Enhancing Portuguese Variety Identification with Cross-Domain Approaches

Hugo Sousa[1,3*], **Rúben Almeida**[2,3,4*], **Purificação Silvano**[5,6], **Inês Cantante**[5,6], **Ricardo Campos**[3,7,8], **Alípio Jorge**[1,3]

[1]Faculty of Sciences, University of Porto, Porto, Portugal
[2]Faculty of Engineering, University of Porto, Porto, Portugal
[3]INESC TEC, Porto, Portugal
[4]Innovation Point - dst group, Braga, Portugal
[5]Faculty of Arts and Humanities, University of Porto, Porto, Portugal
[6]Centre of Linguistics, University of Porto, Porto, Portugal
[7]Department of Informatics, University of Beira Interior, Covilhã, Portugal
[8]Ci2 - Smart Cities Research Center, Tomar, Portugal
hugo.o.sousa@inesctec.pt, ruben.f.almeida@inesctec.pt

Abstract

Recent advances in natural language processing have raised expectations for generative models to produce coherent text across diverse language varieties. In the particular case of the Portuguese language, the predominance of Brazilian Portuguese corpora online introduces linguistic biases in these models, limiting their applicability outside of Brazil. To address this gap and promote the creation of European Portuguese resources, we developed a cross-domain language variety identifier (LVI) to discriminate between European and Brazilian Portuguese. Motivated by the findings of our literature review, we compiled the PtBrVarId corpus, a cross-domain LVI dataset, and study the effectiveness of transformer-based LVI classifiers for cross-domain scenarios. Although this research focuses on two Portuguese varieties, our contribution can be extended to other varieties and languages. We open source the code, corpus, and models to foster further research in this task.

1 Introduction

Discriminating between varieties of a given language is an important natural language processing (NLP) task (Joshi et al. 2024). Over time, populations that share a common language can evolve distinctive speech traits due to geographical and cultural factors, including migration and the influence of other languages (Raposo, Vicente, and Veloso 2021). Recently, this importance became even more pronounced with the advent of variety-specific large language models, where variety discrimination plays a pivotal role (Rodrigues et al. 2023). Whether in the pre-training, fine-tuning, or evaluation phase, having a highly effective system to discriminate between varieties reduces the amount of human supervision required, accelerating the production of curated mono-variety datasets (Öhman et al. 2023). However, developing such a system presents considerable challenges. Classifiers often struggle to identify linguistically relevant features, showing a tendency to be biased towards non-linguistic factors, such as named entities and thematic content (Diwersy, Evert, and Neumann 2014). Consequently, these classifiers exhibit limited transfer capabilities to domains not represented in the training set, significantly restricting their utility in multi-domain applications (Lui and Baldwin 2011; Nguyen et al. 2021).

A language in which variety identification is particularly challenging is Portuguese. It is spoken by more than 200 million people worldwide and serves as the official language of eight nations on five continents, each with its own variety. However, 70% of Portuguese speakers are Brazilian citizens[1], which implies that resources labeled as Portuguese are dominated by this language variety. Another important characteristic of Portuguese is that, unlike languages where differences are predominantly phonological, such as those of the North Germanic family (Holmberg and Platzack 2008), the widespread of Portuguese has fostered considerable phonological, morphological, lexical, syntactic and semantic variations between Portuguese varieties (Brito and Lopes 2016). In the development of language models, for example, this variety divergence has practical implications; models trained in Brazilian Portuguese generate texts that are markedly distinct from those trained in other Portuguese varieties (Rodrigues et al. 2023). This fact restrains the adoption of these models outside of Brazil in domains where formal non-Brazilian text is required, as is the case of legal and medical applications. This underscores the practical importance of developing effective LVI systems that can be deployed in production.

In this study, we describe the development of a cross-domain LVI classifier that discriminates between Brazilian and European Portuguese. To accomplish that, we start with a comprehensive listing of Portuguese LVI resources. The lack of multi-domain corpora motivated us to compile our own dataset. This corpus was then used in the development of our LVI classifier. For the training procedure we devised a training protocol that takes into account the cross-domain

*These authors contributed equally.

[1]Statistic inferred form Wikipedia https://en.wikipedia.org/wiki/Portuguese_language#Lusophone_countries

capabilities of models during evaluation. Furthermore, we also study the impact of masking named entities and thematic content embedded in the training corpus, a process named delexicalization (Lui et al. 2014). To summarize, the contributions of this work are the following:

1. We introduce a novel cross-domain, silver-labeled LVI corpus for Brazilian and European Portuguese, compiled from open-license datasets;
2. We examine the impact of different levels of delexicalization on the overall effectiveness of LVI models;
3. We propose a training protocol for LVI models that yields better generalization performance;
4. We release the first open-source Portuguese LVI model, providing a valuable resource for future research and practical applications.

The remainder of this paper is organized as follows: Section 2 offers a comprehensive literature review on the state-of-the-art in Portuguese LVI. In Section 3, we introduce our compiled dataset, PtBrVarId, and present relevant statistics along with the results of a manual evaluation of the quality of the dataset. Section 4 describes the training protocol and the models developed, including the baselines and benchmarks used for comparison. The results are presented in Section 6, followed by a discussion of future research directions in Section 7.

