EconBERTa - teaching computer to understand language of economics.

Piotr Antoniak

3rd December, 2021

Abstract

Working with text data, also called Natural Language Processing, has a quite long history in the field of economics. However, latest advances in dealing with this unstructured type of data are used rarely. This work introduces EconBERTa - Large Language Model trained from RoBERTa checkpoint on economics texts. I show that adapting this model on a large data set (4GB of uncompressed text - 4.1B characters, 1B tokens, 660M words), yields a model that is quite good at "understanding" economics texts and can be used as a starting point for multiple tasks such as classification, sentence and documents similarity and summary creation. The entire model, as well as notebooks to train and evaluate its performance are available at [link].

1 Introduction

2 Data

2.1 Textual Data

3 Methodology

References

Devlin, Jacob et al. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv: 1810.04805 [cs.CL].

Kingma, Diederik P. and Jimmy Ba (2014). *Adam: A Method for Stochastic Optimization*. arXiv: 1412.6980 [cs.LG].

Appendices

Appendix A:

- aaa