Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)
IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4

Cyberbullying Detection using Pre-Trained BERT Model

Jaideep Yadav
Computer Engineering
Delhi Technological University
Delhi, India
jdp02041997@gmail.com

Devesh Kumar
Computer Engineering
Delhi Technological University
Delhi, India
deveshkumarkblock@gmail.com

Dheeraj Chauhan
Computer Engineering
Delhi Technological University
Delhi, India
chauhan.dheeraj11@gmail.com

***Abstract*—Cyberbullying is spread across various social media platforms. It is a wrong deed in which the victim is harassed by receiving the derogatory / provocative / sensitive images or text messages by the bully. Detection of such message/post in such large platforms is very difficult and may sometimes lead to false detection. Recently, deep neural network based models have shown significant improvement over traditional models in detecting cyberbullying. Also, new and more complex deep learning architectures are being developed which are proving to be useful in various NLP tasks. Google researchers has recently developed a language learning model called BERT, which is capable of generating contextual embeddings and is also able to produce task specific embeddings for classification. A new approach is proposed to cyberbullying detection in social media platforms by using the novel pre-trained BERT model with a single linear neural network layer on top as a classifier, which improves over the existing results. The model is trained and evaluated on two social media datasets of which one dataset is small size and the second dataset is relatively larger size.**

Keywords—cyberbullying; detection; deep neural network models; embedding; pre-trained BERT; social media

I. INTRODUCTION

The internet is growing rapidly in the world. It has given rise to new opportunities in every field whether it is business, education, entertainment, or sports. Today the internet is everywhere, isn't it? But there are two sides to a coin. The internet also has its own disadvantages. One of the major disadvantages is cyberbullying on different online platforms. Cyberbullying is a very big problem in today's society. Recent studies show that 36.5% of people feel that they have been cyberbullied in their lifetime, and 17.4% [2] of people accepted that this has happened to them in the last 30 days. Along with the internet, cyberbullying activities also increase rapidly. These numbers are more than double what they were in 2007. 87% of young online users have accepted that they are witnesses of some kind of cyberbullying occurring online. So, this is the time to stop cyberbullying [2].

Dealing with cyberbullying on online platforms is a very challenging task because it happens in different ways such as by using photographs, using toxic comments, using videos, etc. Here it is continued only with textual bullying activities. To stop this, technology is built to detect cyberbullying activities automatically and also removes them. Detection of these bullying comments also a very tedious task because of the classification of these comments is quite subjective. In the casual conversation of friends, some comments might look like a bully but actually they are not.

To detect these bullying comments various deep learning models are used by the researchers. Also, different word embedding techniques are used to achieve more accuracy. Previous studies on Cyberbullying detection proposed traditional machine learning models and deep learning models with different word embedding techniques (such as GloVe [15], SSWE [16]) where deep learning based models outperformed the traditional machine learning models. A new method of cyberbullying detection is proposed and improved over the results of the previous models by using a pre-trained BERT [11] model, which is a novel technique created by the Google AI research labs and proved to outperform on several NLP tasks.

II. RELATED PAST WORKS

In the past work, authors and researchers have experimented with various machine learning models (such as logistic Regression, Naive Bias, Random Forest, Support Vector Machine) for the detection of cyberbullying. For this, they have taken datasets from a variety of social media platforms (like FormSpring, Twitter, Wikipedia etc.). They have also proposed deep learning models and claimed that deep learning based models outperformed machine learning based models for this classification problem.

Rui Zhao and Kezhi Mao [3] used a new representation learning technique. In this method, they used Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) developed via semantic extension of the deep learning model stacked denoising autoencoder. P. Zhou, et. al. [4] proposed attention-based B-LSTM technique, S. Bhoir, et. al. [5] presented a research of various word embedding techniques based on various parameters. Banerjee [6] used CNN with GLoVe embeddings to achieve a higher accuracy.

Sweta Agrawal and Amit Awekar [7] presented a systematic work in which they have experimented with four different DNN models: Convolutional Neural Network (CNN), Long Short Term Memory Neural Network (LSTM), Bi-directional LSTM and BLSTM with attention. In these Deep learning models, they have also used some different word embedding techniques-random, GloVe, and SSWE.