2 Related Work

Corpora

Despite the numerous works developed in the LVI task, the first gold-labeled dataset that includes Portuguese corpora, the DSL-TL corpus (Zampieri et al. 2024), was only introduced in 2023. This dataset used crowdsourcing to annotate approximately 5k Portuguese documents. The corpus are not only labeled as "European" and "Brazilian" Portuguese, but also a special "Both or Neither" label to signal those documents with insufficient linguistic marks to be considered part of one of these varieties.

Prior to the release of this dataset, the evaluation process was often performed in silver-labeled data, collected using domain-specific heuristics. For instance, in the journalistic domain, it is common to assume the language variety of a document based on the newspaper origin; Brazilian newspapers' articles are assigned a Brazilian Portuguese label, while Portuguese ones are assigned a European Portuguese label (Silva and Lopes 2006; Zampieri and Gebre 2012; Tan et al. 2014). In the social media domain, a similar approach is frequently used. Castro, Souza, and Oliveira (2016) used geographic metadata collected on Twitter to assign a language variety to each document based on the authors location. Unfortunately, many of these Portuguese LVI resources are no longer available online. This limitation motivated us to collect and open-source our training data.

Modeling Approaches

The high effectiveness of N-gram-based systems observed in language identification studies (McNamee 2005; Martins and Silva 2005; Chew et al. 2009), a task closely related to LVI, motivated the application of these methods in the context of LVI. To this day, this approaches are still employed, with several submissions to the VarDial workshop[2] – which compiles most of the recent studies in the LVI task – achieving high effectiveness. Notable examples include Italian with an F_1 score of 0.90 (Jauhiainen, Jauhiainen, and Lindén 2022), Uralic with an F_1 score of 0.94 (Bernier-Colborne, Leger, and Goutte 2021), and Mandarin with an F_1 score of 0.91 (Yang and Xiang 2019).

The adoption of transformer-based techniques (Vaswani et al. 2017) in LVI has not been as rapid as in other NLP tasks. Recently, some studies have leveraged monolingual BERT-based models to fine-tune LVI classifiers for Romanian (Zaharia et al. 2020) and French (Bernier-Colborne, Leger, and Goutte 2022). However, in none of these cases were transformers capable of outperforming N-gram-based techniques, only achieving a F_1 score of 0.65 in Romanian and 0.43 in French. Similar results have been reported for different languages using other deep learning techniques, such as multilingual transformers (Popa and Stefănescu 2020), feedforward neural networks (Çöltekin and Rama 2016; Medvedeva, Kroon, and Plank 2017), and recurrent networks (Guggilla 2016; Çöltekin, Rama, and Blaschke 2018).

In the specific case of Portuguese, older studies have relied on N-gram-based techniques to achieve results above 90% accuracy on silver-labeled benchmarks (Silva and Lopes 2006; Zampieri and Gebre 2012; Goutte, Léger, and Carpuat 2014; Malmasi and Dras 2015; Castro, Souza, and Oliveira 2016). However, it has been noted that evaluating on silver-labeled corpora is reliability (Zampieri and Gebre 2014), and preliminary results obtained on the gold-labeled DSL-TL corpus (Zampieri et al. 2024) revealed more modest performance, with F_1 scores below 70%. Additionally, contrary to observations in silver-labeled evaluations (Medvedeva, Kroon, and Plank 2017), the current state-of-the-art result for Portuguese LVI on the DSL-TL benchmark (0.79 F_1 score) comes from fine-tuning a collection of BERT-based models (Vaidya and Kane 2023).

Cross-Domain Capabilities

Lui and Baldwin (2011) revealed that N-grams based techniques had limited cross-domain capabilities for the language identification task. Despite the good results of N-gram-based models when the train and test domain overlap (above 85% accuracy), the results also show that the effectiveness decreased as much as 40% when both sets do not match. In order to address this phenomenon, the authors have devised a feature selection mechanism that later opened the door to the development of the first cross-domain language identification tool, the `langid.py` (Lui and Baldwin 2012).

In the context of French LVI, Diwersy, Evert, and Neumann (2014) used unsupervised learning to demonstrate that, despite the good results reported by N-grams based-methods (above 95% accuracy), the feature learned by these models reveal no interest from a linguistic point of view. In-

[2]https://aclanthology.org/venues/vardial/

stead, classifiers relied on named entities, polarity and thematics embedded in the training corpus to support its inference process (Ex: If "Cameroun" was mentioned in the document, the model assigned a French-Cameroonian label to it).

In light of these facts, the mass adoption of these architectures in the context of LVI, creates urgency for finding solutions to surpass this limitation. In this study, we extend the knowledge about the cross-domain capabilities of N-gram-based models, while presenting the first results for transformer architectures.

3 PtBrVarId Dataset

In this section we introduce the PtBrVarId, the first silver-labeled multi-domain Portuguese LVI corpus. This resource resulted from the compilation of open-license corpora from 11 European (EP) and Brazilian (BP) sources over six domains, namely: **Journalistic**, **Legal**, **Politics**, **Web**, **Social Media** and **Literature**. The following sections describe how the dataset was created.

