Multimedia Systems (2022) 28:1897–1904
https://doi.org/10.1007/s00530-020-00710-4

SPECIAL ISSUE PAPER

CyberBERT: BERT for cyberbullying identification

BERT for cyberbullying identification

Sayanta Paul[1] · **Sriparna Saha**[1]

Received: 9 July 2020 / Accepted: 23 October 2020 / Published online: 11 November 2020

Abstract
Cyberbullying can be delineated as a purposive and recurrent act, which is aggressive in nature, done via different social media platforms such as Facebook, Twitter, Instagram, and others. A state-of-the-art pre-training language model, BERT (Bidirectional Encoder Representations from Transformers), has achieved remarkable results in many language understanding tasks. In this paper, we present a novel application of BERT for cyberbullying identification. A straightforward classification model using BERT is able to achieve the state-of-the-art results across three real-world corpora: Formspring (∼ 12k posts), Twitter (∼ 16k posts), and Wikipedia (∼ 100k posts). Experimental results demonstrate that our proposed model achieves significant improvements over existing works, in comparison with the slot-gated or attention-based deep neural network models.

Keywords Cyberbullying · Language model · Deep learning · BERT

1 Introduction

Online social media platforms allow people to share and express their thoughts and feelings freely and publicly with others. This can appear as an assortment of tech-empowered exercises, e.g., photo sharing, blogging, social gaming, social video sharing, business networks, comments & reviews, and many others. The information available over these social media is a rich resource for sentiment analysis or inferring other increasing uses and abuses. This increasing growth of social networking introduces continuous harassment and stalking which is commonly referred to as cyberbullying [1]. Broadly cyberbullying can come up of different forms such as racism (e.g., facial features, skin colour), sexism (e.g., male, female), physical appearance (e.g., ugly, fat), intelligence (e.g., ass, stupid), and so on. Sometimes, this act of cyberbullying is anonymous[1], i.e., quite hard to trace, which has intense and devastating effects. Therefore, detecting cyberbullying at its initial stage is a crucial step to prevent this act and also to avoid any fatal incidents caused by it. In recent years, researchers have focused on developing different machine learning and deep learning-based methods for solving the cyberbullying problem.

Classifying texts into specific categories is an ideal problem in Natural Language Processing (NLP). The important intermediate steps involve neural architecture design and data representation using word embeddings. This deep language representation has always been a crucial factor for efficient text categorization. Bidirectional Encoder Representations from Transformers (BERT) [2] was proposed in recent years, and it was successfully used in developing several state-of-the-art models for a wide variety of NLP tasks, including question answering (SQuAD v1.1), natural language inference, and others [2]. It was designed to pre-train deep bidirectional representations from text sequence by jointly conditioning on both left and right context in all layers. The working principle of BERT is twofold. Initially, BERT has been pre-trained on a large amount of unlabeled text, and then, it is fine-tuned for any particular task using labeled data.

Although BERT has achieved amazing results in many natural language understanding tasks, e.g., next sentence prediction, language inference, and so on, here, in this work, we have shown how BERT can be deployed to accomplish cyberbully detection task. Furthermore, BERT is fine-tuned

✉ Sriparna Saha
sriparna.saha@gmail.com

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihta, Bihar, India

[1] https://cyberbullying.org/.

to enhance its performance on cyberbully detection task. The main contributions of this paper are as follows:

- We propose state-of-the-art results for detecting cyberbullying by simply fine-tuning BERT.
- Also, we investigate that BERT can be further represented into an elementary neural model that provides competitive accuracy at a low computational cost with custom layer-wise learning rate and catastrophic forgetting.
- We evaluate our proposed framework and achieve the new state-of-the-art results on three real-world corpora.

The organization of this paper is as follows: a brief survey of previous works solving the cyberbullying detection task has been exhibited in the following section. Section 3 describes the details of the dataset used throughout this work. The proposed frameworks have been explained in Sect. 4. Experimental evaluation along with the obtained results is presented in Sect. 5. The conclusion of the work is elucidated in Sect. 6.

2 Related works

Identifying cyberbullying over social media has been an increasingly trending issue over the past few years. A great number of research activities have been published, trying to address this problem in social networks, and in various forms. In the literature, several works can be found towards providing an effective solution for identifying cyberbullying via text-based, image-based, or video-based, sometimes incorporating multimodality, as well.

