Enhancing Portuguese Varieties Identification with Cross-Domain Approaches

Anonymous ACL submission

Abstract

Recent advances in natural language processing (NLP) have significantly raised expectations for generative models to produce coherent text across diverse languages varieties. In the particular case of the Portuguese language, a predominance of Brazilian Portuguese corpora online induces linguistic traces on those models, limiting their adoption outside 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. The findings of the literature review process motivated us to compile PtBrVarId, a cross-domain LVI corpus, and to study how transformer-based LVI classifiers can be optimised to perform in a cross-domain scenario. Our most effective model, a PtBrVarId fine-tuned version of BERT, sets a new state-of-the-art result of 0.84 F_1-Score on the DSL-TL corpus, the LVI reference benchmark. This result was obtained while maintaining state-of-the-art (SOTA) results (above 0.90 F_1-Score) in the cross-domain scenario. Although this research is focused 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 NLP task (Joshi et al., 2024). Over time, populations sharing a common language can evolve distinctive speech traits due to geographical and cultural factors, including migration and the influence of other languages (Raposo et al., 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). Be it on the pretraining, 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 frequently struggle to identify linguistically relevant features, showing a tendency to be biased towards non-linguistic factors, such as named entities and thematic content (Diwersy et al., 2014). Consequently, these classifiers exhibit limited transfer capabilities to domains not represented in the training set, significantly restricting their utility in multi-domain applications (Sharoff et al., 2010; Lui and Baldwin, 2011).

A language where variety identification is particularly challenging is Portuguese. It is spoken by over 200 million people worldwide and serves as the official language of eight nations across five continents, each one with its one variety. However, 88% of Portuguese speakers are Brazilian citizens, making most of the resources labelled as Portuguese being dominated by this variety. Another important characteristic of Portuguese is that, unlike languages where differences are predominantly phonological, such as those in the North Germanic family[1], the widespread dispersion of Portuguese has fostered considerable phonological, morphological, lexical, syntactic, and semantic variations among Portuguese varieties (Scherre and Duarte, 2016; Kato and Martins, 2016; Brito and Lopes, 2016; Silva, 2013). In LLM development, for example, this variety divergence has practical implications; models trained on Brazilian Portuguese generate texts that are markedly distinct from those trained on 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. For example, legal and medical applications. This underscores the practical importance of developing effective

[1] https://shorturl.at/cRTY8

LVI systems that can be deployed into production and, consequently, to democratize the access to effective LLMs in lower resourced varieties.

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 one. Our multi-domain corpus contains more than 200M silver-labelled tokens. Additionally, a small set of 25k tokens was manually annotated by three linguists to measure the quality of the silver-labelling scheme. The model development began with an evaluation of the cross-domain capabilities of various LVI architectures. Then, we studied the impact of masking the named entities and thematic content embedded in the training corpus by replacing it by its NER/part-of-speech categories, in a process named delexicalization (Lui et al., 2014). We tested different delexicalization probabilities during the hyperparameter tuning process to find the one that optimizes LVI cross-domain effectiveness. To summarise, the contributions of this work are the following:

1. We introduce a novel multi-domain silver-labelled LVI corpus for Brazilian and European Portuguese, compiled from datasets originally designed for a broad range of NLP tasks;

2. We present a comprehensive evaluation of SOTA LVI models across six domains, assessing their effectiveness and identifying areas for improvement, shedding light on the adaptability and effectiveness of existing models when applied to different domains;

3. We study the impact of different levels of delexicalization on the overall effectiveness of LVI models.

4. We open-source[2] the code used to develop this research along with the most effective models and a demo[3] that exploits the explainability technique LIME (Ribeiro et al., 2016).

2 Related Work

The VarDial workshop[4] compiles many of the recent studies developed in the LVI task. In the following subsections, we list these and other resources that include, to some extent, Portuguese LVI resources.

2.1 Corpora

Despite the numerous works developed in the LVI task, the first gold-labelled dataset that includes Portuguese corpora, the DSL-TL corpus (Zampieri et al., 2023), was only introduced in 2023. Prior to the release of this dataset, the training, and evaluation process was often performed in silver-labelled 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's; Brazilian newspapers' articles are assigned a Brazilian Portuguese label, while Portuguese ones are assigned a European Portuguese label (Da Silva and Lopes, 2006; Zampieri and Gebre, 2012; Tan et al., 2014). In the social media domain, a similar approach is frequently used. (Castro et al., 2016) used geographic metadata collected by Twitter/X to assign a language variety to each document based on author's localization.

Many of these Portuguese LVI resources (Da Silva and Lopes, 2006; Zampieri and Gebre, 2012; Castro et al., 2016) are no longer available online. This limitation coupled with prior concerns regarding the reliability of evaluation processes founded on silver-labelled corpora (Zampieri and Gebre, 2014) motivated the introduction of DSL-TL (Zampieri et al., 2023). This dataset used crowdsourcing to annotate approximately 5k Portuguese documents. It includes not only European and Brazilian Portuguese documents, 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.

