University of São Paulo
Institute of Mathematics and Statistics
Bachelor of Computer Science

TugaHue

Use of machine learning and transformers to differentiate the national variates of the Portuguese language

Gabriela Villela Noriega de Queiroz
("Gabriela Noriega")

Final Essay

mac 499 — Capstone Project

Supervisor: Prof. Dr. Marcelo Finger

São Paulo
2023

```
@MastersThesis{tugahue2023,
  author = {given=Gabriela, prefix={Villela}, family=Noriega, suffix={
     de Queiroz}, full={Gabriela Villela Noriega de Queiroz}},
  title = "TugaHue",
  subtitle = "Use of machine learning and transformes to differentiate
      the national variates of the Portuguese language",
  school = "Universidade de São Paulo",
  year = "2023",
  type = {Bachelor's Thesis},
  month = "12",
}
```

Ficha catalográfica elaborada com dados inseridos pelo(a) autor(a)
Biblioteca Carlos Benjamin de Lyra
Instituto de Matemática e Estatística
Universidade de São Paulo

Noriega, Gabriela
TugaHue: Uso de aprendizado de máquina e transformers para distinguir as variedades nacionais da língua portuguesa / Gabriela Villela Noriega de Queiroz; orientador, Marcelo Finger. - São Paulo, 2023.
123 p.: il.

Trabalho de Conclusão de Curso (Graduação) - Ciência da Computação / Instituto de Matemática e Estatística / Universidade de São Paulo.
Bibliografia

1. CE610.4.1X. 2. CH792.7.6. I. Finger, Marcelo. II. Título.

Bibliotecárias do Serviço de Informação e Biblioteca
Carlos Benjamin de Lyra do IME-USP, responsáveis pela
estrutura de catalogação da publicação de acordo com a AACR2:
Maria Lúcia Ribeiro CRB-8/2766; Stela do Nascimento Madruga CRB 8/7534.

I dedicate this project to my AuDHD for it both compelled me and permitted me to do ninety percent of the work in the last month or so.

Abstract

Gabriela Noriega. **TugaHue:** ***Use of machine learning and transformers to differentiate the national variates of the Portuguese language***. Capstone Project Report (Bachelor). Institute of Mathematics and Statistics, University of São Paulo, São Paulo, 2023.

In this project, two mBERT and two BERTimbau models were fine tuned with two different datasets and evaluated over the DSL-TL dataset in order to assess their applicability for the task of automatic dialect identification with the two primary dialects/variates of concern being Brazilian Portuguese and European Portuguese. The training datasets used were the CETEN and CETEM journalistic *corpora* as well as a dataset comprised of a few international documents and their different official translations made by Brazil and Portugal. The results showed that BERTimbau based models outperfomed their mBERT counterparts by 80.44% vs 78.41% F1 score in the best case and 80.44% vs 57.60% F1 score in the worst case.

Keywords: BERT. dialect identification. Portuguese dialects.

Resumo

Gabriela Noriega. **TugaHue:** ***Uso de aprendizado de máquina e transformers para distinguir as variedades nacionais da língua portuguesa***. Monografia (Bacharelado). Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2023.

Neste TCC, dois modelos mBERT e dois modelos BERTimbau foram afinados (*fine tuned*) com dois datasets diferentes e avaliados sobre o dataset DSL-TL para avaliar sua aplicabilidade para a tarefa de identificação automática de dialetos/variedades da língua portuguesa, sendo as duas variedades de interesse a brasileira e a europeia. Os datasets de treinamento empregados foram os *corpora* jornalísticas CETEN e CETEM bem como um dataset formado por alguns poucos documentos internacionais e suas diferentes traduções oficiais feitas por Brasil e Portugal. Os resultados mostraram que os modelos baseados em BERTimbau se saíram melhores do que os modelos baseados em mBERT com 80.44% vs 78.41% no F1 score no melhor caso e 80.44% vs 57.60% no F1 score no pior caso.

Palavras-chave: BERT. identificação de dialeto. variantes da língua portuguesa.

Resumén

Gabriela Noriega. **TugaHue:** ***Uso de aprendizaje de máquina y transformers para la distinguir las variedades nacionales de la lengua portuguesa***. Monografia (Bacharelado). Instituto de Matemáticas y Estadística, Universidad de São Paulo, São Paulo, 2023.

