Covid Fake News Detection using Masked Language Models

Stefan Negulescu, Andrei Zenoveiov, Florin Brad
University of Bucharest, Romania

Introduction

Faster communication and worldwide connectivity has enabled fake news to proliferate on social media. This phenomenon grew during the Covid-19 pandemic, when many lies, rumours and ill-intended posts have been rapidly spread on social media.

Fake news detection is usually treated as a binary classification problem and has been tackled with classical machine learning algorithms: support vector machines, Naïve Bayes, logistic regression, k-nearest neighbors.

To avoid feature engineering, deep learning models like CNN, LSTM and Transformer-based models have also been recently applied.

We tackle this problem with state-of-the art methods, namely BERT-based finetuning and prompt engineering. The fully trained DistilBERT outperforms null prompt fine-tuning and a strong bag-of-words baseline. We also use explainable AI techniques from the Captum library1, such as Integrated Gradients (IG) and attention scores to reveal the most important words in fake and normal tweets.

Dataset

We used the Covid-19 Fake News Dataset, which consists of 12700 Covid-19 related posts, collected from various social media platforms. The classes are balanced, 52.34% of posts being real news, and 47.66% consisting of fake news.

Attribute	Fake	Real	Combined
Unique words	19728	22916	37503
Avg words per post	21.65	31.97	27.05
Avg chars per post	143.26	218.37	182.57

Table 2: Numeric features of the dataset

Attribute	Fake	Real	Combined
Unique words	19728	22916	37503
Avg words per post	21.65	31.97	27.05
Avg chars per post	143.26	218.37	182.57

Table 2: Numeric features of the dataset

Approaches

The baseline model is a MLP trained over a bag-of-word embeddings representation of the sequence. It uses Glove embeddings of size 200 and it has hidden layer of size 200 with a ReLU non-linearity.

Next, several BERT-based models have been finetuned: DistilBERT (which is a smaller version of BERT) and Covid-BERT, which is a model pretrained on 22 million COVID-19 related tweets. To perform classification, we use a linear layer on top of the h[CLS] sequence representation.

For DistilBERT we train the whole network, but also perform two ablations, where we train the linear parameters and the biases or just the linear parameters.

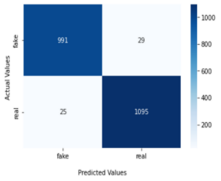


Figure 1: Confusion matrix

Quantitative results

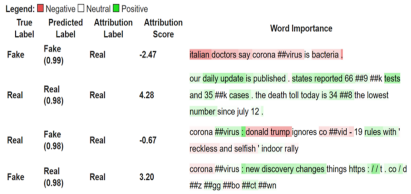
Results in Table 1 show that DistilBERT-uncased has the best performance, close to the top score reported with BERT. Though pretrained on tweets, the CovidBERT model performed slightly worse than DistilBERT.

Model	trained params	F1
DistilBERT	all	97.5%
DistilBERT	linear	94%
DistilBERT	biases+linear	97%
Covid-BERT	all	97%
BERT [19]	all	98%
bag-of-word embeddings	all	92%

Table 1: Results

Freezing all the DistilBERT backbone performed the worst out of all the frozen ablations (94%).

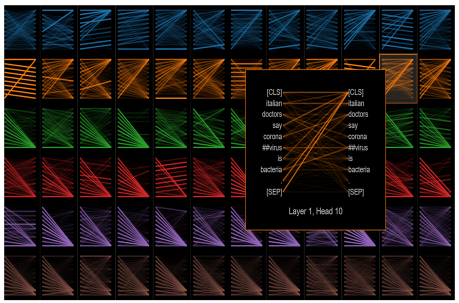
Fig. 1 presents the confusion matrix on the test data.



Qualitative results

In Fig 2 we inspect the most important words for prediction according to the attribution scores from IG. We notice that words in the normal example (reported, published, cases are related to official reporting, while words in the fake example (doctors, corona, bacteria) are related to medicine. The third example is the most confidently predicted false positive example, for which several words (donald, trump, reckless) lower the total attribution score, leading to a fake prediction. In the last example, words associated with real news ('new', 'discovery') trigger a wrong prediction

To consolidate the intuition that each label is correlated with certain topics, we measure the top 100 most important words in the test set for each category, by summing attention scores or attribution scores for each word. We observe a strong correlation between these 2 scores, the resulted sets almost overlapping for each category. We notice in Table 2 that the TP and FP predictions have similar most important words, as well as TN and FN predictions. In addition, the most important words for fake and real predictions have little overlap, consisting primarily of stopwords and neutral words.

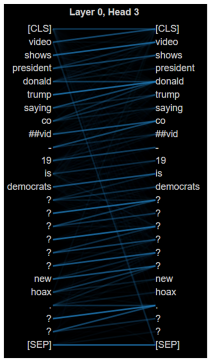


Conclusion

In this experiment, we demonstrated strong performance on Covid-19 fake news detection using a finetuned DistilBERT model. We qualitatively show that important words related to fake news have a distinct topic than the ones related to normal examples (medicine vs official statements). Moreover, links prove to be a spurious feature when testing on cleaned tweets, suggesting that they should be appropriately handled during preprocessing.

Future Work

For future work, better results could be obtained using an updated dataset. Moreover, different methods of tokenizing the dataset (e.g using prompts) could improve overall performance.



References

Hu Zhang, Zhuohua Fan, Jia-heng Zheng, and Quan-ming Liu. An improving deception detection method in computer-mediated communication. J. Networks,7(11):1811–1816, 2012

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. pages 4171–4186, 2019

Sushma Kumari. Nofake at checkthat!2021: Fake news detection using BERT. CoRR, abs/2108.05419,2021

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui,Zhoujun Li, and Philip S. Yu. TI-CNN: convolutional neural networks for fake news detection. CoRR,abs/1806.00749, 2018