Corpora Compiled

Training machine learning and deep learning models requires a robust and well-labeled training corpus. However, manually labeling such a corpus is often laborious, time-consuming and expensive. To address this challenge in our research, we opted for a silver labeling approach.

In the context of the VID task, silver labeling involves identifying texts where the variety can be inferred with a reasonable degree of confidence based on the documents metadata. In the following paragraphs we describe the data sources used in each textual domain along with the heuristics that supported the silver-labelling step. It is important to note that we were careful to only use sources that were permissive for academic research.

Journalistic As a source of news corpus we use two resources available at Linguateca (Santos 2014), namely: CETEMPublico (Rocha and Santos 2000) and CETEMFolha. The CETEMPublico corpus contains news articles from the Portuguese newspaper "Público" while the CETEMFolha contains news from the Brazilian newspaper "Folha de São Paulo". The geographic location of the newspaper is used to label the Portuguese variety.

Literature The literature domain relies on three data sources that index classics of Portuguese literature: the Gutenberg project[3]; the LT-Corpus (Généreux, Hendrickx, and Mendes 2012); and the Brazilian literature corpus[4]. The author's nationality was used to label the documents as European or BP.

Legal The Brazilian split from the legal corpora was compiled from RulingBR (de Vargas Feijó and Moreira 2018) which contains decisions from the Brazilian supreme court ("Supremo Tribunal Federal") between 2011 to 2018. The European split was built from the DGSI website[5] which provides access to a set of databases of precedents and to the bibliographic reference libraries of the Portuguese Ministry of Justice.

Politics For the politics domain we used the manual transcriptions of political speeches in both the European Parliament (Koehn 2005) and the Brazilian Senate (Cezar 2020). The document's origin was used to infer the label for the Portuguese variety.

Web For the web domain, corpora were extracted from OSCAR (Suarez, Sagot, and Romary 2019). To define the labels, we began by identifying domains ending in `.pt` or `.br`. From this list, we manually curated a set of the 50 most frequent domains ending in `.pt` and 50 domains ending in `.br`. The documents from OSCAR associated with these curated domains were then used in our corpus.

Social Media The social media corpora derives from three data sources. For BP we used the Hate-BR (Vargas et al. 2022) dataset, which was manually annotated for train hate speech classifiers, and a compilation of fake news spread in Brazilian WhatsApp groups (da Cunha 2021). Regarding EP, the tweets collected by Ramalho (2021) were filtered based on the tweets' metadata location. Tweets whose location is not part of Wikipedia's list of Portuguese cities[6], were discarded.

Despite the dataset proposed being silver-labeled, some of their components are extracted from high-quality manually annotated corpora that offer sufficient guarantees of belonging to a single language variety. For example, the Europarl corpus (Koehn 2005), is composed of manual transcriptions in EP of political speeches made in European Parliament, therefore it is very unlikely to find any marks of BP in such corpus.

Data Cleaning

To reduce noise in the corpus, we implemented a dedicated data cleaning pipeline. The process starts with basic operations to remove null, empty, and duplicate entries. We then employ the `clean-text` tool[7] to correct Unicode errors and standardize the text to ASCII format. For the Web domain, an additional step is taken using the `jusText` Python package[8] to filter out irrelevant sentences and remove boilerplate HTML code. Finally, outliers within each domain are identified and removed based on the interquartile range (IQR) of token counts, calculated using the `nltk` word tokenizer for Portuguese[9]. Texts falling below the first quartile minus 1.5 times the IQR, or above the third quartile plus 1.5 times the IQR, are discarded. This approach effectively eliminates documents that are either too short or too long for their respective domains.

Table 1 presents the statistics for the corpus obtained after applying the filtering pipeline. The final corpus comprises

[3]https://www.gutenberg.org/browse/languages/pt\#a4827

[4]https://www.kaggle.com/datasets/rtatman/brazilian-portuguese-literature-corpus

[5]https://www.dgsi.pt

[6]https://en.wikipedia.org/wiki/List_of_towns_in_Portugal

[7]https://github.com/jfilter/clean-text

[8]https://github.com/miso-belica/jusText

[9]https://www.nltk.org/api/nltk.tokenize.word_tokenize.html

	Label	Tokens					Docs
		Min	Max	Avg	Std	Count	Count
Journalistic	🇵🇹	16	475	131.29	61.45	189,506,320	1,443,422
	🇧🇷	18	560	81.09	39.11	27,077,538	333,903
Literature	🇵🇹	16	186	77.20	37.39	1,859,660	24,090
	🇧🇷	17	185	72.55	36.19	3,805,896	52,458
Legal	🇵🇹	16	139	51.63	24.43	152,717,737	2,957,980
	🇧🇷	20	124	47.53	22.11	221,167	4,653
Politics	🇵🇹	20	798	258.32	173.39	7,203,739	27,887
	🇧🇷	21	796	276.97	177.60	1,012,586	3,656
Web	🇵🇹	22	2042	517.96	414.72	22,598,587	43,630
	🇧🇷	15	2075	539.66	463.16	23,913,771	44,313
Social Media	🇵🇹	3	646	18.94	9.85	44,758,304	2,363,261
	🇧🇷	6	51	17.11	10.17	94,177	5,504

Table 1: Summary statistics of the PtBrVarId corpus, including the minimum, maximum, average, standard deviation, and count of tokens, as well as the number of documents for each domain and label.

