

Cyberbullying Detection and Handling: A Deep Learning Approach

A Anandha Krishnan¹, Akhil C Jose², Allen Jiji³, Anand Edwin A⁴ and Fasila K A⁵

¹⁻⁵Muthoot Institute of Technology and Science, Puthencruz, India

Email: {18cs054, 18cs276, 18cs050, 18cs053, fasilaka}@mgits.ac.in

Abstract—In the present information age, there's been an increase in the use of social media, as a result, there has been a parallel increase in cyberbullying instances. Cyberbullying can cause many drastic and negative impacts on a person's life, eventually leading to depression or suicide. Detection and prevention of such comments/posts are very difficult as the model used must be reliable. These traditional models may sometimes lead to false detection. With the arrival of contextual embeddings into the scene, the performance of Natural Language Processing (NLP) models has significantly increased. We propose a solution to cyberbullying by building an extension that uses a deep learning model in the back to detect and hide these toxic comments from the user.

Index Terms— cyberbullying, detection, pre-trained BERT, LSTM and Natural Language.

I. INTRODUCTION

The internet usage is increasing exponentially and in the pandemic and post-pandemic scene, the internet has turned into an essential service required by people. With the increase in usage of the internet, crimes in cyberspace have been increasing as well. One such crime is cyberbullying. Cyberbullying is defined as a form of bullying that takes place through the use of electronic communication. Cyberbullying has turned out to be a great problem now that countries have implemented laws preventing cyberbullying. It is more common among people within the age range of 13 to 28 [1]. Cyberbullying is faced by all parts of the internet using society from common people to celebrities. Cyberbullying can affect the victim in multiple ways like anxiety, and depression and can even lead to suicidal thoughts.

Cyberbullying can occur in multiple forms, it can be verbal, textual, or in the form of images. We aim to handle cyberbullying in textual form. To solve the problem of cyberbullying, we aim to create a browser extension that hides the cyberbullying comments/posts from the end-user. The extension classifies the comments/posts with the help of a deep learning model in the back.

Cyberbullying detection is a complex problem and can't be solved with simple rule-based approaches. For example, all toxic comments don't contain swear words and all swear words need not be toxic. The problem requires an understanding of the comment. This led to the usage of machine learning and deep learning approaches to the problem. The related past works prove that deep learning gave a better performance on the problem of cyberbullying detection. There are many models based on both machine learning and deep learning to classify toxic comments. Earlier models used built-in word embeddings or pre-trained non-contextual embeddings like word2vec, GloVe, etc. Now with the availability of a new type of word embeddings called contextual embeddings, the performance of Natural Language Processing (NLP) models has significantly

increased [2]. One such contextual embedding model is the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model, a technique created by Google Artificial Intelligence research labs [13].

We propose a new model that uses BERT based embedding technique with Bidirectional long-short term memory (BiLSTM) to classify toxic comments. We propose a browser extension that uses the above deep learning model in the back to detect and hide the toxic comments from the end user's screen. The extension is designed to work on social media sites like Facebook, Instagram, Twitter, and Youtube. These are the major platforms where users suffer from cyberbullying [14].

II. METHODOLOGY

A. Dataset

The dataset used is Jigsaw unintended bias in toxicity classification dataset from Kaggle. The dataset is obtained from an application called Civil Comments. The dataset contains 1804874 rows with comments with a mean length of 300 characters. It is a commenting plugin for independent news sites. The dataset contains many features, but the relevant features are comment text and target, respectively, the comment and the labeled toxicity score, which ranges from 0 to 1. With 0 being the least toxic and 1 being the most. Comments with targets greater than 0.5 are considered toxic. The dataset is heavily imbalanced with toxic comments comprising just 8% of the total rows.

B. Data Preprocessing

Various preprocessing techniques were applied to the comment text.