En ese proyecto, dos modelos mBERT y dos modelos BERTimbau foran afinados (*fine tuned*) con dos datasets distinctos y evaluados sobre el dataset DSL-TL para evaluar su aplicabilidad para la tarea de identificación automática de dialectos/vareidades de la idioma portugués, sendo las dos variedades de interés la brasileña y la europea. Los datasets de treinamiento empleados form los *corpora* perodistas CETEN y CETEM así como un dataset formado por algunos pocos documientos internacionales y sus differentes traduciones oficiales hechas por Brasil y Portugal. Los resultados muestran que los modelos baseados em BERTimbau foram mejores que los baseados em mBERT con 80.44 % vs 78.41 % en el F1 score en el mejor caso y 80.44 % vs 57.60 % en el F1 score en el peor caso.

Palabras-clave: BERT. identificación de dialectos. dialectos del portugués.

List of Acronyms

ANN	Artificial Neural Network
BERT	Bidirectional Encoder Representations from Transformers
IME	USP's Institute of Mathematics and Statistics
JSON	Javascript Object Notation
ML	Machine Learning
SAR	Special Administrative Region
ULMFiT	Universal Language Model Fine-tuning for Text Classification
USP	University of São Paulo

List of Figures

List of Tables

Contents

Chapter 1

Goals and motivation

This project aims at assisting the clean up of the Carolina corpus[3] by using machine learning techniques to identify misplaced texts in said corpus as it is intended not as a general Portuguese language corpus but as a Brazilian Portuguese only corpus.

While the Portuguese language is spoken by more than 265 million people[10] and is an official language of 9+1 countries[1], the three biggest national variates are Brazilian Portuguese (BP), European Portuguese (EP), and Angolan Portuguese (AP).

Of these, only the first two are of primary concern as there doesn't seem to be a lot of public online texts from the Angolan variate.

The differences between these the Brazilian and European Portuguese variates are most notable when it comes to phonetics. The Carolina corpus, however, is only for written texts. While there were some significant spelling differences between the different variates, the Portuguese Language Orthographic Agreement of 1990 (AO90) has greatly reduced such differences. Note that, despite it being signed in 1990, the agreement only came into force in Brazil and in Portugal in 2009. Meanwhile, Angola has refused to ratify the agreement.

This means that any AI model for the automatic discrimination of Portuguese variates must rely very little on spelling a lot more on word choice and order. A further implication here is that any dataset used for training must include texts in multiple variates that verse very closely about the same things so as to prevent a repeat of that US Army tank-detection NN fiasco[2].

[1] The +1 is because Macau is not technically a country but rather a Chinese SAR (Special Administrative Region).

[2] That episode is almost certainly an urban legend but I chose to mention it anyway because I believe it's a great cautionary tale.

Chapter 2

Literature Review and Basic Concepts

2.1 Basic concepts

2.1.1 Bag of words

A bag-of-words is basically a multiset of words that occur in a text. That is, for any given text, its bag-of-words is the collection that contains all the words in said text and that associates them to the number of times they appeared in said text.

Let's make an example with the preamble to the American Declaration of Rights and Duties of Men [8]. Its text is:

> All men are born free and equal, in dignity and in rights, and, being endowed by nature with reason and conscience, they should conduct themselves as brothers one to another.
>
> The fulfillment of duty by each individual is a prerequisite to the rights of all. Rights and duties are interrelated in every social and political activity of man. While rights exalt individual liberty, duties express the dignity of that liberty. Duties of a juridical nature presuppose others of a moral nature which support them in principle and constitute their basis.
>
> Inasmuch as spiritual development is the supreme end of human existence and the highest expression thereof, it is the duty of man to serve that end with all his strength and resources.
>
> Since culture is the highest social and historical expression of that spiritual development, it is the duty of man to preserve, practice and foster culture by every means within his power.
>
> And, since moral conduct constitutes the noblest flowering of culture, it is the duty of every man always to hold it in high respect.

The corresponding bag-of-words, JSON encoded, is:

```
{
   "a": 3,
   "activity": 1,
   "all": 3,
   "always": 1,
   "and": 12,
   "another": 1,
   "are": 2,
   "as": 2,
   "basis": 1,
   "being": 1,
   "born": 1,
   "brothers": 1,
   "by": 3,
   "conduct": 2,
   "conscience": 1,
   "constitute": 1,
   "constitutes": 1,
   "culture": 3,
   "development": 2,
   "dignity": 2,
   "duties": 3,
   "duty": 4,
   "each": 1,
   "end": 2,
   "endowed": 1,
   "equal": 1,
   "every": 3,
   "exalt": 1,
   "existence": 1,
   "express": 1,
   "expression": 2,
   "flowering": 1,
   "foster": 1,
   "free": 1,
   "fulfillment": 1,
   "high": 1,
   "highest": 2,
   "his": 2,
   "historical": 1,
   "hold": 1,
   "human": 1,
   "in": 5,
   "inasmuch": 1,
   "individual": 2,
   "interrelated": 1,
```

```
      "is": 6,
      "it": 4,
      "juridical": 1,
      "liberty": 2,
      "man": 4,
      "means": 1,
      "men": 1,
      "moral": 2,
      "nature": 3,
      "noblest": 1,
      "of": 12,
      "one": 1,
      "others": 1,
      "political": 1,
      "power": 1,
      "practice": 1,
      "prerequisite": 1,
      "preserve": 1,
      "presuppose": 1,
      "principle": 1,
      "reason": 1,
      "resources": 1,
      "respect": 1
      "rights": 4,
      "serve": 1,
      "should": 1,
      "since": 2,
      "social": 2,
      "spiritual": 2,
      "strength": 1,
      "support": 1,
      "supreme": 1,
      "that": 3,
      "the": 10,
      "their": 1,
      "them": 1,
      "themselves": 1,
      "thereof": 1,
      "they": 1,
      "to": 5,
      "which": 1,
      "while": 1,
      "with": 2,
      "within": 1,
   }
```

