TugaHue: Use of machine learning and transformes to differentiate the national variates of the Portuguese language

Gabriela Villela **Noriega** de Queiroz
<gabri.vnq@usp.br>
Supervisor: Prof. Marcelo Finger

IME-USP Institute of Mathematics and Statistics of the University of São Paulo

TugaHue: Uso de aprendizado de máquina e *transformers* para distinguir as variedades nacionais da língua portuguesa

Gabriela Villela **Noriega** de Queiroz
<gabri.vnq@usp.br>
Orientador: Prof. Marcelo Finger

IME-USP Instituto de Matemática e Estatística da Universidade de São Paulo

Introduction

In this project, two pretrained large language models (LLMs) and two training datasets were combined into four machine learning models for language and dialect classification.

The two LLMs used were BERT multilingual (mBERT) and BERTimbau, an mBERT variant that was trained with more Portuguese language texts.

The models were evaluated over the DSL-TL dataset which is a composed of tweets by Brazilian and Portuguese authors.

The main conclusion was that BERTimbau is noticeably better than mBERT at distinguishing between Brazilian and European Portuguese.

Introdução

Neste projeto, dois grandes modelos de linguagem (LLMs) pretreinados e dois conjuntos de dados (datasets) para treinamento foram combinados em quatro modelos de aprendizado de máquina para classificação de língua e dialeto.

Os dois LLMs emrpegados foram o BERT multilingue (mBERT) e o BERTimbau, uma variante do mBERT que foi treinada com mais textos na língua portuguesa.

Os modelos foram avaliados sobre o dataset DSL-TL que é composto de tweets por autores brasileiros e portugueses.

A principal conclusão foi que o BERTimbau é notavelmente melhor que o mBERT em distinguir entre o português brasileiro e europeu.

Motivation

This project aims to assist in the cleanup of the Coralina *corpus* which is the largest publicly available *corpus* of Brazilian Portuguese. However some of the texts in it have quotations in foreign languages and others lack a clear author. That is, the *corpus* is currently contaminated with other languages and other Portuguese dialects.

Motivação

Esse projeto tem por objetivo auxiliar na limpeza do *corpus* Carolina que é o maior *corpus* de português brasileiro publicamente disponível. Entretanto, alguns dos textos nele tem citações em línguas estrangeiras e outros carecem de autor claro. Ou seja, o *corpus* está atualmente contaminado com outras línguas e outras variedades do português.

Methodology

This project was coded in Python and used the Hugging Face library. The model training was performed on Google Colab Pro on instances with a T4 GPU.

The first dataset used was nicknamed CETEN-CETEM and it is a balanced random sample of the CETEMPublico and the CETENFolha *corpora*. Both are comprised of newspaper excerpts on the same topics and roughly of the same time period.

The second dataset was nicknamed Coraline (a pun on Carolina) and it is a comprised of three international agreements in the original English, Spanish, and French versions as well as the official translations into Portuguese by the governments of Brazil and Portugal.

The models trained on CETEN-CETEM were nicknamed 2L-BERTimbau and 2L-BERT (here BERT and mBERT are synonymous). The 2L refers to the fact that they are binary classifiers. The models trained on Coraline were nicknamed 6L-BERTimbau and 6L-BERT. The 6L refers to the fact that they are multilabel classifiers with six outputs (en, es, fr, pt, pt-BR, and pt-PT).

Metodologia

Este projeto foi programado em Python e usou a bilioteca Hugging Face. O treinamento dos modelos foi realizado no Google Colab Pro em instâncias com uma GPU T4.

O primeiro dataset utilizado foi apelidado de CETEN-CETEM e é de uma amostra balanceada dos *corpora* CETEMPublico e CETENFolha. Ambos são compostos de excertos de jornais sobre os mesmos tópicos e, aproximadamente, do mesmo período.

O segundo dataset foi apelidado de Coraline (um trocadilho com Carolina) e é composto de três acordos internacionais em suas versões originais em inglês, espanhol e francês bem como como suas traduções oficiais para o português pelos governos do Brasil e de Portugal.

Os modelos treinados sobre o CETEN-CETEM foram apelidados de 2L-BERTimbau e 2L-BERT (aqui BERT e mBERT são sinônimnos). O 2L se refere ao fato de que eles são classificadores binários. Já os modelos treinados sobre o Coraline foram apelidados de 6L-BERTimbau e 6L-BERT com o 6L se referindo ao fato de que são classificadores multi-rótulo com seis saídas (en, es, fr, pt, pt-BR, pt-PT).