Natural Language processing (NLP) and other language technologies have shown their potential performance for solving problems like detection of hate-speech, fake news, harassment, cyberbullying, and abusive language. Many significant architectures and other approaches have been introduced, e.g., Karthik et al. [3] proposed detection of cyberbullying (specifically sexual, racism content) over YouTube dataset using Naive Bayes and SVM classifiers; their system achieved 80.20% and 68.30% accuracy, respectively, on sexuality and racism contents. Reynolds et al. [4] presented a social media optimization technique over Formspring dataset which indicates whether a particular post is of type bully or non-bully. This system achieved 78.5% accuracy. In 2015, Djuric et al. [5] came up with paragraph-to-vector-based distributed representations of comments over Yahoo social media that can easily detect hate-speech achieving a sound accuracy of 80%. In 2017, Badjatiya et al. [6] introduced precise cyberbullying identification, i.e., Racism, Sexism, and others using CNN and LSTM architecture over Twitter data which achieved 93% F1-score. Recently, in 2020, Balakrishnan et al. [7] developed an automatic cyberbullying detection mechanism using Twitter users' psychological features which include personalities, sentiment, and emotion. Raisi et al. [8] expanded the objective task by introducing rapidly changing vocabulary of social interactions. They constructed an objective function based on participant-vocabulary consistency to detect online bullying from Twitter and Ask.fm, an online anonymous Q&A platform. In [9], Squicciarini et al. studied role of user demographics and social network features for identifying and characterizing cyberbullying. In particular, their approach involves modeling peer pressure and social dynamics with analytical models from MySpace and Formspring social media platform.

Over recent years, BERT [2] has become widely used and efficient representation model that achieves state-of-the-art performance on sentence-level and token-level tasks, outperforming many task-specific architectures. Aggarwal et al. [10] showed that the fine-tuned BERT model is robust enough to perform significantly well on the downstream task of classification of news articles to identify fake news. The applicability of BERT has been spread over different domain, addressing various task-specific objectives, such as BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) [11], and Stacked-DeBERT (Stacked Denoising Bidirectional Encoder Representations from Transformers) [12]. Lee et al. deployed BERT (BioBERT) [11] for extracting valuable information from biomedical literature which outperforms previous state-of-the-art models in a variety of biomedical text mining tasks when pre-trained on biomedical corpora. Sergio et al. have used BERT for classifying incomplete data into predefined categories, which significantly improved robustness in skewed data, when compared to existing systems, by designing a encoding scheme in BERT, which is apparently a powerful language representation model entirely based on attention mechanisms, referred as Stacked-DeBERT [12].

It can also be seen in recent times that the architecture of BERT is being widely used to identify toxic comments, hate speech, and several other offensive language categorization tasks. For example, Mozafari et al. [13] introduced a transfer learning-based approach on the existing pre-trained language model BERT by leveraging syntactical and contextual information of all transformers of BERT for detecting hate speech. Pavlopoulos et al. [14] showed a perspective based fine-tuning of BERT to identify offensive language from the conversation over different social media platforms.

3 Data description

The datasets used here are publicly available and consist of manually annotated data which can be used for cyberbullying detection. We have mainly explored three popular social

Table 1 Dataset statistics

Dataset	# Posts				No. of class
Twitter [15]	Racism	Sexism	None	Total	3
	1937	3117	11036	16090	
Wikipedia [16]	Attack	Not attack		Total	2
	13590	102274		115864	
Formspring [4]	Bully	Non-bully		Total	2
	776	11997		12773	

Table 2 Dataset statistics (after oversampling)

Dataset	# Posts				No. of class
Twitter [15]	Racism	Sexism	None	Total	3
	9685	9351	11036	30072	
Wikipedia [16]	Attack	Not attack		Total	2
	81540	102274		183814	
Formspring [4]	Bully	Non-bully		Total	2
	6984	11997		18981	

