Data Mining Technology for Business and Society
Homework 3

Beatrice Nobile (1908315) Katsiaryna Zavadskaya (1847985)

June 2020

Introduction

This paper reports the study of NLP on a specific task, i.e. fact checking, using the BERT pre-training and the LAMA and FEVER repositories. In particular the project has been divided into three main parts:

- **Preprocessing**: This task implied the identification of entities, which had to comply with two strict requirements, namely (1) being single-token entities, and (2) belong to the BERT vocabulary.
- **Prediction**: Here we tested the predictions of BERT by means of the LAMA, and analysed its performance in terms of prediction accuracy.
- **Training a Binary Fact-Checking Classifier**: Finally, we tried to use BERT contextual representations, to train the classifier ourselves on the FEVER train set. For this task we used Linear Discriminant Analysis, and the results obtained featured the best accuracy among all tasks, hence we used this strategy on the final test set.

The overall results of the project are summarized in Table 1, while below we are going to analyze in a few more details the work done.

Tasks	1.1	1.2	1.3	2
Accuracy	0.54979	0.57573	0.39730	0.67635

Table 1: Results on Prediction Accuracy

Task 1: Prediction

As previously said, in this task we are analyzing BERT's performance through the LAMA repository in understanding what is the right token under the [MASK]. That is, we have masked the entities identified during the preprocessing, and then tried to see if the pre-trained language model could guess the correct token. Now, this analysis has been further divided into three steps, i.e. firstly we analyzed the accuracy of the first prediction, then of the first ten, and then the overall trend by using probability a threshold. The results, as reported in Table 1, were 0.55 for the first prediction and 0.58 for the top 10. We tried to optimize this result by adjusting for the character case, lower-casing only slightly improved the accuracy in task 1.2, from 0.574688 to 0.57573, while it had no effect in task 1.1. We also tried stemming, yet again given that we are dealing with entities (i.e. mainly names), it did not result in any improvement in the recorded accuracy. Finally, we analyzed the overall behaviour of the accuracy using a fixed probability threshold. The result of this latter analysis can be found in Figure 1. What this figure shows is that naturally, as we relax the requirements for the predictions, we get less accurate results, that is as the probability threshold increases, the accuracy decreases. The reported "best accuracy" for this analysis has been the accuracy at 0.05 as a threshold, which is where we see the elbow of the graph. In fact, we register a sharp decrease from 0.91 to 0.39 when we

decrease the threshold from 0 to 0.05, while the decrease between thresholds 0.05 and 0.1 is only from 0.39 to 0.32 in accuracy. We also studied the threshold at 0.025, but that would be too stringent, cause in certain cases it wouldn't even consider the first predicted token. Hence, we have considered as acceptable the 0.05 threshold, which gives us the highest acceptable accuracy level.

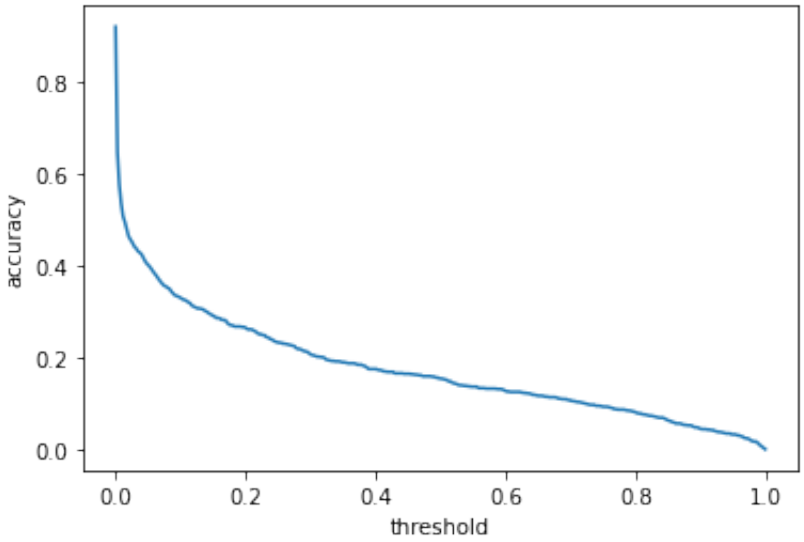

Figure 1: Accuracy as a function of the probability threshold

Task 2: Training a Binary Fact-Checking Classifier

In this task we have trained the model ourselves by means of a Linear Discriminant Analysis (LDA). The choice fell onto this classifier because it provides an optimal separation of the dimensions considered, thus an automatic fine-tuning of the hyperplane separating the two labels in a faster and more efficient way. Indeed, we obtained an accuracy of 0.67635 using as a solver for the LDA the singular value decomposition (svd), making it even more computationally efficient. The results of the training can be found in the file with the final predictions for the test set.