2.1.2 Transformers, BERT and BERTimbau

Transformers are a class of machine learning models introduced in the 2017 paper "Attention Is All You Need" [11]. They were developed for machine translation and their strength comes from their capacity to model dependencies between distant words/tokens without relying on recurrent layers which are slow and expensive to train.

BERT (Bidirectional Encoder Representations from Transformers) is a class of pre-trained models introduced in 2018 [2]. They are capable of modelling the dependency between tokens in both directions and thus can better handle language tasks. The fact that they are pre-trained is massive as it allows people to just fine-tune the model to the task at hand which is much cheaper and quicker as it involves training just a few layers as opposed to training the whole model from scratch.

BERTimbau is an mBERT-like model for Brazilian Portuguese which was trained using the brWaC (Brazilian Web as Corpus) corpus [9].

2.1.3 Model metrics

In this project, the four main metrics of interest are: the F1 score, accuracy, precision, and recall. They are defined, for binary classification, as follows:

$$\text{Accuracy} = \frac{\text{True positives} + \text{True negatives}}{\text{True positives} + \text{False positives} + \text{True negatives} + \text{False negatives}} \tag{2.1}$$

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{2.2}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \tag{2.3}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

For multiclass and multilabel classification, this project uses "micro averaging". That is, the number of true and false positives and negatives for each class are just summed and then plugged into the formulas.

2.2 Literature Review

The problem of automatically classifying dialects or close languages is a well known one.

In 2014, researchers from the University of Zagreb achieved 98% accuracy at identifying the language of tweets from four very similar south-slavic languages: Bosnian, Croatian, Montenegrin and Serbian [7]. They tested multiple classifiers and found out that MultinomialNB with 320 features (i.e. words) worked best. They also discovered that only 470 words from any given user were necessary to accurately classify the user's language.

In 2019, researchers from the Southeastern Louisiana University made a dialect classifier to distinguish between the Southern US and the New England dialects [6]. Their best accuracy and F-score were, 57.86% and 58.51%, respectively.

Both of these papers used data from tweets and employed a bag-of-words technique.

A very different approach was taken by Demszky et al. to distinguish Indian English dialects in 2020. They used minimal pairs and a pretrained BERT transformer. One of their most surprising conclusions, was that it is often possible to make dialect feature classifiers from as few as ten examples (five with the feature and five without the feature) [1]. Their metrics varied per feature and dataset but their best Macro-AUC was 79%.

In 2022, a group of researchers from three different Chinese universities achieved 96.67%, 97.5%, and 98% accuracy for Portuguese variate classification using BERT, CNNs, and ULMFiT, respectively [4]. They trained and evaluated their models on the PAN 2017 Author Profiling training corpora which is composed of one hundred tweets per author and a thousand authors per language.

In 2023, a group of researchers from the USA, Finland, and the UK created a manually labelled dataset of sentences in European Portuguese, Brazilian Portuguese, British English, American English, Peninsular Spanish, and Argentinian Spanish. They achieved F-scores of 75,6% and 47,7% for Brazilian and European Portuguese respectively on an mBERT pretrained model. [12]

Chapter 3

Methodology

3.1 Datasets

As previously alluded in the introduction, the choice of dataset is a key issue for this project.

Upon searching the web for existing corpora, the most promising result was the Pluricentric Corpus of the Portuguese Language (CPLP Corpus)[5] which includes texts from various countries including the three big ones of our interest. However, despite the cited article's promise of making it available through TEITOK, it seemed completely inexistent online.