7,304,438 documents, predominantly from the EP segments of the Journalistic, Legal, and Social Media domains. Regarding the number of tokens, we observe that, with the exception of the Journalistic domain, the distribution between documents labeled as EP and BP within each domain is similar.

A comparison across the domains reveals that the Web domain contains the highest average number of tokens per document, whereas the Social Media domain has the lowest, averaging around 18 tokens per document. This disparity is significant for the development of variety identification models, as distinguishing between language varieties in shorter texts is more challenging due to the limited linguistic cues available. Therefore, the Social Media domain is expected to pose more difficulties than the Web domain, where longer texts provide more opportunities to identify distinguishing features of EP and BP.

It is also important to note that the dataset is highly unbalanced across all domains except the Web domain. This imbalance should be carefully considered when training models using this dataset to ensure robust and unbiased effectiveness.

Quality Assurance

To ensure the quality of the silver-labeling process, we asked three linguists to manually annotate 300 documents, focusing on two key aspects:

Variety The linguists were asked to determine the variety of the text. They had three options: EP, BP, or "Undetermined" for cases where no variety-specific linguistic features were available.

Domain The linguists were also tasked with identifying the domain to which each sentence belonged. They could choose from the six domains used in this research, or select "Undetermined" if the domain could not be clearly identified.

For the sampling process, we randomly selected 50 documents from each domain in our corpus, with an equal split of 25 documents silver-labeled as EP and 25 as BP.

Table 2 presents the agreement between the three annotators using three metrics:

Fleiss's Kappa (Fleiss 1971): Measures the agreement between annotators beyond chance, with values ranging from 0 (no agreement) to 1 (perfect agreement).

Majority Rate: Indicates the percentage of texts where two out of three annotators agree on an annotation.

Accuracy: Assesses how often the majority vote between annotators matches the automatic annotation. It is important to remark that the cases where the labeled agreed by the annotators is "Undetermined" we count both silver-labels (EP and BP) as correct since the text is in fact valid in both varieties.

	Metric	Result
Variety	Fleiss' Kappa	0.57
	Majority Rate	0.95
	Accuracy	0.86
Domain	Fleiss' Kappa	0.69
	Majority Rate	0.94
	Accuracy	0.76

Table 2: Agreement among the three annotators regarding language variety and textual domain.

The results obtained show that the agreement is higher for the textual domain aspect than for the language variety. However, the variety aspect still achieves a Fleiss' Kappa of 57%, which, for three annotators with three labels, can be considered moderate agreement. Upon closer inspection of the results, we found that the Fleiss' Kappa is lower in the Literature, Social Media, and Legal domains (see Table 3). For the Social Media domain, we found the disagreement to be mainly driven by the short length of the texts, with "Undetermined" representing 42% of the labels the annotators agreed on. The same was found for the legal domain, which has the second lowest average tokens per document, where the "Undetermined" represents 34% of the labels the annotators agreed on. In the Literature domain, the disagreement is mainly attributed to the corpus consisting of contemporary books, which often blend linguistic features from both European and BP, making it difficult to assign a definitive variety label.

In Table 3 we detail the annotation agreement metrics per domain for the manually label subset of the PtBrVId corpus. The table shows statistics for the Fleiss' Kappa with all the labels and the Fleiss' Kappa when the entries for which one of the annotators marked the entry as "Undetermined". To complete the table we also show the percentage of entries for which at least one annotator labeled as "Undetermined".

Nevertheless, a majority consensus among the annotators is almost always achievable (over 90% of the times) in both aspects. Furthermore, this majority is strongly aligned with the automatic annotations, with agreement between the annotators and the silver labels exceeding 75%.

In addition to releasing the full annotations provided by each annotator, the documents for which a majority vote could be determined are included in the *test* partition of our dataset. For documents labeled as "Undetermined" by the annotators, the original silver label was used as the final label. The complete dataset is publicly accessible on HuggingFace[10]. This dataset offers an opportunity for an in-depth study of the cross-domain capabilities of various LVI techniques, with a particular focus on the application of pretrained transformers, which is the main focus of this paper.

[10]https://huggingface.co/datasets/liaad/PtBrVId

Domain	Metric	Result
Literature	Fleiss' Kappa	0.23
	Fleiss' Kappa$_{wo/u}$	0.51
	Undetermined Rate	0.36
Legal	Fleiss' Kappa	0.46
	Fleiss' Kappa$_{wo/u}$	0.73
	Undetermined Rate	0.34
Politics	Fleiss' Kappa	0.78
	Fleiss' Kappa$_{wo/u}$	0.87
	Undetermined Rate	0.10
Web	Fleiss' Kappa	0.67
	Fleiss' Kappa$_{wo/u}$	0.84
	Undetermined Rate	0.20
Social Media	Fleiss' Kappa	0.53
	Fleiss' Kappa$_{wo/u}$	0.94
	Undetermined Rate	0.42
Journalistic	Fleiss' Kappa	0.72
	Fleiss' Kappa$_{wo/u}$	0.90
	Undetermined Rate	0.04

Table 3: Extended per-domain analysis of annotator agreement. We present Fleiss' Kappa for all three labels, as well as Fleiss' Kappa excluding the "Undetermined" documents (Fleiss' Kappa$_{wo/u}$). The "Undetermined Rate" rows shows the percentage of documents for which at least one annotator labeled as "Undetermined".

