Detection of Hate Speech in Bangla Language in Social Media Using BERT

Sakib Ahmed, Abu Nayeem Tasneem, Mosarrat Rumman, Najib Murshed
Mofaqkhayrul Islam, Sahiba Tasneem and Md. Tarikul Islam
Department of Computer Science and Engineering
Brac University
Mohakhali, Dhaka - 1212, Bangladesh
{sakib.ahmed,abu.nayeem.tasneem,mosarrat.rumman,md.najib.murshed}@bracu.ac.bd
{mofaqkhayrul.islam,sahiba.tasneem,md.tarikul.islam}@bracu.ac.bd

Abstract

As people are getting more access to social platforms, cyber bullying is also increasing significantly. Although much research has already been done to detect hate speech in social media, very little progress has been made in Bangla language. Therefore we aim to propose a model to detect hate speech in bengali using transformer based model BERT (Bidirectional Encoder Representations from Transformers) by extending "Bengali Hate Speech Dataset 2.0" and using ELECTRA in our pretraining model as it provides better results in performing downstream tasks.

1 Introduction

Rising exposure to social media brings about interaction with people from different walks of life. Although the increased interactions brings forth many positive experiences, there has also been a rise of negativity fueled sentences or "hate speech" expressing extreme detest towards people of different race, religion, gender, and many more. Filtering hate speech is a relatively new challenge, moreover, very little progress was made on such work in Bengali. To tackle this problem, we propose a context based language model, BERT, which is trained using word embeddings, to identify hate speech made in Bengali accurately.

2 Literature Review

At present, the state of the art research for hate speech detection has reached the point where researchers utilize the power of advanced architectures such as transfer learning. For example, in their paper Farha et al.(Abu Farha and Magdy, 2020), researchers compared deep learning, transfer learning, and multitask learning architectures on the Arabic hate speech dataset. There is also research in identifying hate speech from a multilingual dataset using the pre-trained state of the art models such as mBERT and xlm-RoBERTa in the paper Baruah et al.(Baruah, 2020). The Deep Learning Models have been remarkably paid-off to enhance the accuracy of predicting hate speech specifically by extracting particular semantic features of hatred speech comments(Zhang and Luo, 2018).

3 Methodology

3.1 Dataset

Bengali is severely low resourced in terms of datasets ready for NLP tasks. For Fine-Tuning of the BERT model and using it in Downstream tasks it needs to be trained with labeled data. In their paper Md. Rezaul K. et al.(Karim et al., 2020) have mentioned about "Bengali Hate Speech Dataset 2.0" where they have 6418 labeled dataset of hate comments. We propose to extend the dataset using open source software to scrape web pages for hate speech. Their dataset was primarily constructed using comments from youtube only, we propose taking data from, facebook, tiktok, instagram etc social media websites. After collection of the dataset we propose to normalize the dataset by removing the hashtags, emojis, links, and other tags. Then we annotate the dataset in following categories: political, personal, geopolitical, religious, and gender abusive hate. After that we propose training WordPiece vocabulary to create tokens.

3.2 Pre-Training of BERT

The input embeddings into the BERT model is the sum of Token embeddings, segment embeddings and position embeddings. For pre-training our BERT model, a training sample of embeddings is fed to the BERT with maximum sequence length of 512, [CLS],[E1],[E2]....,[SEP],.....,[En]. Here [CLS] tokens are the first token of the sequence which is a classifier and [SEP] separates two segments in a sentence. In our BERT model, we aim to use ELECTRA(Clark et al., 2020) model instead of the usual Masked Language Model (MLM). ELECTRA uses Replaced Token Detection for pre training. In ELECTRA a generator is fed with an input with masked tokens and asked to predict the masked tokens. A Discriminator is fed with the same input replacing the masked tokens with the generated tokens and asked to identify whether the tokens are generated or original .Electra provides significantly better performance in carrying out downstream tasks in terms of computational efficiency(Bhattacharjee et al., 2021).

3.3 Fine-Tuning of BERT

For fine tuning our pre-trained BERT a fully connected Feed-forward layer is added to the top of the last layer of the pretrained BERT, along with a softmax activation layer for classification. The BERT outputs the same number of tokens as the input, but only the [CLS] token from the output is fed to the fully connected layer and the classification layer as shown in figure 1.

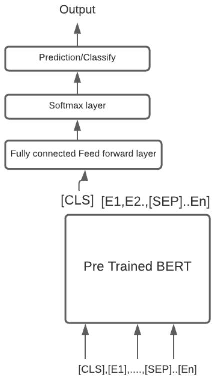

Figure 1: The fine-tuning of the BERT to classify hate category.

The softmax layer outputs a probability of the sequence to belong to a particular hate category and then it is compared with actual value and Cross Entropy loss is calculated.

3.4 Experiment

The fine tuning procedure can be carried out to other pre-trained BERT models such as BanglaBERT(Bhattacharjee et al., 2021) and Multilingual BERT by Google Research (https://github.com/google-research/bert/blob/master/multilingual.md) and the results can be compared to see the efficiency of our pretrained BERT model.

4 Potential Challenges

There are some challenges to implement the idea.

- Bengali is a low resource language with unique script - Might result in very few generated tokens
- Bengali Spelling - Incorrect Bengali spelling will be an obstacle to create proper word embedding.
- Due to having small vocabulary size accuracy of the model might be compromised

Conclusion

State of the art Neural Network models are used to identify ”Hate Speech” in social media in different languages. Our goal is to make a large data set of Bangla ”Hate Words” and train our BERT model to also understand the context those words are being used in a social media post or comment section to fight against hate crime and cyber bullying.

References

Ibrahim Abu Farha and Walid Magdy. 2020. Multi-task learning for Arabic offensive language and hate-speech detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 86–90, Marseille, France. European Language Resource Association.

Das Kaushik Barbhuiya Ferdous Dey Kuntal Baruah, Arup. 2020. Aggression identification in English, Hindi and Bangla text using BERT, RoBERTa and SVM.

Abhik Bhattacharjee, Tahmid Hasan, Kazi Samin, M. Rahman, Anindya Iqbal, and Rifat Shahriyar. 2021. Banglabert: Combating embedding barrier for low-resource language understanding. *ArXiv*, abs/2101.00204.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

Md. Rezaul Karim, Sumon Dey, and Bharathi Raja Chakravarthi. 2020. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. *ArXiv*, abs/2012.14353.

Ziqi Zhang and Lei Luo. 2018. Hate speech detection: A solved problem? the challenging case of long tail on twitter.