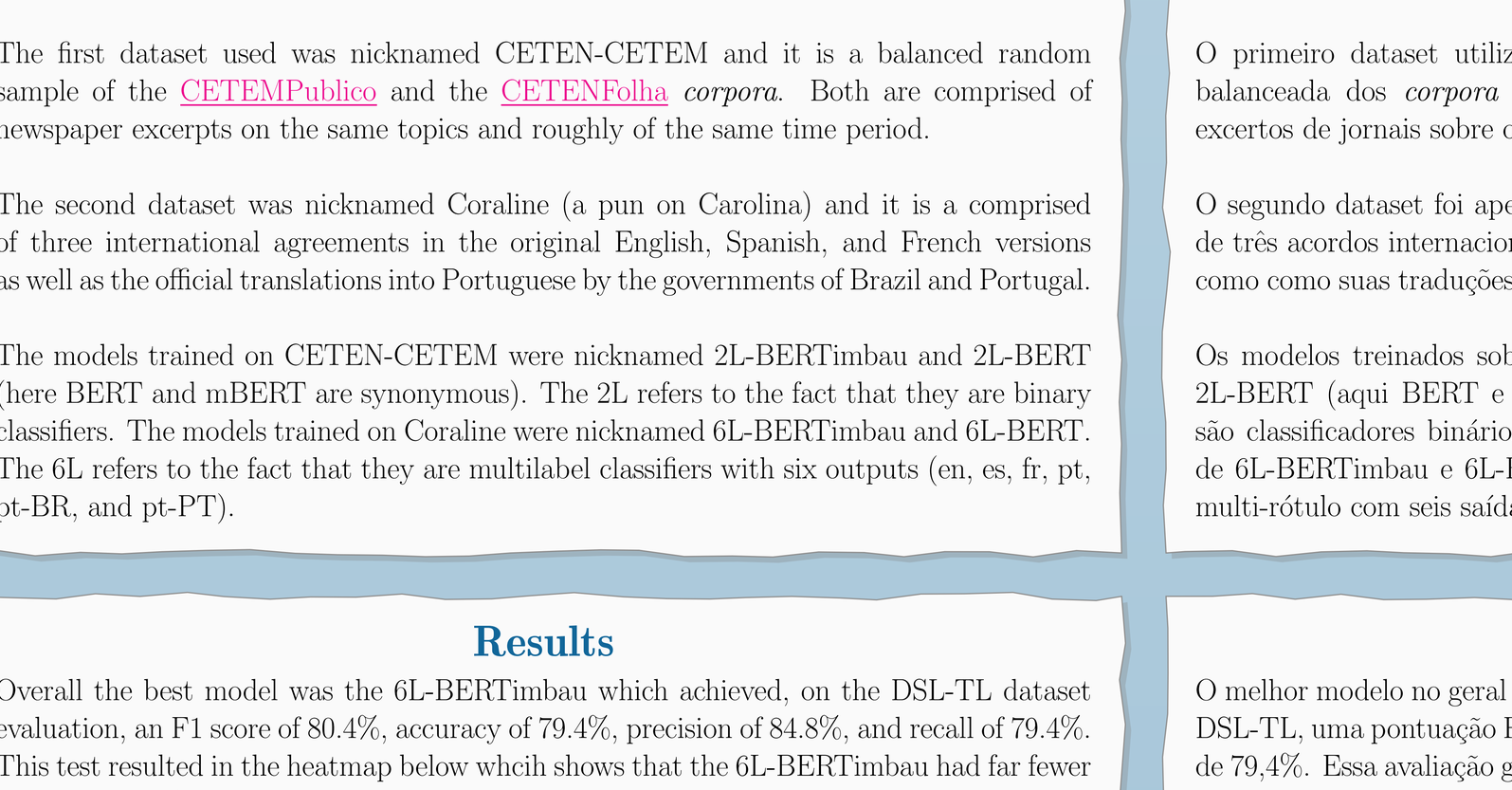
Results

Overall the best model was the 6L-BERTimbau which achieved, on the DSL-TL dataset evaluation, an F1 score of 80.4%, accuracy of 79.4%, precision of 84.8%, and recall of 79.4%. This test resulted in the heatmap below whcih shows that the 6L-BERTimbau had far fewer false positives for European Portuguese than for Brazilian Portuguese.

Resultados

O melhor modelo no geral foi o 6L-BERTimbau o qual atingiu, na avaliação sobre o dataset DSL-TL, uma pontuação F1 de 80,4%, acurácia de 79,4%, precisão de 84,8%, e sensibilidade de 79,4%. Essa avaliação gerou o mapa de calor abaixo o qual mostra que o 6L-BERTimbau teve bem menos falsos positivos para o português europeu do que para o brasileiro.

Heatmap / Mapa de calor



Digital Version / Versão Digital

https://linux.ime.usp.br/~gabrivnq/mac0499/poster.pdf