The best existing corpora for this project were the CETEMPúblico and the CETENFolha which are from Portugal and Brazil, respectively. Both are journalistic corpora from large newspapers and the latter was specifically designed to be comparable to the former, making them near ideal for this machine learning problem. There is the issue of one being much larger than the other as shown in table 3.1 but this can be easily bypassed through random sampling of an equal number (but not equal proportion) of excerpts from each corpora. This resulting dataset will be referred as the CETEN-CETEM dataset.

Trait	CETEMPúblico	CETENFolha
Text excerpts	1 504 258	34 094
Paragraphs	2 571 735	688 400
Sentences	7 082 094	1 597 807
Words	191 687 833	25 475 272

Table 3.1: *A comparison of basic stats for the CETEMPúblico e CETENFolha corpora.*

Additionally, the author constructed a corpus from international treaties and their official translations into Portuguese which differ between Brazil and Portugal as each country translates them separately. Because the Carolina corpus was found to contain some English language paragraphs in quotations from Portuguese texts, it was decided to include the English, French, and Spanish versions of these treaties into the corpus as it would help to filter out texts not in Portuguese from the Carolina corpus.

The treaties used for this corpus, nicknamed the Coraline corpus, were:

- The International Covenant on Civil and Political Rights (ICCPR) (1966).
- The Comprehensive Nuclear-Test-Ban Treaty (CTBT) (1996).
- The Paris Agreement (2015)

The total amount of text is small indeed, however, the fact that this is a nearly parallel corpus means that any model trained on it will have to pickup on small linguistic features as opposed to relying on other things such as topics or place names.

3.2 Preprocessing

3.2.1 CETEN-CETEM

To make the corpora more balanced, a quarter of a million paragraphs were randomly chosen from each dataset for training and evaluating the models.

3.2.2 Carolina

For evaluation over the Carolina corpus, each except was divided into paragraphs according to the `<p>` tag in the XML source files. Those paragraphs were then stripped of leading whitespace and deduplicated.

Finally, the dataset was randomly sampled so as to reduce it to 1/40th of its size before being ran on all four trained previously models.

This reduction was done to save GPU processing time.

3.3 AI Models

For each of the two final datasets (Coraline and CETEN-CETEM), two very similar models were trained, one using mBERT[1] and another using BERTimbau[2].

The models trained on CETEN-CETEM were simple binary classifiers as the training dataset had only two possible values: Brazilian or European.

Meanwhile those trained on Coraline were designed as multi-label classifiers with six labels: English, French, Spanish, Portuguese, Brazilian Portuguese, European Portuguese. In practice, all Portuguese language samples had two labels, one for Portuguese and another for the specific variate of Portuguese. This unusual splitting was chosen in the

[1] Model page on Hugging Face: https://huggingface.co/bert-base-multilingual-cased.

[2] Model page on Hugging Face: https://huggingface.co/neuralmind/bert-base-portuguese-cased.

hopes it could be used to detect variates of Portuguese that were neither Brazilian nor European.

Nearly all work was done on Google's Colab with the Hugging Face Python library running on instances with a T4 GPU.

Model Name	Number of labels	Pretrained base	Fine tuned over	Trained over
6L-BERTimbau	6	BERTimbau	Coraline	5 epochs
6L-BERT	6	mBERT	Coraline	5 epochs
2L-BERTimbau	2	BERTimbau	CETEN-CETEM	2 epochs
2L-BERT	2	mBERT	CETEN-CETEM	2 epochs

Table 3.2: *Summary of the models trained during this project.*

Chapter 4

Results

4.1 Training

The training loss curves (Figures 4.1 and 4.3) show what one would expect for this kind of ML problem.

When it comes to accuracy and F1 scores, BERTimbau was better than BERT throughout the training as depicted in figures 4.2 and 4.4.

However, the matrix confusion in figure 4.5a appears to show that BERT was better than BERTimbau at distinguishing Brazilian and European Portuguese. This is, however, a by-product of the conversion from float scores to boolean labels. When the threshold is lowered from 0.9 (subfigure 4.5a) to 0.7 (subfigure 4.5b), we see that BERTimbau was better than BERT.

4.2 Performance on other datasets

When tested against the DSL-TL dataset (see figure 4.6 and table 4.1), 6L-BERTimbau performed well although it was more likely to misclassify a Brazilian text as being of the European variate than the reverse. On the other hand, 6L-BERT preformed so poorly that it had more false positives than true positives for European Portuguese classification.

One possible explanation for that aforementioned 6L-BERTimbau's misclassification tendency is that it is a by-product of the official Portuguese grammar being based more on the European variate. Thus, whenever Brazilian authors want to "follow the rules", they end up sounding more like their Portuguese counterparts than their fellow countryfolk.

