Fake News Classification with BERT

Andrey Malakhov, Alessandro Patruno, Stefano Bocconi
andrey@zephyros.solutions,alex@zephyros.solutions,stefano@zephyros.solutions

ABSTRACT

This paper describes the usage of the BERT family transformers for the multi-class classification task "FakeNews: Corona virus and 5G conspiracy" track. This is a Natural Language Processing based Fake News detection challenge organized by MediaEval. It demonstrates how one can benefit from using pretrained transformers for tweet discrimination.

1 INTRODUCTION

We investigated the application of the Bidirectional Encoder Representations from Transformers (BERT) model for a classification task on a Twitter dataset. The task is described in the MediaEval webpage [1] and is discussed in Pogorelov et al. 2020 [2]. The dataset is composed of two parts, the full-text tweets content and a sequence of images representing graph networks. The dataset is described in Schroeder et al. 2019 [3]. Due to limited amount of time available, we decided to work only on the full-text tweets and discarded the network graph information. We trained a classifier on top of the pretrained BERT base version in order to discriminate between 5G conspiracy, other conspiracy or non-conspiracy tweets.

2 RELATED WORK

The analysis is mainly based on the BERT model, which is described in Devlin et al. 2018 [4]. The BERT model adds bidirectionality to the language model, which is based on (but different from) the unidirectional architecture of the original Transformer paper (Vaswani et al. 2017) [5].

3 APPROACH

This section describes steps that were applied in order to train the model.

3.1 Data preprocessing

From the train data the tweet text was extracted, we didn't use any metadata, the model is purely trained on text. The first step was to exclude the hash sign (#) from the text. This is necessary for the case when hashtags are incorporated into the tweet text as a part of speech, for example:
*This is an **#example** of a **#tweet** with **#hashtags** within the **#text***
The second step is to replace user mentions in tweets, for example in case of a reply to multiple users. In this case the length of the text can be inflated and this can lead to problems given the sequence size limitation for the BERT model. Also this is done to eliminate the lack of information in the username for our task. The example below illustrates the replacement:
original tweet: ***@username1, @username2, @username3*** test message
processed tweet: ***username*** test message
The third step is to use the BERT *tokenizer*. The tokenizer prepares the text to be fed to the model by splitting the sentence in tokens, adding special tokens (for example to mark the beginning of a sentence), padding the sentence, etc. We limited the length of the tweets up to 256 tokens so that when the text is shorter than that, it will be padded accordingly. If the length of the tweet is larger than 256, the sentence is truncated. Very few tweets exceeded this length, so that this choice shortens the computational time without a significant loss of information and therefore appears to be justified.

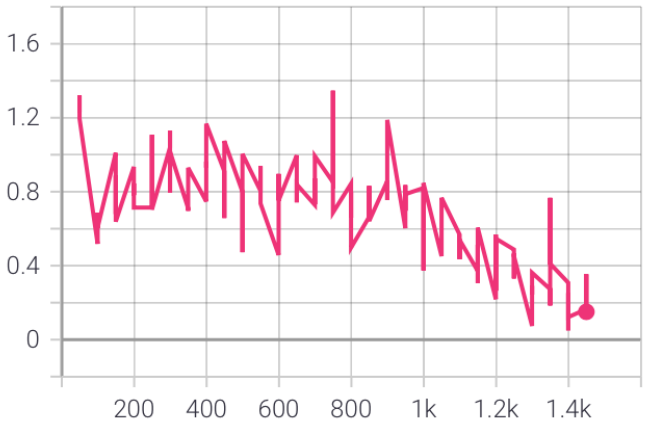

Figure 1: Training Loss

3.2 Classification model

The model uses a simple linear layer on the output of the BERT base with three output neurons. Adding a second linear layer does not improve the results significantly. The optimizer for the model was chosen to be ADAM without weight decay for bias and normalization layers.

4 RESULTS AND ANALYSIS

The model was trained for 20 epochs. In Figure 1 and Figure 2 you can see the log-loss per training iteration for both training and validation datasets.

We chose the epoch for our final model based on accuracy per class, in the end the epoch 19 was chosen with a confusion matrix in Table 1. The application of the model on the test set gave us a final score of 0.400846 when using the official metric of the MediaEval task.

MediaEval'20, December 14-15 2020, Online

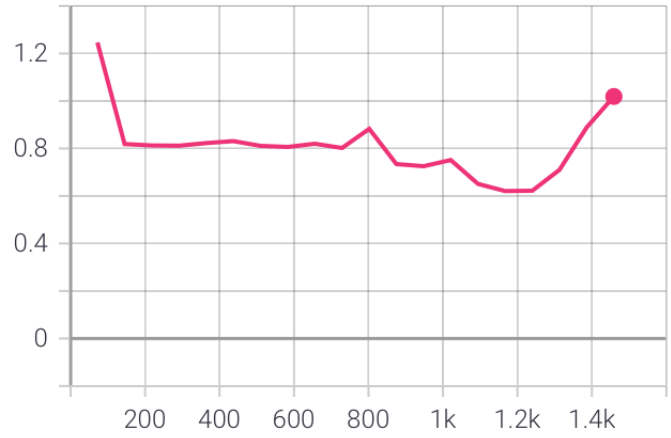

Figure 2: Validation Loss

predicted_target target	5G conspiracy	Other conspiracy	Non-conspiracy
5G conspiracy	150	29	39
Other conspiracy	37	46	50
Non-conspiracy	92	113	608

Table 1: Confusion matrix for validation set

5 DISCUSSION AND OUTLOOK

We can see that even with a simple classifier we can reach a significant accuracy on the raw text tweets. The following improvements can be done:

(1) Use multiple folds over train data to get a robuster model
(2) Increase complexity of the classifier with several layers
(3) Use BERT Large as a foundation instead of BERT base
(4) Incorporate meta data from tweets as number of replies, likes, retweets, mentions, etc. to the classifier part
(5) Use an ensemble of transformers (ROBERTA, ALBERT, BERT, etc.)

REFERENCES

[1] https://multimediaeval.github.io/editions/2020/tasks/fakenews/
[2] Pogorelov et al., 2020, "FakeNews: Corona Virus and 5G Conspiracy Task at MediaEval 2020", MediaEval 2020 Workshop
[3] Schroeder et al. 2019, "FACT: a Framework for Analysis and Capture of Twitter Graphs.", In 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), pp. 134-141. IEEE, 2019.
[4] Devlin et al. 2018, 2017, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", Arxiv 1810.04805
[5] Vaswani et al., 2017, "Attention is all you need", Arxiv 1706.03762