2.2 Techniques Used

The high efficiency observed in various LID studies, coupled with the similarity to the LVI task, suggested the application of these methods in the context of LVI. In particular, n-gram-based techniques (McNamee, 2005; Martins and Silva, 2005; Chew et al., 2009) which had previously revealed SOTA effectiveness in the LID task (↑ 90.0% Accuracy). Therefore it is not uncommon to observe recent studies submitted to VarDial employing these techniques applied to different language varieties: Italian (0.90 F_1 Jauhiainen et al., 2022); b) Uralic

[2] https://shorturl.at/npBIO
[3] https://shorturl.at/inN36
[4] https://aclanthology.org/venues/vardial/

(0.94 F_1 Bernier-Colborne et al., 2021) or c) Mandarin (0.91 F_1 Yang and Xiang, 2019), to cite just the most recent ones.

The adoption of transformer-based techniques (Vaswani et al., 2017) in LVI has not been as fast as in other NLP tasks. Recently, some works have emerged leveraging mono-lingual BERT-based models to fine-tune LVI classifiers in Romanian (0.65 F_1 Zaharia et al., 2020) and French (0.43 F_1 Bernier-Colborne et al., 2022). In none of these cases; however, transformers were capable of outperforming n-gram-based techniques. Similar challenges have also been reported for different languages using other deep-learning techniques: a) Multilingual transformers (Popa and Ștefănescu, 2020); b) Feed-forward neural networks (Medvedeva et al., 2017; Çöltekin and Rama, 2016); c) LSTMs (Guggilla, 2016); d) RNNs (Çöltekin et al., 2018).

In the particular case of Portuguese (Table 1), older studies have relied on n-grams-based techniques to obtain results above 90% accuracy on silver-labelled benchmarks. The preliminary results obtained in the gold labelled DSL-TL corpus revealed, however, more modest results (below 0.70 F_1). Additionally, contrarily to what was often observed in silver-labelled evaluation (Medvedeva et al., 2017), the current SOTA result for Portuguese LVI in the DSL-TL benchmark (0.79F_1-score) is a deep-learning based method (Vaidya and Kane, 2023). More precisely, a fine-tuned version of Portuguese BERT, BERTimbau (Souza et al., 2020). Even though the results are not easy to compare because of different benchmarks and metrics used, the differences between gold and silver-labelled evaluations illustrate how limited of current SOTA Portuguese LVI classifiers can be.

2.3 Cross Domain Capabilities: Delexicalization

Focusing on cross-domain effectiveness of LVI classifiers. (Lui and Baldwin, 2011) revealed that n-grams based techniques had limited cross-domain capabilities for the LID task. Despite the good results of these models when both the train and test domain overlap (↑85% accuracy), the effectiveness decreased up to ↓40% when both sets don't match. In order to address this phenomenon, the author has devised a feature selection mechanism that later opened the door to the development of the first cross-domain **LID** tool, the langid.py (Lui and Baldwin, 2012).

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

Similar concerns had previously been pointed in other NLP tasks like genre classification (Sharoff et al., 2010) for n-gram based methods. In spite of these facts, the mass adoption of these architectures in the context of LVI, create urgency of 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 transformers architectures.

As far as our knowledge extends, the feature selection described above (Lui and Baldwin, 2011) and the *delexicalization* method (Lui et al., 2014) were the only techniques proposed to overcome these limitations. The concept of delexicalization proposes that each input token be replaced by its part-of-speech (POS) tag as a means of masking the thematics embedded within the training corpus. Nevertheless, previous usage of this technique presented significant effectiveness reductions (Lui et al., 2014: ↓ 0.25 F_1-score; Sharoff et al., 2010: ↓ 14.46% accuracy). We thus believe it is useful to study how intermediate levels of delexicalization impact the overall effectiveness of these models. Additionally, it is also important to clarify how delexicalization affects deep-learning methods. Since feature selection approaches tend to be either redundant or hard to apply to deep learning architectures, delexicalization remains as the only technique proposed in literature to develop neural LVI cross-domain models.

3 Develop an Off-the-Shelf Portuguese LVI Classifier

After reviewing the LVI literature, we conclude there is a lack of multi-domain resources, raising concerns about the true effectiveness of SOTA LVI classifiers. Further studies are also required regarding techniques to promote models' cross-domain effectiveness. To address this situation, we intro-

Study	Technique	Test Set	Bench.
(Da Silva and Lopes, 2006)	N-grams + Clustering	A.D	97.83% Pre.
(Zampieri and Gebre, 2012)	N-grams + Naive B.	A.D	99.00% Acc.
(Goutte et al., 2014)	N-grams + SVM	DSLCC	95.60% Acc.
(Malmasi and Dras, 2015)	N-grams + Ensemble of SVMs	DSLCC	95.54% Acc.
(Castro et al., 2016)	N-grams + Naive B.	A.D	92.71% Acc.
(Zampieri et al., 2023)	**N-grams + Naive B.**	**DSL-TL**	**0.60** F_1
	mBERT	**DSL-TL**	**0.62** F_1
	XLM-R	**DSL-TL**	**0.67** F_1
(Vaidya and Kane, 2023)	**Mixture of BERT Experts**	**DSL-TL**	**0.79** F_1

Table 1: Effectiveness of Portuguese LVI models. The resources in **bold** highlight those that were evaluated in gold-labelled corpora. When the test set has been defined by the respective authors, we represent it with A.D (Author Defined).

duce the first multi-domain Portuguese LVI corpus, the PtBrVarId. This resource creates the opportunity for an extensive study of cross-domain capabilities of different LVI techniques. In particular, pre-trained Portuguese transformers.