media platforms namely Twitter, Formspring, and Wikipedia. For the diversity of data used, each platform has been chosen on the basis of classification of cyberbullying content in a different way than the other. Twitter dataset [15] has three classes, i.e., racism, sexism, and none over total 16090 tweets. Wikipedia corpus [16] contains a total of 115854 instances out of which 106402 English entries are extracted with categorization as toxic or non-toxic based on the content. The corpus of Formspring [4] basically contains Q & A type of data. It has 12773 instances classified into 2 classes, out of which only 776 samples are labelled as cyberbully type making the dataset heavily skewed. Overview of each of the used corpora along with class distribution can be seen in Table 1. To get an unbiased prediction and more fine-grained classification, we have oversampled the minority class. That is, we have replicated data belonging from minority class using SMOTE technique, as described by [17]. Precisely, for Twitter corpus, racism class has been oversampled by 400% of its original size, sexism class by 200%; for Wikipedia corpus—attack class by 500%; for Formspring dataset—bully class by 800%. The exact data samples can be viewed in Table 2. We have used the oversampled corpora for all the experimental analysis, reported in the subsequent sections. Some examples of text from each of the corpora are shown in Table 3.

4 Proposed approach

The internal architecture of BERT is multi-layer bidirectional Transformer encoder, which uses bidirectional language models to learn general language representations. The input representation of BERT has the ability to represent both a single sentence or a pair of sentences in one token sequence. Here, a "sequence" is referred to the input token sequence to BERT, which can be a single sentence or multiple sentences packed together. The first token of every sequence is always a special classification token, represented by ([CLS]). The final hidden state that corresponds to the special token is used as the aggregate sequence representation for any classification task [2]. For a given input token, the corresponding representation of input will be assembled by considering corresponding token, segment, and position embeddings together. Here, in this work, we have used the same input representation as introduced in [2]. A high-level overview of the input representation can be seen in Fig. 1.

BERT comprises of hundreds of millions of parameters, while preceding baseline models use much less parameters and also perform much faster. This is why, BERT comes with a huge computational cost. To minimise the cost of the network, we have used a simpler version of BERT, called knowledge distillation [18] method. This method compresses information learned by a large model to a comparably small model. Essentially, any deep neural network

Table 3 Example of experimental data

Corpus	Text/Posts	Class
Twitter	Agha_Memnun @BreitbartNews This is what Christians and other non Muslims have had to deal with for 1400 years in the Middle Eas	Racism
	T @LeVi_Krueger28: If women ruled the world.. There would be no war. At all. Just a bunch of countries that wouldn't talk to each other!	Sexism
	As long as she realizes she's not gonna look as pretty as she usually works. This character is kind of a mess	None
Wikipedia	These damn morons will never understand the seriousness of the present situation	Attack
	I did not ask any further questions and considered the matter closed	Not attack
Formspring	Q: WTF U WAZ just single who yu wiff nw babe? A: Yuu. Actt Likee Itt Takee Day'ss o2 Gett Witt Somebodyy . I Couldaa Gott Witt Somebodyy 1o Mins. Afterr I Saidd I Wuzz Singlee . Lol	Bully
	Q: have u ever been in love? A: no i guess not	Non-bully

