FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

David Burian

Document embedding using Transformers

Institute of Formal and Applied Linguistics

Supervisor of the master thesis:	Jindřich, Libovický Mgr. Ph.D.
Study programme:	Computer Science
Study branch:	Artificial Inteligence

Prague 2023

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources. It has not been used to obtain another or the same degree.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In .... date ....

Author's signature

Dedication.

Title: Document embedding using Transformers

Author: David Burian

Institute: Institute of Formal and Applied Linguistics

Supervisor: Jindřich, Libovický Mgr. Ph.D., Institute of Formal and Applied Linguistics

Abstract: Abstract.

Keywords: text embedding document embedding transformers document classification document similarity

Contents

Introduction

This is an introduction.

1. Tasks

In this chapter we will describe the task we used to evaluate our model and compare it to the rest.

1.1 Classification

1.1.1 IMDB Sentiment Analysis

In this task the model is asked to classify a piece of text based on sentiment. The texts are anonymized reviews from the Internet Movie Database[1] site collected together with their human-annotated labels that classify the text into two categories: positive and negative. The resulting dataset is commonly referred to as IMDB classification or sentiment dataset Maas et al. [2011].

The dataset is split evenly to test and train set, each having 25000 reviews. The label distribution in both sets is uniform, each of of the two labels is represented by 12500 reviews.

As can be seen from the figure Figure 1.1 the reviews are quite short with only 13.56% being longer than 512 RoBerta tokens.

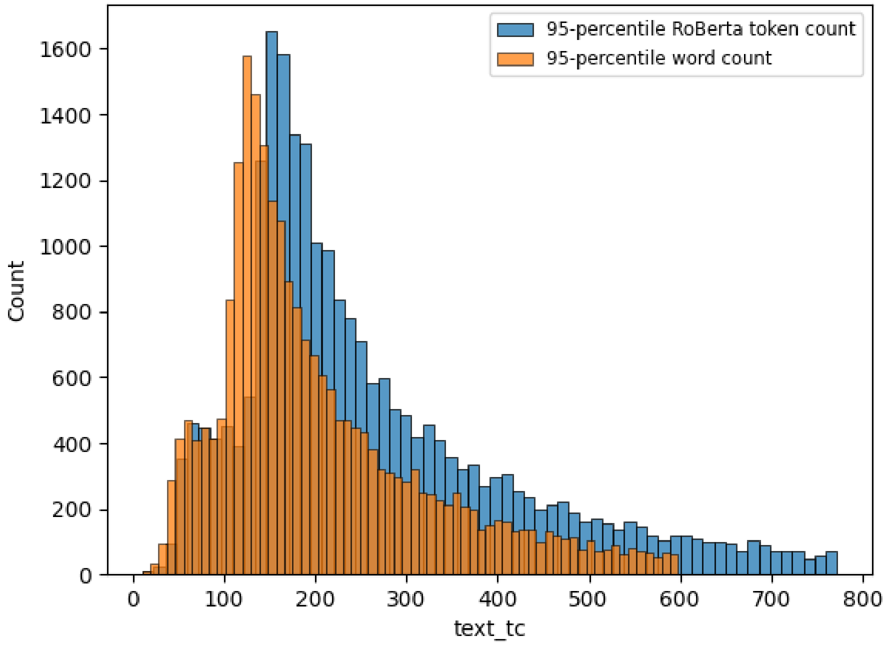

Figure 1.1: Word count and token count distribution of 95-percentiles of reviews. The tokens are generated using RoBerta's pretrained tokenizer from HuggingFace

[1] www.imdb.com

Conclusion

Bibliography

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `https://aclanthology.org/P11-1015`.

List of Figures

List of Tables

A. Attachments