4 Experimental Setup

In this study, we investigate the effectiveness of fine-tuning a transformer-based model for the Portuguese LVI task. We employ an iterative methodology to identify the optimal strategy for combining training corpora from various domains into a unified training process. Our primary objective is to evaluate cross-domain effectiveness and the generalization capabilities of our models.

Models & Baselines

For the transformer-based model, we use BERTimbau with 334 million parameters (Souza, Nogueira, and Lotufo 2020). BERTimbau is the result from fine-tuning the original BERT model (Devlin et al. 2019) on a Portuguese corpus.

To establish a baseline for comparison with the BERT model, we employ N-grams combined with Naive Bayes classifiers. This choice is motivated by the proven effectiveness of such models in previous LVI studies across various Indo-European languages, including Portuguese (Zampieri and Gebre 2012).

Cross-Domain Training Protocol

To ensure that our model generalizes effectively across different domains, we define a two-step training protocol. Step one is used to find the best hyperparameters to train the model so ensure the generalization capability of the model. In this step, the model is trained on a single domain from the PtBrVid corpus and validated on the remaining domains (excluding the one used for training). The hyperparameters yielding the best performance in this cross-domain validation are then used in step two to train the model across all domains combined.

Delexicalization of the corpus is treated as a hyperparameter in our approach. We adjust the probabilities of replacing tokens found by Named Entity Recognition (NER) and Part-of-Speech (POS) tagging with the generic label (such as `LOCATION` or `NOUN`), varying these probabilities incrementally from 0% to 100% in 20% steps. It is important to note that delexicalization is applied exclusively to the training set. The validation set remains unaltered, simulating a real-world scenario where the input text is not modified. We leave the study of the impact of delexicalizing the validation set on the effectiveness of the model for future research.

Train & Validation Data

As referred above the PtBrVId dataset is used to train the models. However, before using for the training, we leave 1,000 documents of each domain for the validation of the model, 500 of each label.

In the step one of our training protocol, we use 8,000 documents from each domain (4,000 from each label) to train the models. We found this sample size to be enough for the models to converge and ensure fast iteration in the training process.

For step two of our training protocol, we compile all the documents from the PtBrVid corpus including the ones used for validation in step one. To avoid the training being dominated by the more represented domains, we undersample the dataset so that all labels from all domains are equally represented. At this step, the manually annotated set from PtBrVId set is used to keep track of the generalization loss.

Benchmarks

In our evaluation, we use two benchmarks: the DSL-TL and FRMT datasets. As mentioned above, the DSL-TL dataset is the standard benchmark for distinguishing between EP and BP, annotated with three labels: "EP", "BP", and "Both". For our purposes, we exclude documents labeled "Both" since our training corpus does not contain that label. This results in a test set comprising 588 documents for BP and 269 for EP. The FRMT dataset (Riley et al. 2023) has been manually annotated to evaluate variety-specific translation systems and includes translations in both EP and BP. We adapt this corpus for the VID task, resulting in a dataset containing 5,226 documents, with 2,614 labeled as EP and 2,612 as BP.

5 Implementation Details

NER and POS tags were identified using spaCy[11]. The BERT model was trained with the `transformers`[12] and `pytorch`[13] libraries, for a maximum of 30 epochs, using early stopping with a patience of three epochs, binary cross-entropy loss, and the AdamW optimizer. The learning rate was set to 2×10^{-5}. In addition, a learning rate scheduler was used to reduce the learning rate by a factor of 0.1 if the training loss did not improve for two consecutive epochs. N-gram models were trained using the `scikit-learn`[14] library. The following hyperparameters were taken into account in the grid search we performed"

- **TF-IDF Max Features:** The number of maximum features extracted using TF-IDF was tested with the following values: 100, 500, 1,000, 5,000, 10,000, 50,000, and 100,000.
- **TF-IDF N-Grams Range:** The range of n-grams used in the TF-IDF was explored with the following configurations: (1,1), (1,2), (1,3), (1,4), (1,5), and (1,10).
- **TF-IDF Lower Case:** The effect of case sensitivity was tested, with the lowercasing of text being either `True` or `False`.
- **TF-IDF Analyzer:** The type of analyzer applied in the TF-IDF process was either `Word` or `Char`.

Regarding computational resources, this study relied on Google Cloud N1 Compute Engines to perform the tuning and training of both the baseline and the BERT architecture. For the baseline, an N1 instance with 192 CPU cores and 1024 GB of RAM was used. For BERT, we used an instance with 16 CPU cores, 30 GB of RAM, and 4x Tesla T4 GPUs. The grid search on N-grams takes approximately three hours under these conditions, while for BERT, it takes approximately 52 hours to complete. The final training took three hours for N-grams and approximately ten hours for BERT.