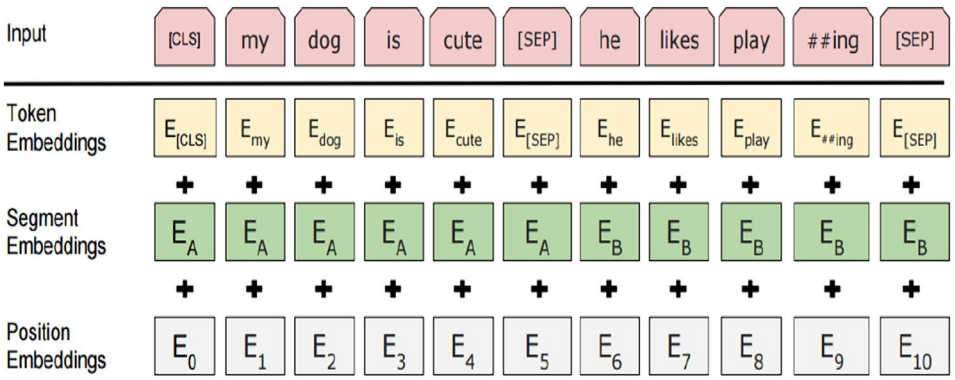


Fig. 1 Input representation of BERT model

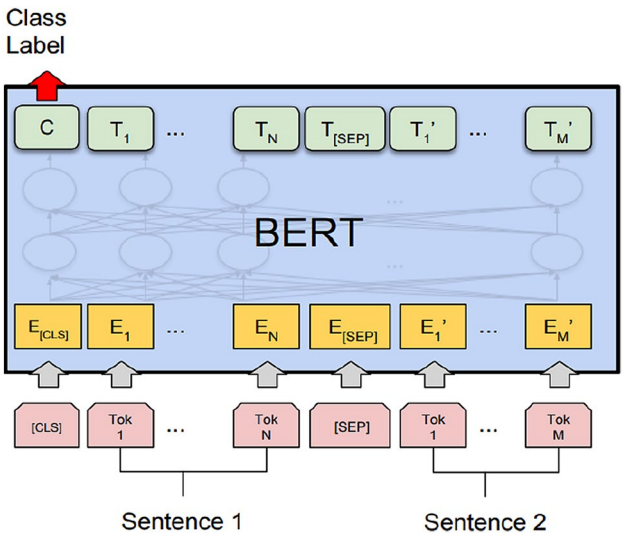


Fig. 2 Fine-tuning procedures for BERT

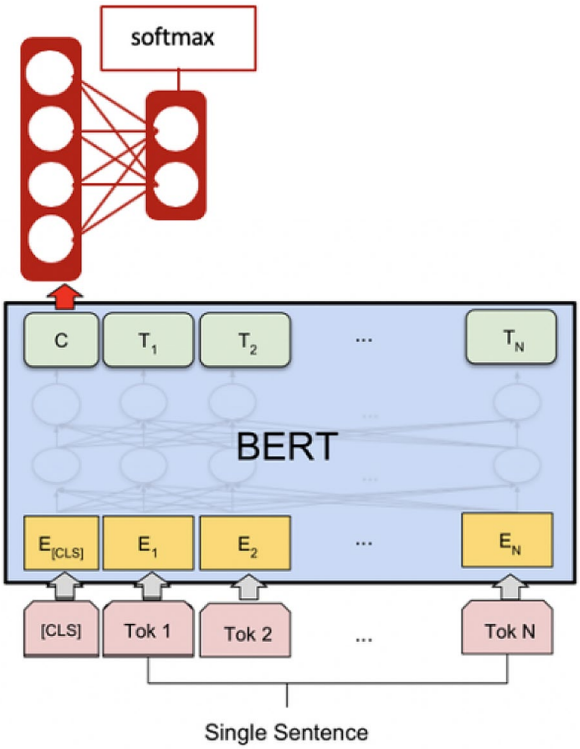


Fig. 3 Overview of our model

computes class-belongingness probability using a softmax output layer as classification layer, which converts the logit z_i, modeling each class into a probability value q_i, and by comparing z_i with the other logits:

$$q_i = \frac{exp(z_i/k)}{\sum_j exp(z_j/k)}, \tag{1}$$

where k is a variable which normally sets to 1. Higher value for k softers probability distribution over classes.

In our case, to make use of $\text{BERT}_{\text{Large}}$ for cyberbullying classification task, a fully connected layer has been added over the final hidden state that corresponds to the [CLS] input token. The model has been further optimized using an additional softmax classifier having parameters, $W \in \mathbb{R}^{K \times H}$, where H refers to dimension of the hidden state vectors and K indicates number of classes, during the fine-tuning phase. Overview of fine-tuning procedure of BERT is shown in Fig. 2.

During distillation, knowledge is transferred to the distilled model by training it on a transfer set and using a soft target distribution for each case in the transfer set that is produced by using the cumbersome model [18]. To distill knowledge from fine-tuned $\text{BERT}_{\text{Large}}$, we carry out knowledge distillation using the training examples to form the transfer set. Figure 3 illustrates an overview of the proposed model.

We have compared our fine-tuned BERT model against CNN, LSTM, BiLSTM with attention layer, and also with two popular traditional machine learning-based text classification models. Machine learning models are trained on tf-idf vectors of the document and those are implemented using

Scikit-Learn 0.22.2. PyTorch 1.4.0 is used as the backend framework. Randomly sampled 80% of the data have been used for training and 10% each has been used for validation and testing, respectively.

5 Experimental evaluation

In this section, we have explained the details of fine-tuning of BERT model and also hyperparameter analysis along with the evaluation of our proposed framework over three aforementioned real-world corpora. Also, we have provided a through-error analysis on misclassified instances.

