Application of BERT Model for Hate Speech Detection

Introduction

Natural Language Processing (NLP), a subfield of Machine Learning, focuses on methods to enable computers to understand, interpret, and generate human language by processing and manipulating textual and linguistic data. In recent years progresses in NLP, has introduced modern-day tools like ChatGPT, which has revolutionised advanced language interactions. Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art NLP model that utilizes the capabilities of Transformers architecture and can perform tasks requiring a complete understanding of a sentence to carry out sentiment analysis or sentence classification, word classification (named entity recognition (NER)), and question answering.[1] BERT based models are increasingly used in various applications and a version of it is selected for this analysis.

Dataset

The dataset is selected from Hugging Face dataset repository which contains labelled train data and unlabelled test data. The data fields are: label (0 denoting 'no-hate-speech' and 1 denoting 'hate-speech') and tweet (text as a string).[2] The training and test dataset has 31,962 and 17,197 entries respectively. Another aspect of the training dataset is that only 2242 tweets were classified as hate speeches. The skewness of the data could lead to misrepresentation in the model training. Therefore, before model training, in the train_test_split ‘stratify’ is applied to include a similar proportion of 2 labels in them.

The dataset preprocessing included: cleaning the input text by removing special characters, symbols, and non-ASCII characters, and normalizing the text to a standard form. This is useful for text preprocessing before tokenization.

Model Selection and Implementation

BERT is developed by Google AI and has been trained on 3.3 billion word corpus and 1.56 billion tokens.[3] The bidirectional nature of BERT means it considers both left and right context for each word, allowing it to capture rich contextual information. At the core of the BERT Transformer architecture are two fundamental components: the masked language model (MLM) and the transformer encoder. The masked language model is self-supervised pertaining to learn the bidirectional representation of contexts where a percentage of the words are masked and the model is trained to predict them. This leads BERT to understand the relationships between words in a sentence. On the other hand, the transformer encoder consists of multi-layered self-attention and feed-forward neural networks.[4]

For this analysis, “bert-base-cased” checkpoint is selected. This is one of the pretrained BERT models provided by the Hugging Face Transformers library. The model has 110 million parameters and consists of 12 transformer layers, making it a relatively smaller BERT variant. This model preserves the cases of the original texts which allows more of the information from the source to be retained.

The model is initiated with the trained weights as part of transfer learning. However, the hyperparameters needed to be adjusted for training so that the model can perform the classification task on our dataset. The fine tuning required a change in optimizer, the learning rate and batch size. AdamW optimser is selected with the learning rate of 1e-8 which is a standard for language models. The advantage of AdamW is that it allows weight decay into the optimisation steps which maintains a stable training prrocess. The data loading pipeline is also optimized.

The loss function is sparse categorical crossentropy and training metric has been used is Accuracy.

Model Evaluation

The model is run for 3 epochs and reached a good fit quickly because of the pre-trained weights from transfer learning.

Test Loss:	0.154
Test Accuracy:	0.956

Classification Report:				
	precision	recall	f1-score	support
0	0.98	0.97	0.98	3237
1	0.66	0.75	0.7	244
accuracy			0.96	3481
macro avg	0.82	0.86	0.84	3481
weighted avg	0.96	0.96	0.96	3481

Table 1: Model training result

Limitations

The training dataset contained 2242 hate speech entries which are only 7% of the total data. The result shows that the model categorises some speech as hate speech which is not necessarily so. This is because the training data label has implicit biases in it. For instance, the below tweets are categorised as hate speech but the perception is subjective as both of them appear to be constructive criticism as opposed to being emotional outbursts of hatred. Overall, the model classified 1416 of 17197 unlabelled test entries as hate speech.

i) @user the uk governments new #anti-semitism definition conflates with valid criticism of #israel | opendemocracy

ii) be careful in criticizing #obama for his decision on #israel &amp; sanctions against #russiahacking , as #liberals will consider this

In conclusion, the model has shown great capabilities to classify a hate speech quite accurately. The bias in the input data is the human element which is beyond the scope of this analysis. This model can be utilized to identify hateful/biased/racist comments and posts on social media platforms.

References

1. Hugging Face. (2023). Hugging Face course on NLP.
Retrieved 16 August, 2023, from https://huggingface.co/learn/nlp-course/chapter1/1?fw=tf

2. Hugging Face. (2023). Dataset: Tweets Hate Speech Detection.
Retrieved 16 August, 2023, from https://huggingface.co/datasets/tweets_hate_speech_detection

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (Year). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Google AI Language.
Retrieved 16 August, 2023, from https://arxiv.org/pdf/1810.04805.pdf

4. Vaswani, A., et al. (2017). Attention Is All You Need. Google Brain.
Retrieved 16 August, 2023, from https://arxiv.org/abs/1706.03762