- The comment texts were converted into lowercase to ease the other preprocessing steps.
- The next step was to remove contractions from the sentence. Contractions are words or combinations of words that are shortened by dropping letters and replacing them with an apostrophe. Contractions in the sentence were replaced by respective expansions.
- HTML tags, URLs, and special characters were removed from the comment text as they do not contribute much to the meaning [17].
- The toxicity labeled was of a continuous value between 0 and 1. We converted the toxicity into discrete values of 0 and 1 with a threshold of 0.5. Comments with toxicity greater than 0.5 were considered toxic.
- Usual text preprocessing techniques like stemming, lemmatizing, and stopwords removal was not used as the proposed method uses the BERT model which captures information about each word in a sentence to get its context. The above-mentioned preprocessing techniques may degrade the model's performance.
- Emoji removal was not done as the BERT model accepts emoji as input and can determine context from that as well.
- Comment texts with character lengths greater than 900 were removed as those were detected as outliers by the Interquartile Range (IQR) method of outlier detection. Also, these comments are not representative of the comments/posts in the platforms we aim to implement the extension.

C. Feature Engineering

Many features were obtained from the given comment text.

- The polarity of the text was determined using the sentiment analysis method. Polarity ranges from -1 to 1. Where -1 denotes the most negative sentiment and 1 represents the most positive.
- Many other features like comment length, the number of words, average word length, the number of unique words, and other features were also determined from the comment text.

D. Data Oversampling

Almost all cyberbullying datasets are imbalanced with a surplus amount of non-bullying comments and a lack of bullying comments. This imbalance can cause a bias in model prediction as it tends to predict the non-bullying comments more often. Various techniques can be used to solve the problem of class imbalance. These include techniques like undersampling, oversampling, synthetic minority oversampling technique (SMOTE), and assigning class weights [18][19]. Undersampling leads to wastage of useful data while oversampling a highly imbalanced large dataset doubles the amount of data which leads to an increase in training time which is one

disadvantage of SMOTE technique. SMOTE is not an effective method for resolving class imbalance in text-based data as the generated text samples may not possess any meaning which can lead to another bias.

Because of the above reasons, we used the class weighting method which penalizes the loss of the minority class and reduces the loss of the majority class. This balances out the bias created by the imbalanced data. By calculations, we assigned the weight for the positive class as 6.1 and 0.54 as the weight for the negative class.

E. Word Embeddings

The performance of an NLP model purely depends on the type of embedding used. So the selection of a good word embedding layer is significant while designing an NLP model. It is proven that pre-trained contextual word embeddings outperform the normal word embedding in all aspects. Our choices for word embeddings were 4 BERT-based models BERT, ALBERT, RoBERTa, and DistilBERT.

TABLE I. COMPARISON OF DIFFERENT WORD EMBEDDINGS

Word Embedding	F1 Score	Accuracy
BERT	0.89	0.92
ALBERT	0.89	0.9
RoBERTa	0.88	0.89
DistilBERT	0.89	0.89

To test the performance of different word embeddings on the dataset, each embedding technique was used as the embedding layer for the model on an undersampled version of the dataset. The corresponding F1 score and accuracy were recorded which is shown in table I. According to the results BERT was selected as the embedding layer because of the best F1 score and accuracy.

F. Model Architecture

The model architecture of the neural network used to classify the comments is shown in Fig. 1. The input layer of the neural network takes in the corresponding text. The input text is passed on to ALBERT embedding layer which converts the sentence into a 2-dimensional vector of dimension ($n, 768$), where n is the number of words in the input sentence. It is followed by a spatial dropout layer with a dropout rate of 0.2 which is used to regularize the model.

The output from the dropout layer is passed to the 2 bidirectional Long short-term memory (LSTM) layers which are chained together. Both the biLSTM layers consist of 128 nodes. The initial layer of the LSTM gives a hidden state to each time step of the input, in another instance the hidden state output is coupled with the state of the cell for the last time step of the input. In biLSTM, the sequence is processed in both directions.