5.1 Training details

Batch size, number of epochs, and learning rate have been optimized while fine-tuning BERT. Here, we have also considered maximum sentence length for each of the corpora (refer to Table 4). To keep the quality of the model sensitive over all the corpora, we have kept uniform set of optimal hyperparameters. On the basis of potential performance on validation set, we have used learning rate of 2×10^{-5}, dropout probability of 0.5, batch size of 16, and respective maximum sentence length for individual corpus. To build an effective transfer set for distillation, we have augmented the training splits of each of the corpora by applying POS-tagged word interchange and random masking [18]. The optimized set of hyperparameters for fine-tuning BERT are shown in Table 5.

Table 4 Maximum sentence and word length for individual corpus

Corpus	MaxSent_Len	Max-Word_Len
Twitter	30	17
Wikipedia	201	10
Formspring	105	10

Table 5 Optimized sets of Hyperparameters

Name of hyperparameter	Value
No. of Transformer block	12
Hidden state size	768
Batch size	16
Learning rate	2×10^{-5}
No. of Epoch	10
Dropout probability	0.5

5.2 Experimental results

We have shown the mean F1 scores across five runs in Table 6. All the results reported in the table except our proposed framework, BERT, are reproduced from the original works [19]. We have also calculated the inference times on the validation sets, with batch size of 16 for all the three corpora. $\text{BERT}_{\text{Large}}$ consumes computational power at least around 30 times more with respect to other models. Table 7 shows the inference time for all the frameworks with respect to each of the corpora. Figure 4a shows the comparison between the prediction quality on the validation set and on the test set. The graph conveys the effectiveness of our proposed model. To demonstrate that our model has been sufficiently trained, we have shown validation accuracy over a number of epochs in Fig. 4b.

5.2.1 Comparative analysis with baselines

To demonstrate that our proposed method is potentially performing, we have shown comparative analysis with baseline models which include both traditional as well as deep learning-based models, as presented in Table 6. We have reproduced the results from the original paper [19].

- Support Vector Machine and Logistic Regression using two data representation methods, namely, character n-gram and word n-gram features.
- CNN which efficiently extracts contextual features from the given textual content with minimal complexity.

Table 6 Results on the validation and test sets. Best values are in bold

Model	Twitter		Wikipedia		FormSpring	
	Val. *F*1	Test *F*1	Val. *F*1	Test *F*1	Val. *F*1	Test *F*1
SVM	0.83	0.80	0.82	0.77	0.79	0.75
LR	0.85	0.81	0.81	0.75	0.81	0.72
CNN	0.96	0.93	0.89	0.81	0.93	0.91
RNN+LSTM	0.87	0.85	0.73	0.61	0.93	0.88
BiLSTM(attention)	0.94	0.93	0.92	0.87	0.94	0.91
$\text{BERT}_{\text{Large}}$	0.96	0.94	0.94	0.91	0.94	0.92

Table 7 Comparison of inference (seconds) on validation sets with batch size 16

Corpus	CNN	RNN + LSTM	BiLSTM (att.)	$BERT_{Large}$
Twitter	0.3T	0.4T	0.7T	25.4T
Wikipedia	3.4T	5.9T	7.9T	211.3T
Formspring	0.01T	0.3T	0.7T	14.7T

- RNN with LSTM as Long Short-Term Memory architecture is an improvement over RNN, capable of learning long-term dependencies.
- BiLSTM with attention layer further increases the amount of input information fed to the network by encoding it in both forward and backward direction. Using bidirectional LSTM, input from both the past and future of the current time step can be utilised. The attention mechanism allows the model learn where to focus upon the input sentence and what it has already produced so far.

5.3 Key observations

All the reported results above are statistically significant as we have performed statistical *t* test [20] at 5% significance level. Thus, to ensure that no ambiguity was introduced during training, the experiments were conducted for five times. In the data description section, we have seen that, for each of the corpora, the instances of bully class are very less in comparison to total number of non-bully instances.

