to design a MT architecture. To learn the contextual information of input tweets from both directions, we have placed an RNN followed by an attention layer as a shared layer in our proposed model. As we know, RNNs are better for long-range semantic dependency-based tasks than some local key phrases.[31] Now most frequently used RNNs are LSTM and GRU. We have kept GRU instead of LSTM in our proposed model because the performances of both GRU and LSTM were almost identical in our dataset, but LSTM took more times than GRU. The intuition behind keeping the attention layer is to assign more weights to the words that contribute the most to the phrase's meaning, which in turn will help in increasing the accuracy of CBD. We have conducted experiments with some other attention mechanisms, such as dyadic attention mechanism,[32] self-attention,[11] but the simple word attention, as proposed in Yang *et al.*'s work,[27] works better in our problem.

TABLE 5. Experimental results of our model with and without a ft path.

Model	Unimodal (Text)		Multimodal (Text + Emoji)	
	Accuracy	*F*1-score	Accuracy	*F*1-score
MT-BERT +VecMap (without FT)	80.86	80.83	81.79	81.59
MT-BERT +VecMap (with FT)	81.41	81.67	82.05	82.27

The use of a FC dense layer as a task-specific layer is a common practice. We have also incorporated the same in our proposed model. As a new concept, in the proposed architecture, a connection to transfer the intermediate task-specific sentiment and emotion features to the last FC layer of the bully channel is added. Our intuition was to investigate how our main task (CBD) gets benefited from the features extracted for solving two auxiliary tasks, namely SA and ER. Table 5 shows the experimental results of our model with and without a feature-transfer (FT) path. From Table 5, we have concluded that transferring the intermediate task-specific features of secondary tasks to the last FC layer of the main task helps in boosting the performance of the main task. This finding is another novelty of our work.

CONCLUSION AND FUTURE WORK

This article investigated the effect of sentiment and emotion information for detecting cyberbullying from code-mixed tweet dataset. As a contribution, we have created a novel multimodal code-mixed tweet dataset, *BullySentEmo*, annotated with bully, sentiment, and emotion labels to determine if sentiment and emotion label information can assist in identifying cyberbully more accurately. We have introduced the *MT-MM-Bert +VecMap*, an attention-based MT multimodal framework for sentiment- and emotion-aided cyberbullying detection. BERT and VecMap have been incorporated into our proposed model for the efficient representations of code-mixed texts, which are very common for multilingual speakers while writing some informal texts. Our developed framework outperforms all ST and unimodal models, with a significant margin

illustrating the importance of utilizing multimodal information and sentiment, emotion information for CBD. The superior performance of our proposed approach, which achieved accuracy values of 82.05%, 58.05%, and 77.87% for three tasks, CBD, ER, and SA, respectively, confirms that combining BERT and Vec-Map to represent word vectors can handle code-mixed data better than a single way representation.

In the future, we would like to develop an unsupervised approach for cyberbullying detection, as data annotation is time-consuming and error-prone work.

ACKNOWLEDGMENTS

Dr. Sriparna Saha would like to thank the Young Faculty Research Fellowship (YFRF) Award, supported by the 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 thank the support of Ministry of Home Affairs (MHA), India, for conducting 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, 2008.
2. C. Myers-Scotton, *Duelling Languages: Grammatical Structure in Codeswitching*. London, 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. B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," 2005, *arXiv:cs/0506075*.
5. D. S. Chauhan, S. 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.
6. T. Saha, A. Upadhyaya, S. Saha, and P. Bhattacharyya, "A multitask multimodal ensemble model for sentiment- and emotion-aided tweet act classification," *IEEE Trans. Comput. Social Syst.*, early access, Jul. 01, 2021, doi: 10.1109/TCSS.2021.3088714.
7. V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment analysis for effective detection of cyber bullying," in *Proc. Asia-Pacific Web Conf.*, 2012, pp. 767–774.
8. R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, no. 1, pp. 41–75, 1997.
9. B. Felbo, A. Mislove, A. Sùgaard, 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*.
10. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
11. A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
12. M. Artetxe, G. Labaka, and E. Agirre, "Learning bilingual word embeddings with (almost) no bilingual data," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 451–462.
13. H. Rosa *et al.*, "Automatic cyberbullying detection: A systematic review," *Comput. Hum. Behav.*, vol. 93, pp. 333–345, 2019.
14. K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the detection of textual cyberbullying," in *Proc. Int. Conf. Weblog Social Media*, 2011, pp. 11–17.
15. 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.
16. K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, 2011, vol. 2, pp. 241–244.
17. R. Pawar and R. R. Raje, "Multilingual cyberbullying detection system," in *Proc. IEEE Int. Conf. Electro Inf. Technol.*, 2019, pp. 040–044.
18. B. Haidar, M. Chamoun, and A. Serhrouchni, "Multilingual cyberbullying detection system: Detecting cyberbullying in arabic content," in *Proc. 1st Cyber Secur. Netw. Conf.*, 2017, pp. 1–8.
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, Pers., Emotions Social Media*, 2018, pp. 36–41.
20. 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.*, 2021, pp. 147–155.
21. T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Mach. Learn.*, vol. 42, no. 1, pp. 177–196, 2001.

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, no. 1, pp. 110–129, 2021.
23. P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, vol. 98. Hoboken, NJ, USA: Wiley, p. 16, 1999.
24. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," 2018, *arXiv:1802.06893*.
25. M. M. Khapra, A. Ramanathan, A. Kunchukuttan, K. Visweswariah, and P. Bhattacharyya, "When transliteration met crowdsourcing: An empirical study of transliteration via crowdsourcing using efficient, non-redundant and fair quality control," in *Proc. Lang. Resour. Eval. Conf.*, 2014, pp. 196–202.
26. K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," 2014, *arXiv:1409.1259*.
27. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2016, pp. 1480–1489.
28. B. Eisner, T. Rocktäschel, I. Augenstein, M. Bošnjak, and S. Riedel, "Emoji2vec: Learning emoji representations from their description," 2016, *arXiv:1609.08359*.
29. Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learning*, 2014, pp. 1188–1196.
30. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
31. W. Yin, K. Kann, M. Yu, and H. Schütze, "Comparative study of CNN and RNN for natural language processing," 2017, *arXiv:1702.01923*.
32. T. Saha, A. Upadhyaya, S. Saha, and P. Bhattacharyya, "Towards sentiment and emotion aided multi-modal speech act classification in Twitter," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 5727–5737.

KRISHANU MAITY is a research scholar with the Department of Computer Science and Engineering (CSE), Indian Institute of Technology Patna, Patna, India. His research interests includes natural language processing specifically code-mixed languages, deep Learning and user behaviors' analysis in social media. Contact him at krishanu_2021cs19@iitp.ac.in.

ABHISHEK KUMAR is working as a research engineer in the Computer Science and Engineering Department of Indian Institute of Technology Patna, Patna, India. His research interests include the domain of NLP, data science, and machine learning applications. Kumar received M.Tech. degree in computer science and engineering. Contact him at abhishek.km23@gmail.com.

SRIPARNA SAHA is an associate professor in Computer Science and Engineering Department of Indian Institute of Technology Patna, Patna, India. Her current research interests include machine learning, multiobjective optimization, natural language processing, and biomedical information extraction. Contact her at sriparna@iitp.ac.in.