The best model from each training dataset were also compared on a random sample of the Carolina corpus and they outputted roughly similar results (see figure 4.8);

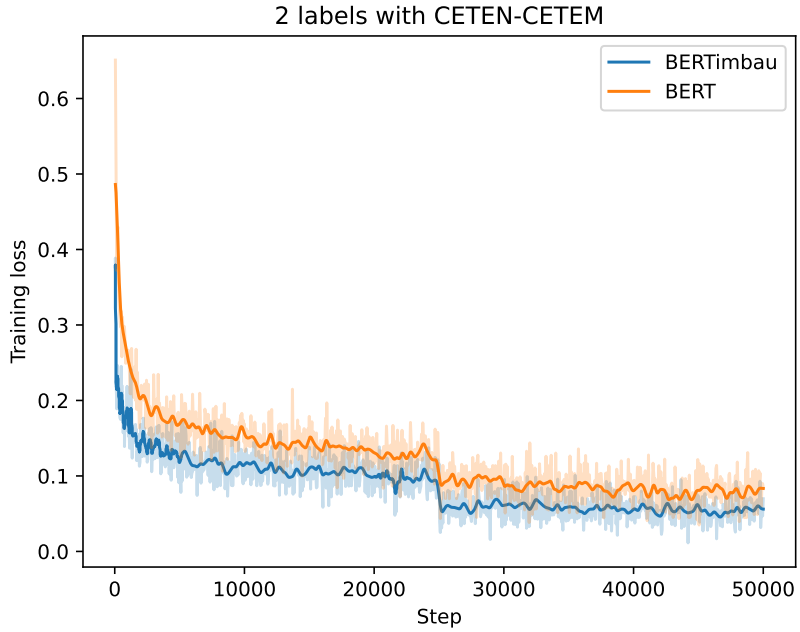


Figure 4.1: *Training loss of 2L-BERTimbau and 2L-BERT.*

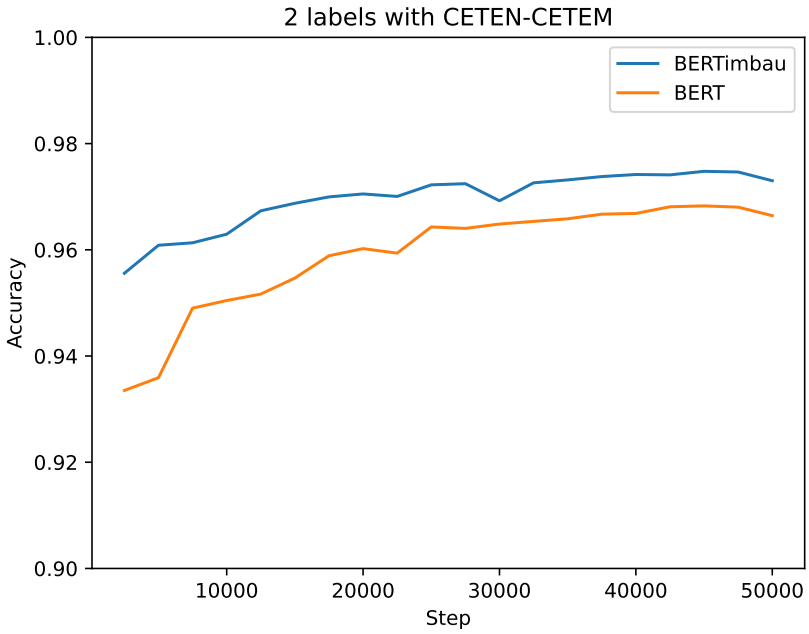