III. DATASETS

In this detection work, two publicly available datasets: Formspring (a Q&A forum) [13] and Wikipedia [14] talk pages (collaborative knowledge repository) are used. Both of these datasets are manually labelled and free to use datasets.

There some problems with these datasets. One problem is that these datasets are imbalance according to the target classes. The number of bullying comments is much less as compared to the non-bullying comments. This will lead to a bias prediction for non-bullying comments and leads to good but fake accuracy. To avoid this problem, the bully data is oversampled to various rates in the dataset and tested with them. Another problem is that the comments in the datasets are of various sizes. This can affect the vocabulary of each dataset. So, it truncated the size of the comment ranked at 95 percentiles in that dataset as mentioned by [7] in their paper. For the smaller comments, padding is used to match the length.

Formspring: This is a question-answer based online platform where an anonymous person can also comment. This dataset contains 12773 question-answer pair comments which are manually annotated by three workers. This dataset contains 776 posts which are marked bully by at least two workers.

Wikipedia: Wikipedia maintains a separate talk page, for each normal Wikipedia page, which contains the history of discussion of all the users who have participated in its editing. This dataset contains 115864 discussion comments which are manually annotated by ten persons. This dataset contains 13590 comments which are labelled as a personal attack (bully).

IV. PROPOSED METHOD

For our cyberbullying detection task, a deep neural network called Transformer is employed as the base in our model. Transformer [12] is a novel neural network architecture based on a self-attention mechanism that is particularly well suited for language understanding. A recently developed novel BERT [11] pre-trained model is used with fine-tuning for our specific task and dataset. BERT is built on top of the Transformer and consists of 12 layers of it in the BERT-base-model (uses transformer encoders only) provided by the authors. Initially, pre-trained BERT is trained on the task-specific dataset to learn the dataset-specific embeddings. Then, it generates contextualized embeddings for the text provided to it as input. This model is a prebuilt model that is free to use and has already defined standard internal training parameters recommended by the authors. A single linear neural network layer is used on top of BERT for the classification purpose which is untrained initially. This layer classifies the sentences based on the number of classes specified (2 in our case).

A. BERT Embeddings

The main activity of a BERT model is to generate word and sentence embeddings (inbuilt pooling) for input to classifiers. BERT has proved to give state-of-the-art results for many NLP related tasks and is used in Google search engines since 2018. As defined in [11], BERT is a technique of pre-training language representations, meaning that it is trained on a general-purpose "language understanding" model on a large text corpus (like Wikipedia), and then used for various downstream NLP tasks. Pre-trained embeddings can either be contextual or context-free, and contextual embeddings can further be categorized as unidirectional or bidirectional. Context-free models such as word2vec, GloVe or SSWE generate a single word embedding representation for each word in the vocabulary. But BERT generates a bidirectional contextual embedding that can lead to different embedding for the same word according to its meaning in the textual context.

The BERT-base-model consists of 12 layers of transformer for generating the final embeddings. Each layer consists of only transformer encoders to encode (generate embeddings) the input data. Initially, the tokenized sentences are passed to the first layer of transformer, which then generates the same number of tokens as output. This is then passed through several layers which also produces the same number of tokens as output. But of course, after passing through each layer it has a changed feature values (weights) as it progresses. For training BERT on a particular dataset, the input is given as individual text sentences. These sentences are converted to BERT specific tokens by using BERT tokenizer. Further formatting is required so that the data becomes ready for training.

It is required to:

1. Add 2 special tokens ([CLS] and [SEP]) to the beginning and end of each sentence.
2. Pad the smaller size sentence with a [PAD] token and truncate the large size sentence to a fixed constant length.
3. Explicitly differentiate sentence tokens from [PAD] tokens with the “attention mask”.

After training the BERT model for a specific number of epochs it generates final embeddings. The [CLS] token embedding from the final layer of the transformer is used for the classification. Also, the word specific embeddings are generated which can be used for word classification tasks. BERT model flow chart is shown in the fig-1.