The output of the chained biLSTM layers is supplied to a global max-pooling layer and a global average pooling layer [8]. These two layers perform the appropriate pooling functions and the output is concatenated into a single vector.

The concatenated outputs are passed into a skip connection which is used to smoothen the cost function surface [6]. This helps reduce the training time by providing easy and smooth convergence. Both the dense layers used in the skip connections consist of 512 nodes. Finally, the output vector from the skip connection is passed to a dense layer with one node and a sigmoid activation on it. This layer acts as the output layer which returns the corresponding toxicity score of the input text. The output layer gives an output between 0 and 1.

III. RESULTS

The model was trained by making use of the backpropagation approach. There are various hyperparameters that influence the model training. For this model, the optimizer used was the Adam optimizer along with a binary cross-entropy loss function [19]. The model was trained for 5 epochs with a learning rate of 0.001.

10% of the dataset was kept apart for testing purposes. The test set has the same ratio of positive and negative classes as the original dataset.

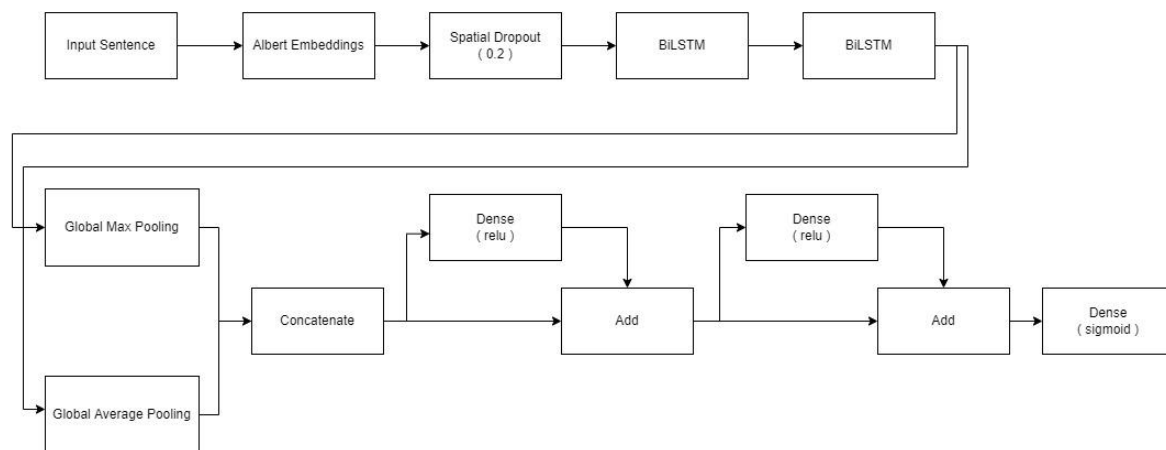


Figure 1. Model Architecture

Here the model was able to achieve a train and test accuracy of 95% and a test F1 score of 0.93. The model was able to detect both bullying and non-bullying comments without bias. However, the model gets the context wrong sometimes, especially in long sentences.

The model was deployed using Tensorflow serve which is used to deploy deep learning and machine learning algorithms in a production environment. The hosted model was connected with the extension which communicates with the deployed models to classify comments that appear on the end-user system.

Once the extension identifies that a comment is bullying, it hides it even before reaching the screen.

IV. CONCLUSION AND FUTURE WORK

Our approach was able to get much better results than traditional rule-based and machine learning approaches. It was also able to outperform deep learning models which made use of old non-contextual embedding techniques. The extension was able to catch almost all cyberbullying comments faced by a normal end-user. While extension performed well on sites like Facebook, Instagram, Youtube, and Twitter, the accuracy was a bit low on Reddit because the dataset didn't represent long comments as they appear on Reddit. The results can be improved by collecting data that represents comments on Reddit. By changing the hyperparameters of models and along with additional data, there may be an increase in the performance of the model. Working on these mentioned areas of improvement is kept for future work.