The development of off-the-shelf LVI tools requires models not only to be effective, but also fast and light inference processes. For that reason, we start our analysis with the smallest Portuguese transformer available, BERTimbau base (Souza et al., 2020), and move towards more complex architectures based on the results obtained. Regarding techniques to promote models' cross-domain effectiveness, we focus our attention on delexicalization (Lui et al., 2014). To obtain a clear picture of the impact of delexicalization in overall models' effectiveness, all the results in this study are presented with its equivalent non-delexicalzed training version.

4 PtBrVarId: Multi-Domain Portuguese LVI Dataset

The development of the first six-domain Portuguese LVI corpus (journalistic, legal, politics, web, social media and literature) started with the compilation of corpora from 11 different data sources. We decided to name our dataset PtBrVarId, since it only considers two labels; European (PT-PT) and Brazilian Portuguese (PT-BR).

The silver-labelling scheme adopted allowed the automatic annotation of over 200M tokens. Additionally, PtBrVarId also includes a small set of manually annotated documents (25k tokens), which we hereafter refer to as **platinum test set**. This test set serves two purposes: a) Probe the quality of the automatic annotation and b) Estimate the cross-domain capabilities of the models developed.

In the following sections, we describe the most important steps during the development of PtBrVarId. These results are complemented with information in Appendix B where more detailed per-domain/per-variety analysis are introduced.

4.1 Compiling Pre-Existent Corpora

In this section, we describe the data sources used in each textual domain together with the heuristics that supported the silver-labelling step. This information is summarised in Table 2.

Literature relies on three data sources that index classics of Portuguese literature: a) The Gutenberg project; b) The LT-Corpus and c) Brazilian Literature corpus. We used the author's nationality to distinguish between European and Brazilian Portuguese books.

Politics compiles manually transcriptions of political speeches in both the European Parliament (Koehn, 2005) and the Brazilian Senate. We rely on the gold-labelled characteristics of these sources to confidently use document's origin to distinguish between both Portuguese varieties.

Journalistic uses the CETEM corpus (Rocha and Santos, 2000) to extract news articles from Portuguese newspaper Público and Brazilian newspaper Folha de São Paulo. The geographic location of the newspaper is used to assume a Portuguese variety.

Social Media corpora derives from three data sources. The manually annotated Brazilian Portuguese hate speech corpus, Hate-BR (Vargas et al., 2022), and a compilation of fake news spread in

Brazilian WhatsApp groups (Cunha, 2021). Regarding European Portuguese, the tweets collected by (Ramalho, 2021) were filtered based on tweets' metadata location. Tweets whose location is not part of Wikipedia's list of Portuguese cities[5], were discarded.

Web corpora was extracted from OSCAR (Ortiz Suarez et al., 2019). We established an allow list of over 100 subdomains for both .pt and .br geographies, composed of informal descriptive websites representative of Web data.

Domain	# Documents	# Tokens
Literature	74k	47M
Legal	29M	133M
Political	650k	5M
Journalistic	200M	1.7M
Web	80k	26M
Social Media	18M	32M

Table 2: Number of documents and tokens for each domain.

4.2 Quality Assurance Process

In Table 3 we present the agreement between the three Portuguese nationals that performed the annotations using Fleiss's Kappa (Fleiss, 1971). Each annotator was asked to label the Portuguese variety and the textual domain in a class balanced sample of 300 documents extracted from the dataset (50 from each domain, 25 European, 25 Brazilian Portuguese); documents without sufficient variety/domain linguistic features could be labelled as **"undetermined"** by the annotators.

Annotation	Metric	Result (%)
Varieties	Fleiss' Kappa	57.0
	Majority Rate	95.3
	Accuracy	85.6
Domain	Fleiss' Kappa	69.0
	Majority Rate	94.0
	Accuracy	76.0

Table 3: Agreement among the three annotators regarding both the documents' language variety and textual domain.

The results were then compared with the automatic annotation to determine: a) How frequent is a 2/3 majority among the annotators possible (Majority Rate) and b) How aligned this majority is with the automatic annotation (Accuracy).

The agreement is higher for the textual domain than about the Portuguese variety. Nevertheless, a 2/3 majority remains almost always possible ($\uparrow$ 90.0%). This majority is also highly aligned with the automatic annotation, with more than$\uparrow$ 70.0% Accuracy. In Table 6 we extend our analysis, presenting per-domain agreement results. We demonstrate that there is a $\downarrow$ 20% Kappa reduction due to introduction of the **"undetermined"** label in the annotation.

Finally, the manually annotated documents where a 2/3 majority was possible were compiled to create the platinum test set.

5 Experimental Setup

5.1 Establish Baselines

The good results reported by LVI studies in different Indo-European languages, including Portuguese (Zampieri and Gebre, 2012), using N-gram combined with Naive Bayes classifiers (Table 1) motivate us to use this technique as baseline to evaluate the effectiveness gains/decreases of the different techniques used in this study. Furthermore, as previously mentioned in Section 2.2, the 0.79 F_1-score result obtained in the DSL-TL corpus serves as a trustworthy benchmark for Portuguese LVI.