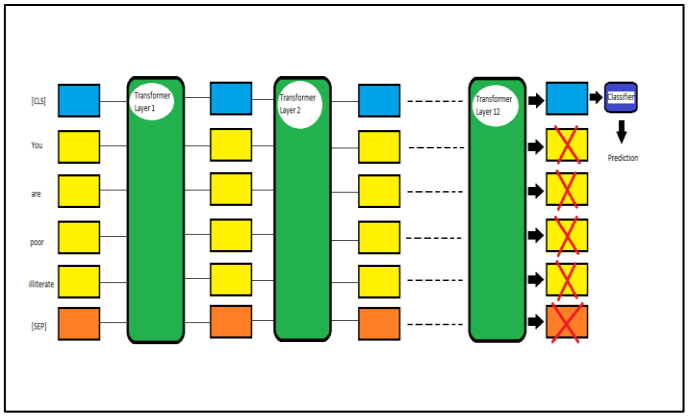

Fig.1 BERT Base model overview

B. Architecture

The model consists of various components/layers constituting its architecture. Following fig-2 shows the overview of the architecture used:

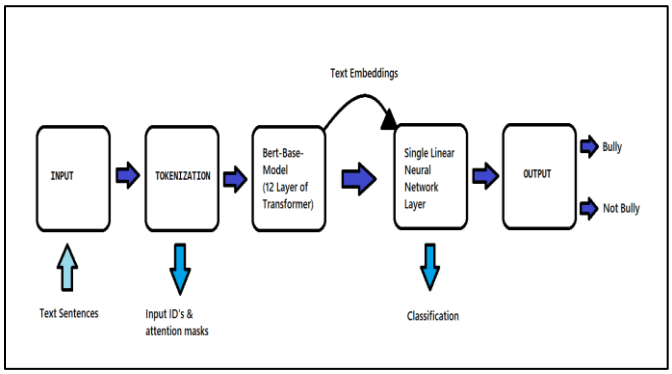

Fig.2 Architeture Overview

The BERT model and the single linear classification layer on top of it are the core layers of the architecture. In our bully detection model, different components/layers are employed, they are as follows:

- Input - The input layer provides the required text sentences in our dataset with the corresponding labels (0-not bully, 1-bully) to the BERT tokenization layer.
- Tokenization - in this layer, tokenizer provided by the BERT model is used. It is used to convert the words in the text to integer tokens determined by the pre-trained vocabulary of BERT.
    - At first, sentences are split into separate tokens for each sentence in the dataset and then padded or truncated to a constant maximum length.
    - Then, the special tokens [CLS] (stands for classification) and [SEP] (used to separate each sentence) are added to the beginning and the end of each sentence respectively.
    - After generating the tokens, they are converted to token_ids, which are integers corresponding to each token. These integer values are already defined values in the pre-trained BERT vocabulary. For example, the token_id value for [PAD] token is defined to be 0 in the vocabulary.
    - At last, attention mask array is created for each sentence. Attention Mask is simply an array of 1s and 0s which helps in differentiating between real tokens and padding tokens. The value 1 indicates a real token, while the value 0 indicates a padding token. Attention mask is useful when the model is trained, so that the model does not include padding tokens in its learning process.
- BERT-Base-Model - the core component of the architecture is the BERT embedding model. We have used the "BERT-base-uncased" model which provides a pre-trained model for lowercased English language and consists of 12 layers of transformer encoders to encode the language data. The model is fine-tuned on the dataset to learn dataset-specific vocabulary and generate the corresponding embeddings. It is provided with token_ids and corresponding attention mask as input for each sentence in the dataset. The special token [CLS] which is also the first token in the input to BERT contains the sentence embedding as a 768 sized vector which is then used to classify the sentence. The max-pooling is already done by the BERT model and that's because it obtained a 768 sized vector as output. This token embedding contains all the essential features of a sentence required for the task. The fig-3 below represents the training and output of BERT model:

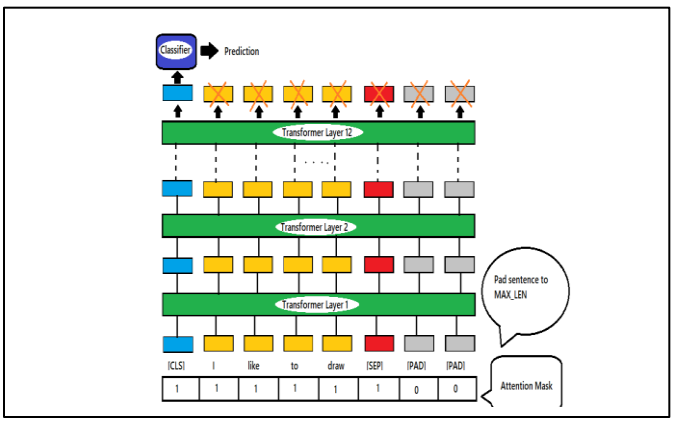

Fig.3 BERT Base model with token_ids and attention masks as input

- Classifier Layer - this layer consists of a single linear layer of neural network which is trained on the embeddings provided as input from the output of the

previous layer. It, then classifies the input based on the classes specified.

- Output - output generated by the classification layer is 0 or 1. Where 0 represents a non-bully sentence and 1 represents a bully sentence.

C. Implementation Details

The proposed approach is implemented using Python programming language and Google Colab runtime environment which is useful for running neural network models on a GPU machine. The transformers package from Hugging Face is used to provides us a PyTorch interface for working with BERT. Currently the Hugging Face library is the most widely accepted and powerful library (with PyTorch interface) for working with BERT models [11].

Two datasets were used to test the model and gather the results for analysis. The first was Formspring dataset with 12773 posts with around 776 marked as bully posts. The second dataset used was Wikipedia dataset with 115864 posts with 13590 marked as bully posts. Also, the Formspring dataset has a huge class imbalance, so, the model is tested on the oversampled dataset. The datasets are pre-processed to clean the not required fields in the data. Only the text and its label were extracted for the classification purpose.

V. RESULTS

The model is trained using various hyperparameters and using the standard values suggested in the BERT [11] paper. The validation loss metric is also used while training the datasets to avoid overfitting of the trained model.

The table-1 below represents the results obtained for the Formspring dataset for various oversampling rates as mentioned in [7]:

TABLE I. RESULTS FOR FORMSPRING DATASET WITH VARYING OVERSAMPLING RATES

Oversampling Rate	Label	Precision	Recall	F1-Score
1 time	Bully	0.64	0.55	0.59
2 times	Bully	0.82	0.91	0.86
3 times	Bully	0.90	0.99	0.94

The performance of the model is evaluated on the formspring dataset with varying oversampling rates. The posts marked as bully are oversampled to the varying rates to increase the number of bully posts. The results show that the performance increases with the increasing number of bully posts as it is able to learn more about the bully sentences.

The table-2 below represents the results obtained for the Wikipedia dataset:

TABLE II. RESULTS FOR WIKIPEDIA DATASET

Label	Precision	Recall	F1-Score
Bully	0.81	0.82	0.81

No oversampling was done for the Wikipedia dataset (even though it has very few posts marked as bully), as it is a large dataset and the BERT model is proved to work well when the dataset is large enough.

The table-3 below compares the results of the study with the DNN models used in [7]:

TABLE III. PERFORMANCE COMPARISON USING F1-SCORE

Dataset	Label	Sweta & Amit's [7] Models	Our Model
Formspring (x1)	Bully	0.51	0.59
Formspring (x3)	Bully	0.91	0.94
Wikipedia	Attack (or Bully)	0.76	0.81

The comparison shows that our proposed model gives an improvement over the past models. The proposed model is mainly compared with the models used by Sweta and Amit [7] as it provides systematic and latest results of the deep neural network based models and also the 2 datasets used are the same.