5.3.1 Statistical significance test

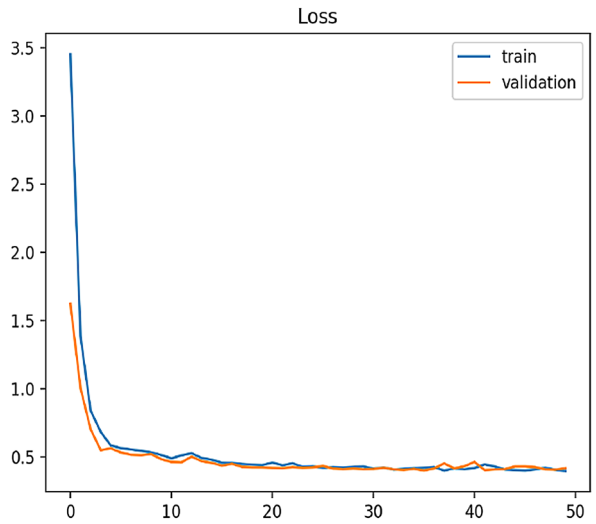
As mentioned in the earlier subsection, results of statistical *t* test of five different runs of our proposed model on the dataset are shown in Table 8.

The *p* values [21] after conducting *t* test on the results of our proposed framework with respect to other models are shown in Table 9. The threshold value of *p* is 0.05. Results show that for majority of the cases the *p* value is less than threshold signifying that the performance improvements attained by our proposed model are statistically significant.

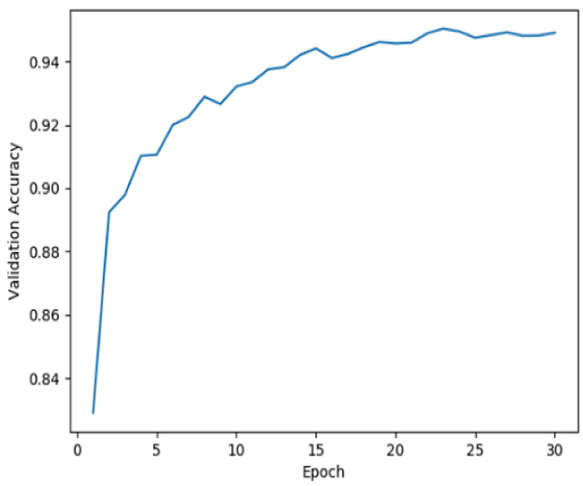
5.3.2 Error analysis

We have manually checked the misclassified instances to perform a thorough error analysis. The proposed models are unable to classify certain posts or comments which contain comments that are likely to appear in other categories, e.g., "This bitch got mad". In the quoted example, the word 'bitch' is an abusive word and is frequently used for different bullying comments, which leads it to be categorized as bully but originally the sentence belongs to non-bully. Few possible reasons behind these misclassifications can be:

- Presence of abusive or swear words in the posts or comments.
- Also, it can be observed that each of the corpora contains non-standard English words which our proposed model is not able to interpret properly to build a vector representation.



(a) Model loss on both train and validation



(b) Model accuracy on validation set

Fig. 4 Learning of our proposed model indicates a good fit

Table 8 F1-score for five runs of each of the models on three corpora

Corpus	Model performance					
	SVM	LR	CNN	RNN + LSTM	BiLSTM (att.)	BERT⋆
Twitter	0.80	0.81	0.93	0.85	0.93	0.94
	0.80	0.81	0.93	0.84	0.93	0.94
	0.80	0.80	0.93	0.85	0.93	0.93
	0.79	0.80	0.92	0.83	0.93	0.93
	0.80	0.81	0.91	0.85	0.92	0.94
Wikipedia	0.77	0.75	0.81	0.61	0.87	0.91
	0.77	0.76	0.81	0.61	0.86	0.90
	0.76	0.74	0.81	0.61	0.86	0.90
	0.78	0.75	0.80	0.61	0.87	0.90
	0.77	0.75	0.81	0.60	0.87	0.91
FormSpring	0.75	0.72	0.91	0.88	0.91	0.92
	0.75	0.71	0.91	0.87	0.91	0.92
	0.74	0.71	0.91	0.88	0.91	0.92
	0.75	0.71	0.90	0.87	0.91	0.93
	0.75	0.72	0.90	0.88	0.90	0.92

* indicates proposed framework

Table 9 p values obtained after comparing our proposed framework with other state-of-the-art models for different datasets

Corpus	SVM	LR	CNN	RNN + LSTM	BiLSTM (att.)
Twitter	8.39e-11	2.79e-10	0.034	4.75e-08	0.035
Wikipedia	6.89e-10	2.28e-10	1.51e-09	1.89e-13	5.68e-13
Formspring	5.42e-12	5.42e-12	0.0009	4.89e-07	0.001