5.2 Cross-domain Evaluation of LVI Classifiers: Three Step Process

The development of an effective cross-domain LVI classifier required us to develop a three-step evaluation process capable of assessing models' cross-domain capabilities. First, each model is evaluated on the silver-labelled validation sets defined for each of the six textual domains.

Then, we used two gold-labelled test sets, the DSL-TL corpus and the "entity bucket adverbial cases" (Riley et al., 2022) of FRMT: Few-shot Region-aware Machine Translation to obtain a trustworthy estimation of the F_1-scores of LVI classifiers. Despite, originally developed by Google to benchmark machine translation systems, the annotations on the FRMT corpus, can be easily transposed to LVI.

Finally, we used the platinum test set to obtain further details on the model's effectiveness. We consider a model to be reliable if it is a cross-domain tool capable of achieving SOTA results

[5] https://shorturl.at/atEIK

in silver labelled data while maintaining its performance levels both in the gold and platinum-labelled test sets.

5.3 Combining Different Textual Domains

In this study, we follow an iterative approach to the problem of finding the best strategy for combining training corpora from different textual domains in a single training process. We started by leveraging under-sampling to combine the six domains into a single training corpus while ensuring class balanced proprieties in this dataset.

5.4 Delexicalization Framework

Previous studies on delexicalization approached the problem with a coarse-grained strategy, replacing the entire input for its POS tags. We believe a fine-grained methodology is required to evaluate the impact of introducing a token replacement probability hyperparameter P_{POS} in the overall effectiveness of the models. Additionally, we propose to replace the named entities (NER) identified using spaCy[6] by its NER tag with a probability P_{NER}.

In this study, we apply delexicalization exclusively to the train set. The evaluation was done without performing any sort of modification to the input text. The goal is to recreate a real world usage scenario, where text is not transformed. We leave as future work (Section 7) measuring the impact of delexicalizing the test set in the models' effectiveness.

5.5 Tuning Delexicalization

We performed hyperparameter tuning to determine the best delexicalization probabilities (P_{POS}, P_{NER}). We performed six parallel grid searches, one for each domain, using a stratified training sample of 5000 documents. Each grid search was evaluated using the five validation sets from the domains different from the training one. The goal is to determine the parameters that optimise cross-domain performance.

Despite our focus on delexicalization, other training parameters were evaluated during grid search. The parameters assessed vary according to the technique under scrutiny; a list of those parameters are presented in Table 5.

In Heatmaps 1 and 2 we report with a probability step of 0.2 the average F_1-scores obtained in the six parallel grid searches for each (P_{POS}, P_{NER}) pair.

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

Figure 1: Hyperparameter tuning results for different levels of delexicalization in the n-grams setting. Each cell represents the F_1-score of the best performing textual domain for for that (P_{POS}, P_{NER}) set of values.

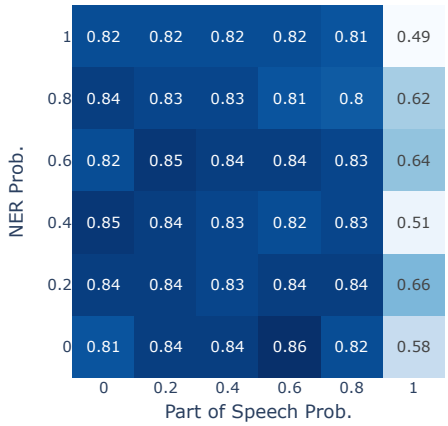

Figure 2: Hyperparameter tuning results for different levels of delexicalization in the BERT finetuning setting. Each cell represents the F_1-score of the best performing textual domain for for that (P_{POS}, P_{NER}) set of values.

The results reveal: a) Marginal gains are possible using intermediate levels of delexicalization; b) High levels of P_{POS} have a negative impact on models' effectiveness; c) BERT-based models present higher effectiveness in the cross-domain scenario than the n-grams. Based on these findings, we decided to proceed to the training stage with a delexicalization version of the training set with ($P_{\text{POS}} = 0.6 \wedge P_{\text{NER}} = 0.0$) in the case of BERT and ($P_{\text{POS}} = 0.2 \wedge P_{\text{NER}} = 0.6$) in the case of n-grams.

6 Results

The following section reports the F_1-scores obtained by N-grams baseline and BERT fine-tuning using the optimized parameters derived from the hyperparameter tuning step (Section 5.5). All the

results are presented together with its equivalent non-delexicalized version; to easily observe how delexicalization affects overall model effectiveness.

6.1 N-Grams

The results presented in Figure 3 clarify the gains delexicalization promotes in n-gram-based approaches. In five out of eight domains this technique was beneficial with a particular focus to the gold-labelled FRMT corpus, where a gain of $\uparrow$ 0.13 F_1-score was achieved.

Even though the experiments were not optimised for the DSL-TL evaluation, our baseline establishes a new benchmark in this corpus of 0.76 F_1-score using non-neural techniques.

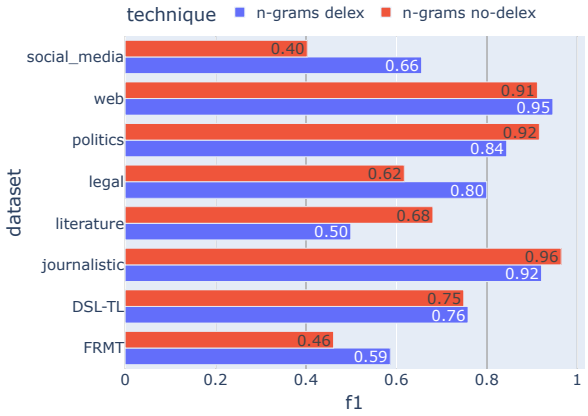

Figure 3: N-grams F_1 effectiveness in silver/gold-labelled test sets.

