Identifying Age-Related Bias in BERT

Degree programme: BSc in Computer Science | Specialisation: Data Engineering
Thesis advisor: Prof. Dr. Mascha Kurpicz-Briki
Expert: Ciril Saner (Bundesamt für Informatik und Telekommunikation)

Machine learning (ML) is a hot topic and widely used in today's society for different tasks like translation, spam filters or recognizing objects. A subfield of ML is natural language processing, in which texts are processed by machines. These machines, however, are prone to varying kinds of bias and can impose unfair results when using them.

Introduction

To train machine learning models, a lot of computational resources and text data are necessary. Because of this, big companies publish pre-trained machine learning models, that only need some fine-tuning before they are ready to be used, making them more accessible. There is, however, a problem with such models. As they are trained on large amounts of text data, they pick up patterns that lie hidden for humans, introducing bias in the process. Research shows that they are subject to different kinds of bias, for instance gender or ethnical bias. Despite numerous studies present, there is a lack of research in terms of age-related bias in pre-trained models. This thesis aims at contributing to the work present in this field.

Goal of the thesis

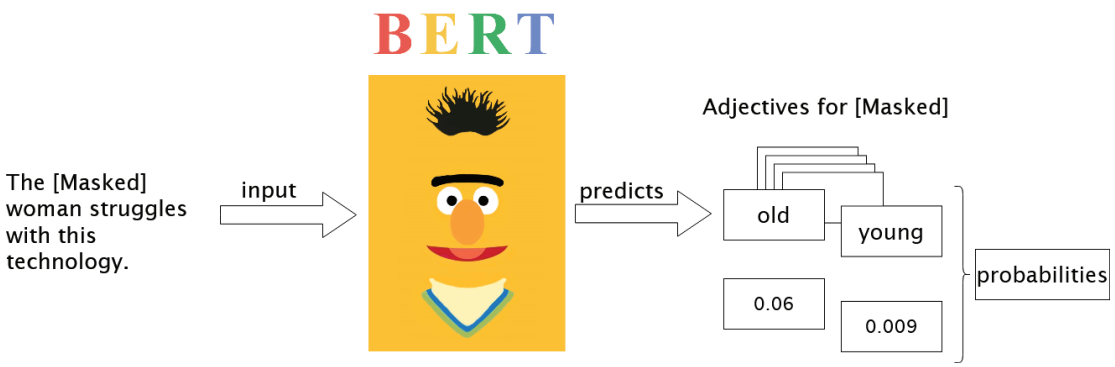
The goal of the thesis is to answer if a pre-trained machine learning model, such as BERT, is affected by age-related bias and if so, by what degree. For this purpose, three different methods were used to evaluate the pre-trained BERT model:

- Contextualized Word Embedding Association Test (CEAT) by Guo and Caliskan – analysis and comparison of word embeddings (numerical representation of words)
- Sentiment analysis – classification of a sentence on a numerical scale, that indicates satisfaction of the writer
- Masked prediction – calculation of probabilities on adjectives occurring for masked words

Results

Out of three experiments conducted, only one indicates the presence of bias in the pre-trained BERT model. Both CEAT and the sentiment analysis fail to reject the null hypothesis, as their respective p-value is bigger than our chosen alpha of 0.05. The third experiment, the masked prediction, however, does indicate the existence of bias. It is important to note that failing to reject the null hypothesis is not the same as accepting it. Meaning, even if two experiments do not indicate the presence of bias, it is wrong to conclude that there indeed is no bias present. This work finds its limitations on the number of sentences used for the experiments, which might directly affect the results and is a point for future improvement.

Marco Wysshaar



Prediction of a masked word