Table 10 Misclassified instances and reasons behind misclassification

Instances	Predicted	Original	Possible reason
@Discerningmumin hamas was elected once and have not had elections since.they have skipped many elections. mursi was going in same direction	Racism	None	As the word 'Mursi' indicate some ethnic fact
Hahah funny how u defend that beiber kid its also funny how u stalked my whole Twitter! nice goin! bitch!	Bully	Non-bully	Due to presence of the words, e.g., 'stalk' and 'bitch'
Hey why you such a bitch ? why thank yuh !	Bully	Non-bully	Same as before

Table 10 contains some instances of misclassification and the reasons for the same.

6 Conclusions and future works

This work forges ahead the state-of-the-art in cyberbullying identification in more fine-grained way over various social media platforms. This paper contributes the following towards effectively identifying cyberbullying to avoid its adverse effects: (a) developing state-of-the-art results for cyberbullying classification by simply fine-tuning BERT, (b) modelling BERT can be further represented into an elementary neural model that provides competitive accuracy at a low computational cost, and (c) empirically evaluating the performance of the frameworks over three different real-world corpora.

In this paper, we have conducted extensive experiments to investigate the fine-tuning approach of BERT for the cyberbullying detection task. The experimental results show that the performance of proposed BERT model is reasonably good. As the further extension of the present work, we are planning to expand the area of social networking sites by providing with our framework as a text-based automatic cyberbullying detection tool along with exploring the usage

of combining extrinsic knowledge with BERT model. The future scope of this work also includes incorporating different modality of information, e.g., images, videos, and audio data.

Acknowledgements 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. Peter, K.S., et al.: Cyberbullying: Its nature and impact in secondary school pupils. J. Child Psychol. Psychiatry **49**(4), 376–385 (2008)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: arXiv preprint arXiv:1810.04805 (2018)
3. Dinakar, K., Reichart, R., Lieberman, H.: Modeling the detection of textual cyberbullying. In: Fifth International AAAI Conference on Weblogs and Social Media (2011)
4. Reynolds, K., Kontostathis, A., Edwards, L.: Using machine learning to detect cyberbullying. Int. Conf. Mach. Learn. Appl. Workshop **2**, 241–244 (2011)
5. Djuric, N., et al.: Hate speech detection with comment embeddings. In: Proceedings of the 24th International Conference on World Wide Web. pp. 29–30 (2015)
6. Badjatiya, P., et al.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759–760 (2017)
7. Balakrishnan, V., Khan, S., Arabnia, H.R.: Improving cyberbullying detection using Twitter users' psychological features and machine learning. Comput. Sec. **90**, 101710 (2020)
8. Raisi, E., Huang, B.: Cyberbullying identification using participant-vocabulary consistency. In: arXiv preprint arXiv:1606.08084 (2016)
9. Squicciarini, A., et al.: Identification and characterization of cyberbullying dynamics in an online social network. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 280–285 (2015)
10. Aggarwal, A., et al.: Classification of fake news by fine-tuning deep bidirectional transformers based language model. EAI Endorsed Transactions on Scalable Information Systems Online First. EAI, Ghent (2020)
11. Lee, J., et al.: BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
12. Sergio, G.C., Lee, M.: Stacked DeBERT: All Attention in Incomplete Data for Text Classification. In: arXiv preprint arXiv:2001.00137 (2020)
13. Mozafari, M., Farahbakhsh, R., Crespi, N.: A BERT-based transfer learning approach for hate speech detection in online social media. International Conference on Complex Networks and Their Applications, pp. 928–940. Springer, Berlin (2019)
14. Pavlopoulos, J., et al.: Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 571–576 (2019)
15. Waseem, Z., Hovy, D.: Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop, pp. 88–93 (2016)
16. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1391–1399 (2017)
17. Chawla, N.V., et al.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
18. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: arXiv preprint arXiv:1503.02531 (2015)
19. Agrawal, S., Awekar, A.: Deep learning for detecting cyberbullying across multiple social media platforms. European Conference on Information Retrieval, pp. 141–153. Springer, Berlin (2018)
20. Dietteric, T.G.: Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. **10**(7), 1895–1923 (1998)
21. Nuzzo, R.: Scientific method: Statistical errors. Nat. News **5067487**(487), 150 (2014)