We have made our codebase open-source[15] to promote reproducibility of our results and to encourage further research in this area.

6 Results

Impact of Delexicalization

Figure 1 depicts the average F_1 scores obtained in the PtBrVid validation set by the N-grams and BERT models, for each (P_{POS}, P_{NER}) percentage pair. The averages are computed across models trained in different domains.

The results suggest that intermediate levels of delexicalization can yield marginal improvements in model effectiveness. However, high levels of P_{POS} adversely affect model performance. This finding is particularly interesting because previous studies have reported significant reductions in effectiveness due to delexicalization (Sharoff, Wu, and Markert 2010; Lui et al. 2014). Notably, these earlier studies fo-

[11] https://spacy.io/models/pt

[12] https://huggingface.co/docs/transformers/

[13] https://pytorch.org

[14] https://scikit-learn.org/

[15] https://github.com/LIAAD/portuguese_vid

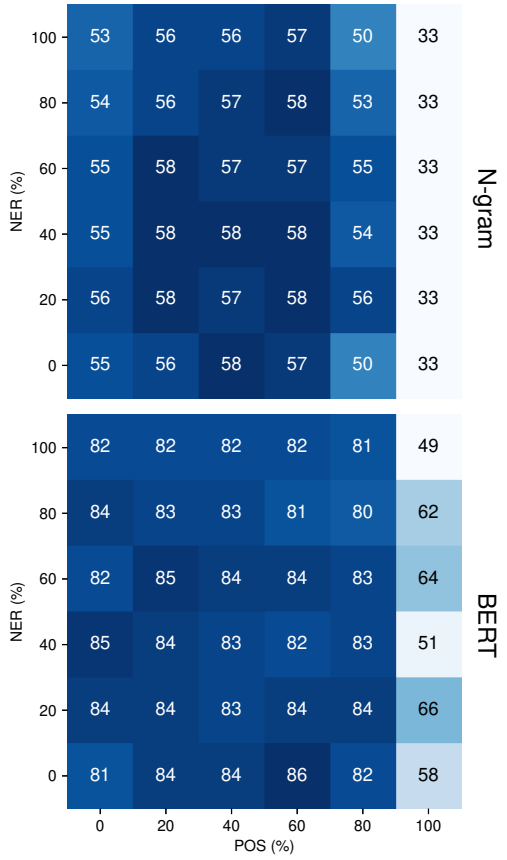

Figure 1: Average F_1 score for each (P_{POS}, P_{NER}).

cused solely on full delexicalization and did not evaluate performance on out-of-domain corpora.

Based on these insights, we proceeded to the second step of our training protocol using a delexicalized version of the training set, with ($P_{\text{POS}} = 0.2$, $P_{\text{NER}} = 0.6$) for the N-gram model and ($P_{\text{POS}} = 0.6$, $P_{\text{NER}} = 0.0$) for BERT models.

Overall Results

This section presents the F_1 scores for the N-gram baseline and BERT fine-tuning models, comparing their performance with and without delexicalization to highlight their impact on the overall effectiveness of the model.

The results in Figure 2 underline the benefits of delexicalization on system effectiveness across both benchmarks and models. Specifically, in FRMT, training in the delexicalized corpus improved the F_1 score by approximately 13 and 10 percentage points for the N-gram and BERT models, respectively.

Upon examining the less pronounced discrepancy in the DSL-TL benchmark, we found it to be largely attributed to the FRMT dataset's entity-specific partition, known as the entity bucket. In this bucket, models trained without delexicalization struggle, as they rely on entities to determine language variety. Given that the FRMT dataset contains the same text in both BP and EP, these models often misclassify pairs of sentences by assigning the same label to both, leading to frequent errors. In the extreme case, they end up getting around half of the labels wrong, which is what happened to the N-gram model, only achieving an F_1 score of 46.95% in this benchmark. This highlights the importance of using delexicalization in the training process. To the best of our knowledge, we are the first to report positive results from the use of delexicalization, which was enabled by the proposed cross-domain training protocol.

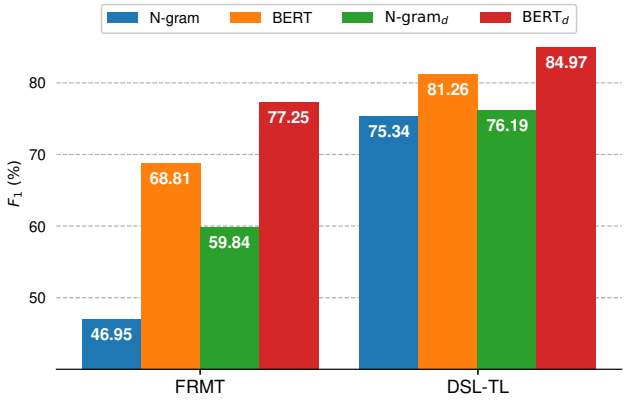

Figure 2: F_1 in FRMT and DSL-TL benchmarks. Models with the subscript d were trained on a delexicalized corpus.