REFERENCES

- [1] Whittaker, Elizabeth & Kowalski, Robin. (2014). Cyberbullying Via Social Media. *Journal of School Violence*. 14. 11-29. 10.1080/15388220.2014.949377.
- [2] Alessio Miaschi and Felice Dell'Orletta. 2020. Contextual and Non-Contextual Word Embeddings: an in-depth Linguistic Investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.
- [3] 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.
- [4] Monirah A. Al-Ajlan, Mourad Ykhlef, "Optimized Twitter Cyberbullying Detection based on Deep Learning", 978-1-5386-4110-1, IEEE-2018.
- [5] He, Kaiming et al. "Deep Residual Learning for Image Recognition." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 770-778.
- [6] N. Mulliqi et al., "The Importance Of Skip Connections In Encoder-Decoder Architectures For Colorectal Polyp Detection," 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 380-384, doi: 10.1109/ICIP40778.2020.9191310.
- [7] M. Munikar, S. Shakya and A. Shrestha, "Fine-grained Sentiment Classification using BERT," 2019 Artificial Intelligence for Transforming Business and Society (AITB), 2019, pp. 1-5, doi: 10.1109/AITB48515.2019.8947435.
- [8] Nagi, Jawad & Ducatelle, Frederick & Di Caro, Gianni & Ciresan, Dan & Meier, Ueli & Giusti, Alessandro & Nagi, Farrukh & Schmidhuber, Jürgen & Gambardella, Luca Maria. (2011). Max-pooling convolutional neural networks for vision-based hand gesture recognition. 2011 IEEE International Conference on Signal and Image Processing Applications, ICSIPA 2011. 342-347. 10.1109/ICSIPA.2011.6144164.

- [9] Huang, W., Song, G., Li, M., Hu, W., Xie, K. (2013). Adaptive Weight Optimization for Classification of Imbalanced Data. In: Sun, C., Fang, F., Zhou, ZH., Yang, W., Liu, ZY. (eds) *Intelligence Science and Big Data Engineering. IScIDE 2013. Lecture Notes in Computer Science*, vol 8261. Springer, Berlin, Heidelberg.
- [10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- [11] S. Liu, H. Tao and S. Feng, "Text Classification Research Based on Bert Model and Bayesian Network," 2019 Chinese Automation Congress (CAC), 2019, pp. 5842-5846, doi: 10.1109/CAC48633.2019.8996183.
- [12] Y. Tian and Z. Zhang, "Text Classification Model Based on BERT-Capsule with Integrated Deep Learning," 2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA), 2021, pp. 106-111, doi: 10.1109/ICIEA51954.2021.9516041.
- [13] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv*, abs/1810.04805.
- [14] Kwan, Grace & Skoric, Marko. (2013). Facebook bullying: An extension of battles in school. *Computers in Human Behavior*. 29. 16–25. 10.1016/j.chb.2012.07.014.
- [15] Gaffney, H., Farrington, D. P., Espelage, D. L., & Ttofi, M. M. (2019). Are cyberbullying intervention and prevention programs effective? A systematic and meta-analytical review. *Aggression and violent behavior*, 45, 134-153.
- [16] Menesini, E., & Spiel, C. (2012). Introduction: Cyberbullying: Development, consequences, risk and protective factors. *European Journal of Developmental Psychology*, 9(2), 163-167.
- [17] Camacho-Collados, J., & Pilehvar, M.T. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. *ArXiv*, abs/1707.01780.
- [18] Yu, Lian and Nengfeng Zhou. "Survey of Imbalanced Data Methodologies." *ArXiv* abs/2104.02240 (2021): n. pag
- [19] Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61, 863-905..
- [20] Kingma, Diederik P. and Jimmy Ba. "Adam: A Method for Stochastic Optimization." *CoRR* abs/1412.6980 (2015): n. pag.