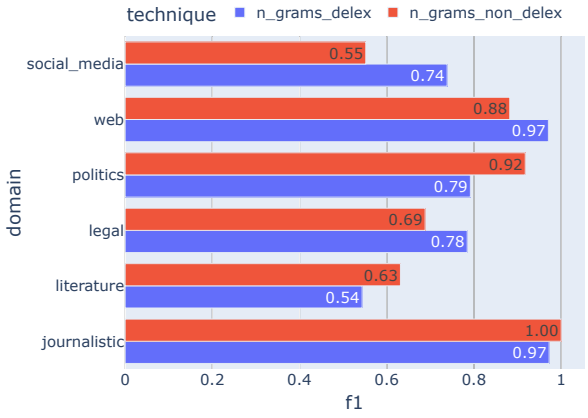

Figure 4: N-grams F_1 effectiveness in the platinum test set.

Importantly, the results obtained in the platinum test set (Figure 4) corroborate the findings mentioned above. In particular, the five domains that benefit from delexicalization overlap the findings of silver-labelled evaluation.

6.2 BERT

The results presented in Figure 5 clarify the overall improvement BERT architectures introduced in the Portuguese LVI task. Consistent results above 0.90 F_1 introduce average gains of $\uparrow$ 0.10 F_1 when compared with the n-grams' baseline.

Regarding the impact of delexicalization, the effectiveness gains/reduction on BERT-based approaches are marginal. Again, the benefits of this technique are more notorious in gold-labelled test sets. Delexicalization helped set a new benchmark on the DSL-TL corpus of 0.84 F_1, an improvement of $\uparrow$ 0.05 F_1 when compared with the current SOTA.

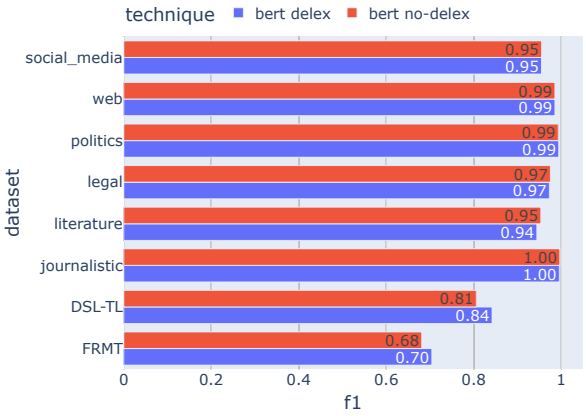

Figure 5: BERT F_1 effectiveness in silver/gold labelled test set.

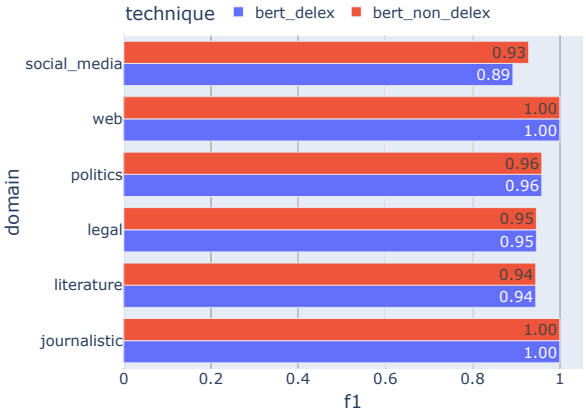

Figure 6: BERT F_1 effectiveness in the platinum test set.

Additionally, the results in the platinum test set (Figure 6), corroborate the findings mentioned above.

6.3 Overall Results

The effectiveness reported by BERT in both silver, gold and platinum labelled data provide sufficient cross-domain capabilities to deliver the first cross-domain LVI tool. Additionally, the fact that both N-grams (0.76 F_1-score) and BERT-based methods (0.84 F_1-score) were able to set SOTA results in the DSL-TL benchmark, even when they were not optimised to do so, sheds lighting on the potential the PtBrVarId corpus introduces in future Portuguese LVI studies.

7 Conclusion & Future Work

In this study, we introduce the first multi-domain Portuguese LVI corpus with over 200 million tokens evaluated by three annotators. We used this corpus to develop the first cross-domain Portuguese LVI model. The model has been obtained by fine-tuning a Portuguese BERT base architecture to deliver a fast, light and reliable tool to discriminate between European and Brazilian Portuguese. The development of this cross-domain architecture employs delexicalization techniques to mask entities and thematics embedded in the training set, increasing the cross-domain capabilities of these models. The F_1-scores obtained on gold labelled data establish a SOTA result of 0.84 F_1-score in the DSL-TL benchmark, illustrating the potential of this tool. The model will now be integrated in other ongoing project headed by our research team that aims to develop a large European Portuguese corpus to support the training of a SOTA European Portuguese LLM.