When comparing the BERT model with the N-gram models, one can observe that the BERT model outperforms the N-gram model across all scenarios, achieving an F_1 score of 84.97% in DSL-TL and 77.25% in FRMT. To support further research and exploration, we have made the BERT$_d$ model available on HuggingFace, inviting the research community to use and build on this work[16].

7 Conclusion & Future Work

In this study, we introduced the first multi-domain Portuguese LVI corpus, which includes more than 7 million documents. Leveraging this corpus, we fine-tuned a BERT-based model to create a robust tool for discriminating between European and Brazilian Portuguese. The training strategy leverages delexicalization to mask entities and thematic content in the training set, thereby enhancing the model's ability to generalize. This approach has potential for adaptation to other language variants and languages.

We have identified two key avenues for future work to further enhance the quality and scope of Portuguese LVI. First, the corpus should be expanded to include other less-resourced Portuguese varieties, particularly African Portuguese. Second, it is crucial to explore the impact of the pre-trained model selection, as the language variety on which the model was originally trained may introduce bias into the LVI classifier.

[16] https://huggingface.co/liaad/PtVId

Acknowledgments

This research is supported by national funding from the Portuguese Foundation for Science and Technology (FCT) under the project with DOI 10.54499/LA/P/0063/2020. The authors also acknowledge the support of the StorySense project (DOI 10.54499/2022.09312.PTDC) and the advanced computing project PTicola (ID CPCA-IAC/AV/594794/2023). Hugo Sousa further acknowledges FCT for funding his PhD grant (ID 2022.14691.BD).

References

Bernier-Colborne, G.; Leger, S.; and Goutte, C. 2021. N-gram and Neural Models for Uralic Language Identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, 128–134. Association for Computational Linguistics.

Bernier-Colborne, G.; Leger, S.; and Goutte, C. 2022. Transfer learning improves French cross-domain dialect identification. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, 109–118.

Brito, A. M.; and Lopes, R. E. 2016. *The Handbook of Portuguese Linguistics*. Wiley, 1st edition. ISBN 9781118791950.

Castro, D.; Souza, E.; and Oliveira, A. D. 2016. Discriminating between Brazilian and European Portuguese National Varieties on Twitter Texts. In *2016 5th Brazilian Conference on Intelligent Systems*, 265–270. IEEE. ISBN 978-1-5090-3566-3.

Cezar, R. F. 2020. Brazilian Presidential Speeches from 1985 to July 2020.

Chew, Y. C.; Mikami, Y.; Marasinghe, C. A.; and Nandasara, S. T. 2009. Optimizing n-gram order of an N-gram based language identification algorithm for 63 written languages. *The International Journal on Advances in ICT for Emerging Regions*, 2.

Çöltekin, Ç.; and Rama, T. 2016. Discriminating Similar Languages with Linear SVMs and Neural Networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, 15–24. The COLING 2016 Organizing Committee.

Çöltekin, Ç.; Rama, T.; and Blaschke, V. 2018. Tübingen-Oslo team at the VarDial 2018 evaluation campaign: An analysis of n-gram features in language variety identification. In *Proceedings of the fifth workshop on nlp for similar languages, varieties and dialects*, 55–65.

da Cunha, L. C. C. 2021. FakeWhatsApp. BR: Detecção de Desinformação e Desinformadores em Grupos Públicos do WhatsApp em PT-BR.

de Vargas Feijó, D.; and Moreira, V. P. 2018. *RulingBR: A Summarization Dataset for Legal Texts*, 255–264.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*, 4171–4186. Association for Computational Linguistics.

Diwersy, S.; Evert, S.; and Neumann, S. 2014. *A weakly supervised multivariate approach to the study of language variation*, 174–204. DE GRUYTER.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76: 378.

Généreux, M.; Hendrickx, I.; and Mendes, A. 2012. A large Portuguese corpus on-line: cleaning and preprocessing. In *Computational Processing of the Portuguese Language: 10th International Conference, PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings 10*, 113–120.

Goutte, C.; Léger, S.; and Carpuat, M. 2014. The NRC system for discriminating similar languages. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, 139–145.

Guggilla, C. 2016. Discrimination between Similar Languages, Varieties and Dialects using CNN- and LSTM-based Deep Neural Networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, 185–194. The COLING 2016 Organizing Committee.

Holmberg, A.; and Platzack, C. 2008. *The Scandinavian Languages*. Oxford University Press. ISBN 9780195136517.

Jauhiainen, T.; Jauhiainen, H.; and Lindén, K. 2022. Italian Language and Dialect Identification and Regional French Variety Detection using Adaptive Naive Bayes. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, 119–129. Association for Computational Linguistics.

Joshi, A.; Dabre, R.; Kanojia, D.; Li, Z.; Zhan, H.; Haffari, G.; and Dippold, D. 2024. Natural Language Processing for Dialects of a Language: A Survey.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, 79–86.

Lui, M.; and Baldwin, T. 2011. Cross-domain Feature Selection for Language Identification. In Wang, H.; and Yarowsky, D., eds., *Proceedings of 5th International Joint Conference on Natural Language Processing*, 553–561. Asian Federation of Natural Language Processing.