Figure 4.2: *Accuracy of 2L-BERTimbau and 2L-BERT.*

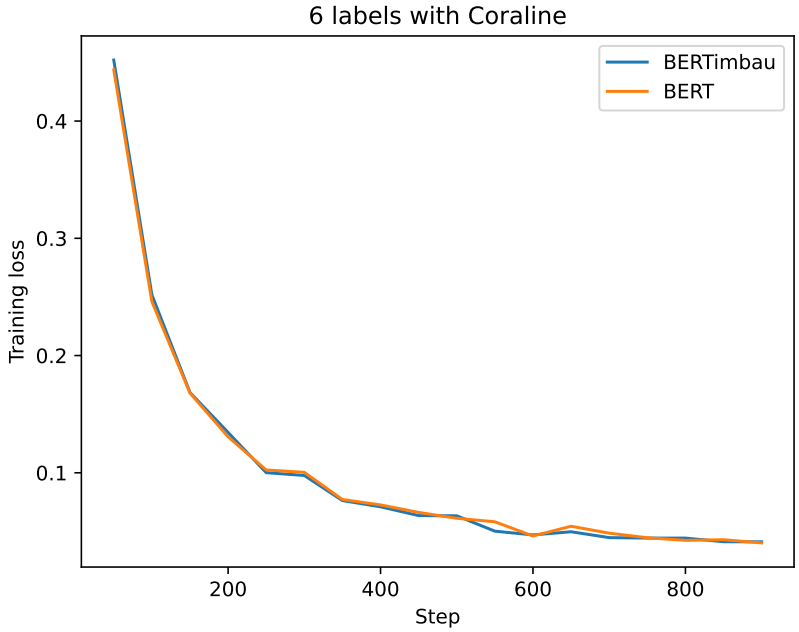


Figure 4.3: *Training loss for 6L-BERTimbau and 6L-BERT.*

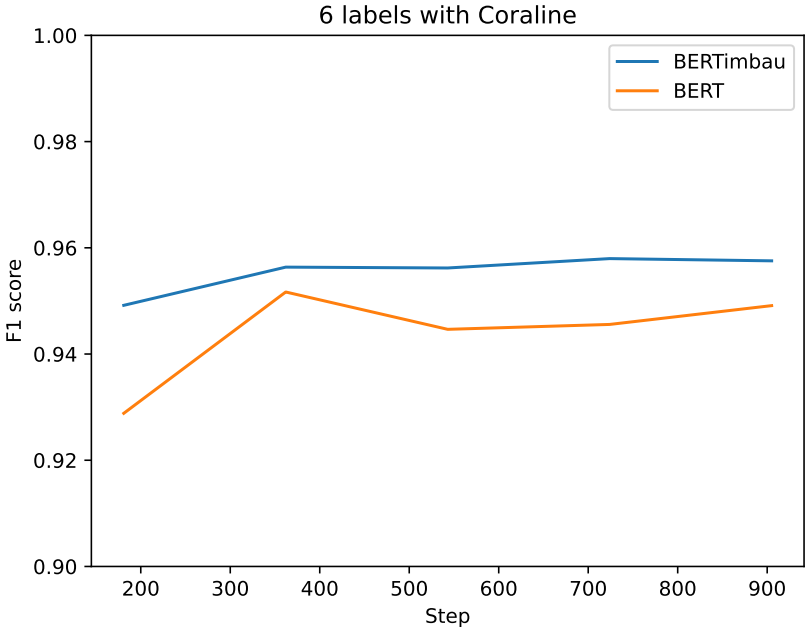
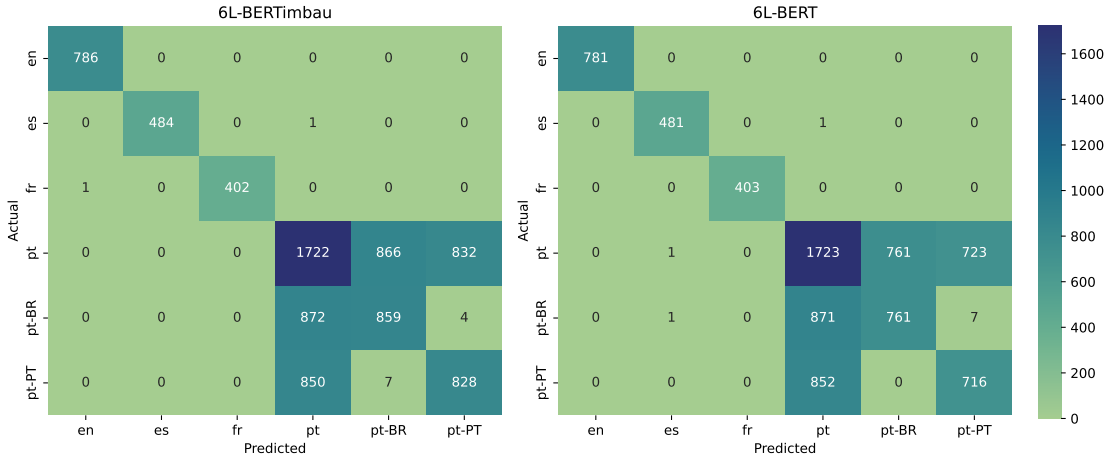
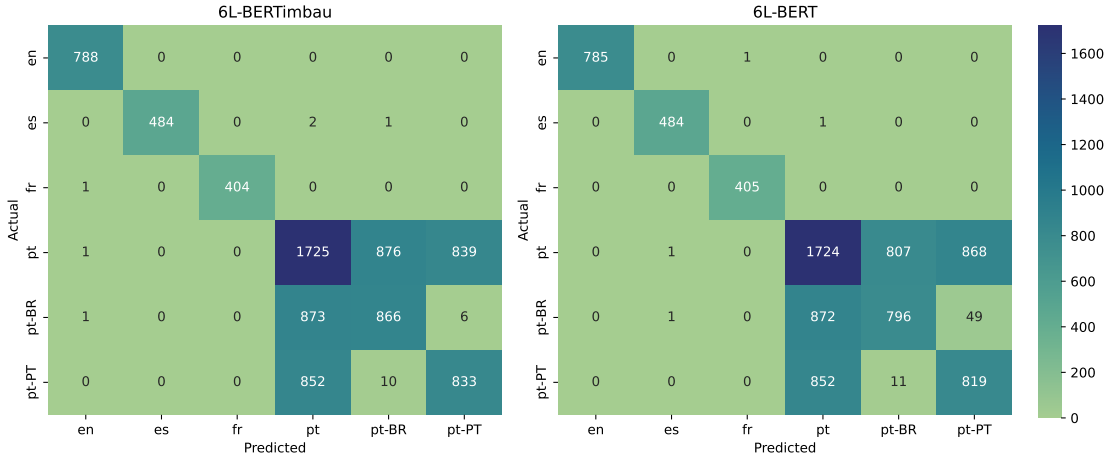