The results from some other existing models are presented in the table-4 below:

TABLE IV. PREVIOUS MODELS RESULTS

Paper	Dataset	Model Used	Performance
[13]	Formspring	SMO, IBK, J48, JRip	Accuracy – 78.5%
[14]	Wikipedia	LR, MLP	AUC – 96.59
[9]	Twitter	RNN, Glove	Accuracy – 81.60%
[6]	Twitter	CNN, Glove	Accuracy – 93.97%
Our model	Formspring	BERT-base-model	Accuracy – 98%
Our model	Wikipedia	BERT-base-model	Accuracy – 96%

VI. CONCLUSIONS AND FUTURE WORKS

Since cyberbullying is already a well-known and well-determined type of bullying in Social media (like Formspring, Twitter, Wikipedia, etc.), many studies and experiments have been conducted by the researchers to detect cyberbullying in such platforms. The old methods of traditional machine learning were also used by the researchers but they proved to be inefficient and inaccurate by time. Then,

recently deep learning based models proved to outperform the previous traditional models. In our proposed approach of using a pre-trained BERT model which is based on the complex and novel deep neural network, Transformer provides a new approach of detecting bullying in different social media platforms. Also, it gives improved results in comparison to the previous models.

The proposed model uses a single linear layer of neural network for classification which can be replaced by the deep neural network models like CNN and RNNs. Also, the model gives much better and stable results if the size of the dataset is large like the Wikipedia dataset gave much better results without the need for the oversampling. The above observations learned from the proposed model leads to the future scope of the research work.

REFERENCES

[1] Jason Brownlee, "How to use Word Embedding Layers for Deep Learning with Keras" on October 4, 2017 in Deep Learning for Natural Language Processing.

[2] Justin W. Patchin, "Summary of Our Cyberbullying Research (2007-2019)" , Cyberbullying Research Centre, July 10, 2019.

[3] Rui Zhao, Kezhi Mao, "CyberBullying Detection based on Semantic-Enhance Marginalize Denoising Autoencoders" IEEE Transaction on Affective Computing, 2015.

[4] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Houng Wei, Haobo Xu "Attention-based Bi-directional Long Short Term Memory Network for Relation Classification" proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 207-212, August 12, 2016.

[5] MS. Snehal Bhoir, Tushar Ghorpade, Vanita Mane "Comparative Analysis of Different Word Embedding Models" IEEE, 2017.

[6] V. Banerjee, J. Telavane, P. Gaikwad and P. Vartak, "Detection of Cyberbullying Using Deep Neural Network", 2019 5th International Conference on Adavnced Computing & Communication System (ICACCS), Coimbatore, India, 2019, pp. 604-607.

[7] Agrawal S., Awekar A. (2018) "Deep Learning for Detecting Cyberbullying Across Multiple Social Media Platforms", In: Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds) Advances in Information Retrieval. ECIR 2018. Lecture Notes in Computer Science, vol 10772. Springer, cham.

[8] Brown, E. Clery and C. Ferguson, "Estimating the prevalence of young people absent from school due to bullying", National Center for Social Research, 2011.

[9] Monirah A., Al-Ajlan, Mourad Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning", 978-1-5386-4110-1, IEEE-2018.

[10] Vandana Nanda Kumar, Binsu C, Kovoor, Sreeja M.U., "Cyber-Bullying Revelation in Twitter Data using Naïve-Bayes Classifier Algorithm" International Journal of Advanced Research in Computer Science. Volume 9, No. Jan-Feb 2018.

[11] Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" arXiv preprint arXiv:1810.04805, 2018 unpublished.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention is all you need" Advances in Neural Information Processing Systems 30 (NIPS 2017).

[13] K. Reynolds, A. Kontostathis, and L. Edwards. "Using machine learning to detect cyberbullying" In ICMLA, pages 241-244, 2011.

[14] E. Wulczyn, N. Thain, and L. Dixon. "Ex machina: Personal attacks seen at scale". In WWW, pages 1391-1399, 2107.

[15] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation" In EMNLP, pages 1532-1543, 2014.

[16] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification" In ACL, pages 1555-1565, 2014.