We identify four future work topics to further improve the quality of Portuguese LVI. First, the expansion of the corpus to other Portuguese varieties with less resources available, namely African. Second, the evaluation of different Portuguese transformers in this NLP task, we are confident that a more complex architecture would improve the results obtained. Third, we look forward to quantise and prune the transformer architecture developed to provide a light weighted, fast, CPU oriented model up to mass adoption by the NLP community. Fourth, we look forward to evaluating the impact delexicalizing the test set can have in the overall effectiveness of the models developed.

Finally, we believe it is paramount to quantify the effort it would take to adapt our experimental setup to other Portuguese varieties/European languages. Regarding Portuguese varieties, since the code developed was designed to easily expand towards them, only small adaptations on the automatic labelling scheme and the manual annotation of an equivalent platinum test set for the new varieties would be required.

In the case of other European languages, additional steps would be necessary. For example, the adoption of other mono-lingual transformers. Nevertheless, a good starting point for such endeavour would be British/American English and Castilian/Argentinian Spanish. Both languages have mono-lingual BERTs to support the task, and are included as part of the DSL-TL corpus, whose annotation is able to provide trustworthy evaluations following our three steps proposal.

Limitations

We identify two main limitations related with the dataset used that engage directly with the work developed. First, despite our efforts, parts of the evaluation are still founded on silver-labelled data. Which, as we mentioned in the paper, is often considered in the LVI literature misleading. Additional manually annotations are desirable to increase the confidence in the results obtained.

Second, many documents collected online do not have sufficient linguistic traces to confidently classify it as a single variety. To surpass this limitation, the DSL-TL corpus introduced the possibility of a "Both/Neither" class to signal those cases. Our silver-labelling process does not take into consideration those cases, which introduces entropy in the training data and could potential negatively impact the overall effectiveness of the models developed with our corpus.

Ethical Considerations

We identify two ethical aspects our work engages with that should be discussed to benefit transparency and open-minded science. First, we compile existing corpora with permissive scientific licensing. We use Brazilian datasets related to hate speech and social media comments in the social media domain. Unfortunately, the lack of respect witnessed in social media transposes to our corpus, with vast amounts of racism, xenophobia, toxic masculinity, and harassment presented in our social media corpus. Also, the silver-label nature of the social media domain is particularly challenging because it often mentions other persons by their names or other unique forms of mentioning; addi-

tional means of anonymization should be implied in a 1.0 Version of our corpus since there is no linguistic gain in incorporating this mentions that can impact negatively the privacy of individuals.

Secondly, it is imperative to mention that our multinational research team is composed of elements from four continents, including Portuguese and Brazilian elements that were consulted during the development of this tool. It was mentioned that in both countries, there are negative attitudes towards the other variant of Portuguese, with small discussions in Portugal claiming the "purity of the language" as a former colonial power and in Brazil claiming the right to the "evolution of a self-linguistic identity" as a new rising multicultural power.

In the past, some literature reviews point to works in this field by Balkan researchers with heavily political intentions. Even though we acknowledge that our research can fuel the discussion on the Portuguese language in this topic, we accept the burden because we believe that the Portuguese language as an all benefits from the difference in variants, not only the European and Brazilian ones, but also the many African variants, and also the Asian variants of Macau and Oceanic's East-Timor. As mentioned in the conclusions, one of the future work points is to extend our work to these variants to create a Portuguese corpus with all existent variants in an actual exercise of diversity rather than nefarious purity discussions.

References

Gabriel Bernier-Colborne, Serge Leger, and Cyril Goutte. 2021. N-gram and neural models for uralic language identification: NRC at VarDial 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 128–134, Kiyv, Ukraine. Association for Computational Linguistics.

Gabriel Bernier-Colborne, Serge Leger, and Cyril Goutte. 2022. Transfer learning improves french cross-domain dialect identification: Nrc@ vardial 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 109–118.

A. M. Brito and R. E. Lopes. 2016. The structure of dps. In W. L. Wetzels, S. Menuzzi, and J. Costa, editors, *The Handbook of Portuguese Linguistics*, 1st edition, pages 254–274. Wiley Blackwell.

Dayvid Castro, Ellen Souza, and Adriano LI De Oliveira. 2016. Discriminating between brazilian and european portuguese national varieties on twitter texts. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 265–270. IEEE.

Yew Choong Chew, Yoshiki Mikami, Chandrajith Ashuboda Marasinghe, and S Turrance Nandasara. 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(2).

Çağrı Çöltekin and Taraka Rama. 2016. Discriminating similar languages with linear SVMs and neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 15–24, Osaka, Japan. The COLING 2016 Organizing Committee.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 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 (vardial 2018)*, pages 55–65.

Lucas Cabral Carneiro da Cunha. 2021. Fakewhatsapp. br: detecção de desinformação e desinformadores em grupos públicos do whatsapp em pt-br.

Joaquim Ferreira Da Silva and Gabriel Pereira Lopes. 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 (CIMCA'06)*, pages 212–212. IEEE.

Sascha Diwersy, Stefan Evert, and Stella Neumann. 2014. A weakly supervised multivariate approach to the study of language variation. *Aggregating dialectology, typology, and register analysis. linguistic variation in text and speech*, pages 174–204.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Cyril Goutte, Serge Léger, and Marine Carpuat. 2014. The nrc system for discriminating similar languages. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 139–145.