Figure 4.4: *F1 score for 6L-BERTimbau and 6L-BERT.*



(a) *Using a threshold of 0.9 to convert from float to boolean.*



(b) *Using a threshold of 0.7 to convert from float to boolean.*

Figure 4.5: *Confusion matrix for 6L-BERTimbau and 6L-BERT.*

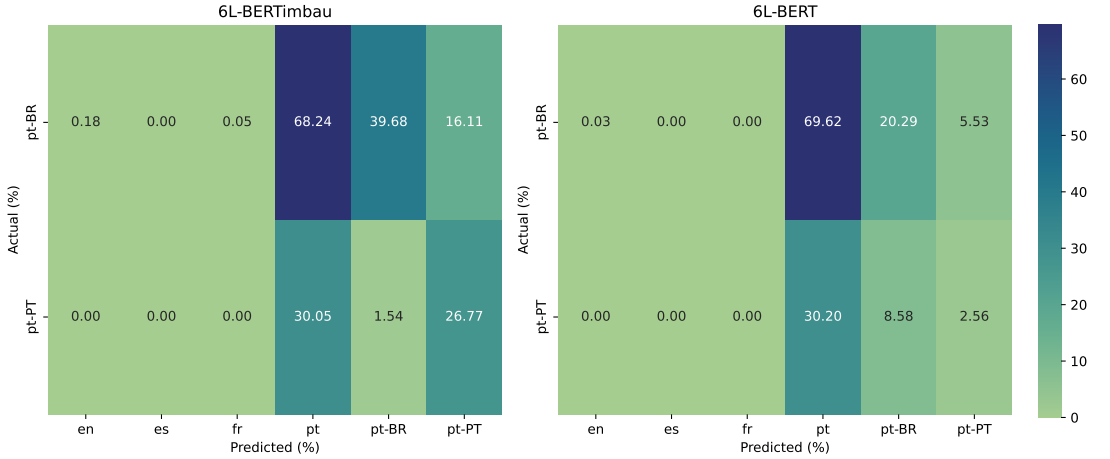


Figure 4.6: *Confusion matrix for 6L-BERTimbau and 6L-BERT when tested against the DSL-TL dataset.*

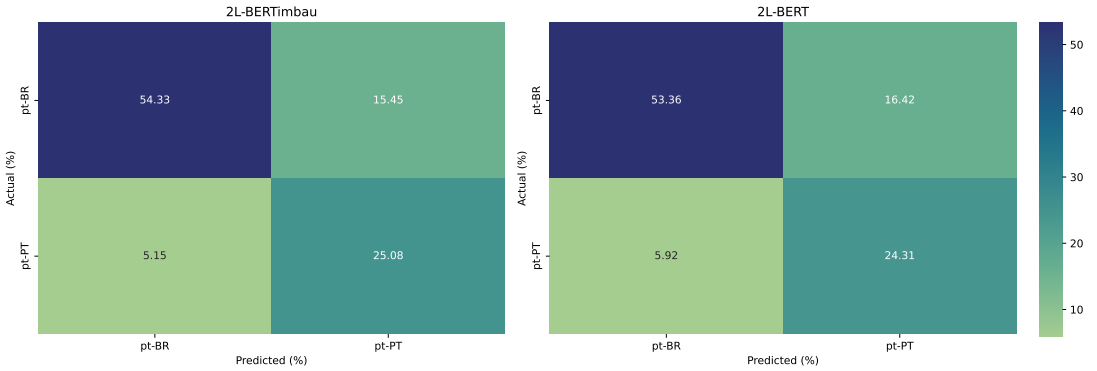


Figure 4.7: *Confusion matrix for 2L-BERTimbau and 2L-BERT when tested against the DSL-TL dataset.*