Lui, M.; and Baldwin, T. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, 25–30.

Lui, M.; Letcher, N.; Adams, O.; Duong, L.; Cook, P.; and Baldwin, T. 2014. Exploring Methods and Resources for Discriminating Similar Languages. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, 129–138.

Malmasi, S.; and Dras, M. 2015. Language Identification using Classifier Ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, 35–43. Association for Computational Linguistics.

Martins, B.; and Silva, M. J. 2005. Language Identification in Web Pages. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, 764–768. Association for Computing Machinery. ISBN 1581139640.

McNamee, P. 2005. Language Identification: A Solved Problem Suitable for Undergraduate Instruction. *J. Comput. Sci. Coll.*, 20: 94–101.

Medvedeva, M.; Kroon, M.; and Plank, B. 2017. When Sparse Traditional Models Outperform Dense Neural Networks: the Curious Case of Discriminating between Similar Languages. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, 156–163. Association for Computational Linguistics.

Nguyen, V.; Karimi, S.; Rybinski, M.; and Xing, Z. 2021. Cross-Domain Language Modeling: An Empirical Investigation. In *Proceedings of the The 19th Annual Workshop of the Australasian Language Technology Association*, 192–200. Australasian Language Technology Association.

Öhman, J.; Verlinden, S.; Ekgren, A.; Gyllensten, A. C.; Isbister, T.; Gogoulou, E.; Carlsson, F.; and Sahlgren, M. 2023. The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling.

Popa, C.; and Stefănescu, V. 2020. Applying Multilingual and Monolingual Transformer-Based Models for Dialect Identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 193–201. International Committee on Computational Linguistics.

Ramalho, M. S. 2021. High-level Approaches to Detect Malicious Political Activity on Twitter.

Raposo, E.; Vicente, G.; and Veloso, R. 2021. *Geografia da Língua Portuguesa*, volume 1, 71–81. Fundacao Galouste Gulbenkian.

Riley, P.; Dozat, T.; Botha, J. A.; Garcia, X.; Garrette, D.; Riesa, J.; Firat, O.; and Constant, N. 2023. FRMT: A Benchmark for Few-Shot Region-Aware Machine Translation. *Transactions of the Association for Computational Linguistics*, 11: 671–685.

Rocha, P. G. M.; and Santos, D. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa.

Rodrigues, J.; Gomes, L.; Silva, J.; Branco, A.; Santos, R.; Cardoso, H. L.; and Osório, T. 2023. *Advancing Neural Encoding of Portuguese with Transformer Albertina PT-**, 441–453. Springer Nature Switzerland. ISBN 9783031490088.

Santos, D. 2014. Corpora at Linguateca: Vision and Roads Taken. *Working with Portuguese corpora*, 219–236.

Sharoff, S.; Wu, Z.; and Markert, K. 2010. The Web Library of Babel: evaluating genre collections. In Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; and Tapias, D., eds., *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. European Language Resources Association.

Silva, J. D.; and Lopes, G. 2006. Identification of Document Language is Not yet a Completely Solved Problem. In *2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce*, 212–212. IEEE. ISBN 0-7695-2731-0.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. *BERTimbau: Pretrained BERT Models for Brazilian Portuguese*, 403–417.

Suarez, P. J. O.; Sagot, B.; and Romary, L. 2019. Asynchronous Pipelines for Processing Huge Corpora on Medium to Low Resource Infrastructures. 9 – 16. Leibniz-Institut f"ur Deutsche Sprache.

Tan, L.; Zampieri, M.; Ljubešic, N.; and Tiedemann, J. 2014. Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora*, 11–15.

Vaidya, A.; and Kane, A. 2023. Two-stage Pipeline for Multilingual Dialect Detection. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, 222–229. Association for Computational Linguistics.

Vargas, F.; Carvalho, I.; de Góes, F.; Pardo, T.; and Benevenuto, F. 2022. HateBR: A Large Expert Annotated Corpus of Brazilian Instagram Comments for Offensive Language and Hate Speech Detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 7174–7183. European Language Resources Association.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008. ISBN 9781510860964.

Yang, L.; and Xiang, Y. 2019. Naive Bayes and BiLSTM Ensemble for Discriminating between Mainland and Taiwan Variation of Mandarin Chinese. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, 120–127. Association for Computational Linguistics.

Zaharia, G.-E.; Avram, A.-M.; Cercel, D.-C.; and Rebedea, T. 2020. Exploring the Power of Romanian BERT for Dialect Identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 232–241. International Committee on Computational Linguistics.

Zampieri, M.; and Gebre, B. 2014. VarClass: An Open-source Language Identification Tool for Language Varieties. In Calzolari, N.; Choukri, K.; Declerck, T.; Loftsson, H.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. European Language Resources Association.

Zampieri, M.; and Gebre, B. G. 2012. Automatic identification of language varieties: The case of Portuguese. In *The 11th Conference on Natural Language Processing*, 233–237.

Zampieri, M.; North, K.; Jauhiainen, T.; Felice, M.; Kumari, N.; Nair, N.; and Bangera, Y. M. 2024. Language Variety Identification with True Labels. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 10100–10109. ELRA and ICCL.