Chinnappa Guggilla. 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 (VarDial3)*, pages 185–194, Osaka, Japan. The COLING 2016 Organizing Committee.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 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*, pages 119–129, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

M. A. Kato and A. M. Martins. 2016. European portuguese and brazilian portuguese: An overview on word order. In W. L. Wetzels, S. Menuzzi, and J. Costa, editors, *The Handbook of Portuguese Linguistics*, 1st edition, pages 15–40. Wiley Blackwell.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th international joint conference on natural language processing*, pages 553–561.

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

Marco Lui, Ned Letcher, Oliver Adams, Long Duong, Paul Cook, and Timothy Baldwin. 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*, pages 129–138.

Shervin Malmasi and Mark Dras. 2015. Language identification using classifier ensembles. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 35–43, Hissar, Bulgaria. Association for Computational Linguistics.

Bruno Martins and Mário J. Silva. 2005. Language identification in web pages. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC '05, page 764–768, New York, NY, USA. Association for Computing Machinery.

Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of computing sciences in colleges*, 20(3):94–101.

Maria Medvedeva, Martin Kroon, and Barbara Plank. 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 (VarDial)*, pages 156–163, Valencia, Spain. Association for Computational Linguistics.

Pedro Javier Ortiz Suarez, Benoit Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Cristian Popa and Vlad Ștefănescu. 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*, pages 193–201, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Miguel Sozinho Ramalho. 2021. High-level approaches to detect malicious political activity on twitter.

Eduardo Raposo, Grasa Vicente, and Rita Veloso. 2021. *GEOGRAFIA DA LÍNGUA PORTUGUESA*, volume 1, page 71–81. Fundacao Galouste Gulbenkian.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2022. FRMT: A benchmark for few-shot region-aware machine translation.

Paulo Alexandre Rocha and Diana Santos. 2000. Cetempúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. *quot; In Maria das Graças Volpe Nunes (ed) V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. *arXiv preprint arXiv:2305.06721*.

Maria Marta Pereira Scherre and Maria Eugênia Lammoglia Duarte. 2016. Main current processes of morphosyntactic variation. *The Handbook of Portuguese Linguistics*, pages 526–544.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The web library of babel: evaluating genre collections. In *LREC*.

R. V. M. Silva. 2013. O português no contexto das línguas românicas. In E. P. Raposo, M. F. Nascimento, M. A. Mota, L. Segura, and A. Mendes, editors, *Gramática do Português, Volume 1*, pages 145–156. Fundação Calouste Gulbenkian, Lisboa.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. 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)*.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 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 (BUCC)*, pages 11–15.

Ankit Vaidya and Aditya Kane. 2023. Two-stage pipeline for multilingual dialect detection. *arXiv preprint arXiv:2303.03487*.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 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*, pages 7174–7183, Marseille, France. European Language Resources Association.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Li Yang and Yang Xiang. 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*, pages 120–127, Ann Arbor, Michigan. Association for Computational Linguistics.

George-Eduard Zaharia, Andrei-Marius Avram, Dumitru-Clementin Cercel, and Traian Rebedea. 2020. Exploring the power of Romanian BERT for dialect identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties*

and Dialects, pages 232–241, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Marcos Zampieri and Binyam Gebrekidan Gebre. 2012. Automatic identification of language varieties: The case of portuguese. In *KONVENS2012-The 11th Conference on Natural Language Processing*, pages 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI).

Marcos Zampieri and Binyam Gebrekidan Gebre. 2014. Varclass: An open-source language identification tool for language varieties. In *LREC 2014: 9th International Conference on Language Resources and Evaluation*, pages 3305–3308.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels. *arXiv preprint arXiv:2303.01490*.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The nordic pile: A 1.2tb nordic dataset for language modeling.

A European and Brazilian Portuguese: Some Constrative Features

The Portuguese language is an Indo-European, Romance and Iberian language with four branches of varieties: European, Brazilian, African and Asian that feature *phonological*, *morphological*, *lexical*, *syntactic*, and *semantic* differences. Although the PT-PT and PT-BR varieties vary across all these linguistic levels, since our dataset considers exclusively written text, we will exclude the phonological differences from our analysis.

At the morpho-syntactic level, the contrast can be observed, for example, in the pronominal system and the structure of nominal, prepositional and verbal phrases. (Scherre and Duarte, 2016) discuss the variation in Brazilian Portuguese of the 2nd person singular (*tu/ você*, 'you') and 1st person plural (*nós/ a gente*, 'we/the people') nominative pronouns. Additionally, (Kato and Martins, 2016) show how the system and the position of clitics behave distinctively: while in PT-PT, the clitics with the role of complement (*o(s), a(s)* ('him', her', 'it') are widely utilized (*e.g. O João viu a Maria/viu-a*, 'John saw Maria/her'), in PT-BR, nominal phrase or the pronoun *ele/ ela* ('he', 'she') are employed instead (e.g. *O João viu Maria/ ela*, 'John saw Maria/ she'). The position of the clitics is a factor of disparity between the two varieties as well because in PT-PT the clitics are by default placed after the verb (enclisis), and in PT-BR they are positioned before the verb (proclisis) (*e.g. Dá-me um computador/ Me dá um computador*, 'Give me a computer').