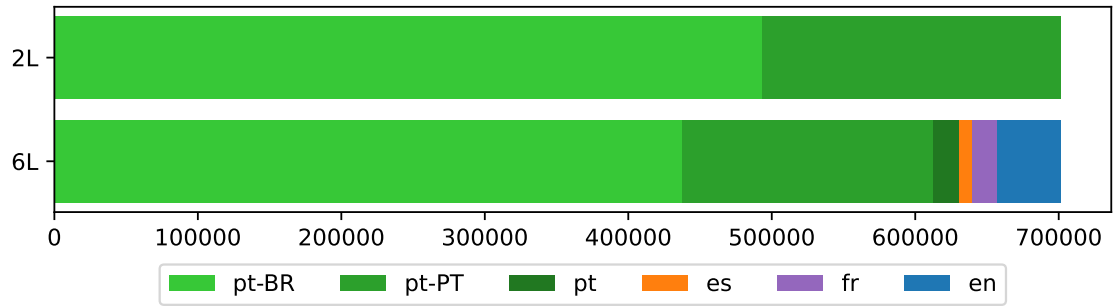


Figure 4.8: *Classification of Carolina corpus sample by 2L-BERTimbau (upper bar) and 6L-BERTimbau (lower bar).*

Model name	F1	Accuracy	Precision	Recall
2L-BERTimbau	80.08%	79.41%	82.44%	79.41%
2L-BERT	78.41%	77.66%	80.85%	77.66%
6L-BERTimbau	80.44%	79.41%	84.79%	79.41%
6L-BERT	57.60%	56.48%	59.19%	56.48%

Table 4.1: *Summary of the metrics of the models when evaluated over the DSL-TL dataset using weighted averaging.*

Chapter 5

Conclusions

The results obtained throughout this project support the conclusions that BERTimbau is a better fit than mBERT for the task of Portuguese language dialect classification.

Additionally, it was shown that small datasets can be effective when they contain nearly parallel texts. That is, different texts (in this project's case translations) that verse about the same topic in the same structure and style.

References

[1] Dorottya Demszky et al. “Learning to Recognize Dialect Features”. In: *CoRR* abs/2010.12707 (2020). arXiv: 2010.12707. URL: https://arxiv.org/abs/2010.12707 (cit. on p. 7).

[2] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (cit. on p. 6).

[3] Marcelo Finger et al. *Carolina: The Open Corpus for Linguistics and Artificial Intelligence*. https://sites.usp.br/corpuscarolina/corpus. Version 1.2 (Ada). 2022 (cit. on p. 1).

[4] Sameeah Noreen Hameed, Muhammad Adnan Ashraf, and Yanan Qiao. “Multi-Lingual Language Variety Identification using Conventional Deep Learning and Transfer Learning Approaches”. In: *Int. Arab J. Inf. Technol.* 19 (2022), pp. 705–712. URL: https://api.semanticscholar.org/CorpusID:251606491 (cit. on p. 7).

[5] Maarten Janssen et al. “The CPLP Corpus: A pluricentric corpus for the common Portuguese spelling dictionary (VOC)”. In: *Proceedings of the 28º EURALEX International Congress*. 2018 (cit. on p. 9).

[6] Aran Komatsuzaki. *One Epoch Is All You Need*. 2019. arXiv: 1906.06669 [cs.LG] (cit. on p. 7).

[7] Nikola Ljubešić and Denis Kranjčić. “Discriminating between closely related languages on twitter”. In: *Informatica* 39.1 (2015) (cit. on p. 7).

[8] OAS – Organisation of American States. *American Declaration of Rights and Duties of Men*. 1948. URL: https://www.oas.org/dil/access_to_information_human_right_American_Declaration_of_the_Rights_and_Duties_of_Man.pdf (cit. on p. 3).

[9] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. “BERTimbau: pretrained BERT models for Brazilian Portuguese”. In: *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*. 2020 (cit. on p. 6).

[10] UNESCO. *207 EX/43 - World Portuguese Language Day*. Sept. 13, 2019. URL: https://en.unesco.org/sites/default/files/accord_unesco_langue_portuguaise_conference_generale_eng.pdf (cit. on p. 1).

[11] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017). arXiv: 1706.03762 (cit. on p. 6).

[12] Marcos Zampieri et al. *Language Variety Identification with True Labels*. 2023. arXiv: 2303.01490 [cs.CL] (cit. on p. 7).