The contrast between the two varieties extends also to the structure of nominal and prepositional phrases. (Brito and Lopes, 2016), for instance, refers to the fact that in PT-PT, the possessive is habitually preceded by a definite article, whereas in PT-BR, it can occur by itself (e.g. *O João viu a minha filha/ minha filha*, 'John saw my daughter'). Moreover, PT-BR allows for the use of a bare singular noun, which is disallowed in PT-PT (e.g. *Ontem vi filme no cinema* (PT-PT×; PT-BR✓), 'Yesterday, I saw a film at the cinema'). The expression of datives with the role of an indirect object is also built differently: whereas in PT-PT, the preposition *a* ('to') is used, in PT-BR the preposition is another one, *para* ('to'), as in *O João contou à Maria/para Maria* ('John told Maria'). Another well-known and documented morpho-syntactic difference lies in the opposition between using the infinitive versus gerund in constructions corresponding to the progressive or secondary predicates. In these cases, PT-BR utilises the gerund while PT-PT resorts to the infinitive (e.g. *O João está a ler/lendo*, 'John was reading'.

It is at the lexical level that the two varieties exhibit the most contrast. Besides the different words to represent the same entity (*hospedeira de bordo/aeromoça*, 'stewardess'), Brazilian Portuguese has much vocabulary with indigenous (*caipira, acajá*, and African (dengo, cafuné) origins. Brazilian lexical richness is also the result of the contact with the languages of numerous immigrants and the easiness in accepting neologisms and loanwords (Silva, 2013).

The phonetic-phonological and prosodic differences are undoubtedly the most noticeable and some impact on orthography. When there is a stressed syllable followed by a nasal consonant at the beginning of the next syllable, the timbre of the stressed vowel varies depending on the variety: in PT-PT [ɔ], [e] and in PT-BR [o], [ɛ]. This phonetic feature is marked in writing with different orthographical signs, as illustrated in words like (*homónimo/homônimo*, 'homonymous') and (*grémio/grêmio*, 'guild'). Another case with consequences to the spelling refers to some consonants that are silent in one variety, but not in the other one, or the other way around, and that, when they are not silent, are represented orthographically (e.g. *facto/fato*, 'fact' and *ato/acto*, 'act'). Finally, in terms of orthography, certain specific words have different spellings in each variety, like (*registo/registro*, 'registry').

B Dataset

B.1 Corpora Compiled

In Table 4 we detail the sources compiled to produce PtBrVarId.

B.2 Corpus Splitting: Train-Test Splits

In Table 7 we present the statistics regarding class distribution and number of tokens on PtBrVarId. The dataset has a problem of class imbalance in many domains, which forced us to apply undersampling techniques to improve the training quality.

C Hyper-parameter Tuning

In Table 5 we list the additional parameters to delexicalization, considered during the grid search process.

Domain	Variety	Dataset	Task	License
Literature	PT-PT	Gutenberg Project [7]	-	CC
		LT-Corpus[8]	-	**ELRA**
	PT-BR	Brazilian Literature[9]	Author Id.	CC
		LT-Corpus[10]	-	**ELRA**
Politics	PT-PT	**(Koehn, 2005)**	**Mac.Translation**	**CC-BY-NC-4.0**
	PT-BR	Brazilian Senate Speechs	-	CC
Journalistic	PT-PT	(Rocha and Santos, 2000)	-	CC
	PT-BR	CETEM Folha[11]	-	CC
Social Media	PT-PT	(Ramalho, 2021)	Fake News Detec.	MIT
	PT-BR	**(Vargas et al., 2022)**	**Hate Speech Detec.**	**CC-BY-NC-4.0**
		(Cunha, 2021)	**Fake News Detec.**	**GPL-3.0 license**
Web	Both	(Ortiz Suarez et al., 2019)	-	CC

Table 4: List of pre-existent corpora compiled to produced the Portuguese LVI corpus.

Parameter	Options
TF-IDF Max Features	100
	500
	1,000
	5,000
	10,000
	50,000
	100,00
TF-IDF N-Grams Range	(1,1)
	(1,2)
	(1,3)
	(1,4)
	(1,5)
	(1,10)
TF-IDF Lower Case	True
	False
TF-IDF Analyzer	**Word**
	Char

Table 5: List of hyperparameters tested besides delexicalization. The usage of bold highlights the best result obtained. The parameters name follows the sklearn convention[12]

D Annotation Results

In Table 6 we detail the annotation agreement metrics per-domain for the gold-labelled subset of the LVI dataset proposed.

The low results in the literature domain are explained by its compilation of non-contemporary books. In the 18th and 19th century, the cultural differences between Portuguese and Brazilian writers were less significant, and therefore it creates additional uncertainty. In a version 0.2 of the dataset, we should integrate contemporary literature to achieve full potential from the models.

E 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, no GPU was needed, and it was used N1 instances with 192 CPU cores and 1024 GB of RAM. While for BERT we used an instance with 16 CPU cores, 30 GB of RAM and 4x Tesla T4. The grid search on n-grams takes approximately three

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

[-2] https://shorturl.at/kANY4

[-1] https://www.kaggle.com/datasets/rtatman/brazilian-portuguese-literature-corpus

[0] https://shorturl.at/moDHN

[1] https://www.linguateca.pt/cetenfolha/index_info.html

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

[3] https://shorturl.at/kANY4

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

[5] https://shorturl.at/moDHN

[6] https://www.linguateca.pt/cetenfolha/index_info.html

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

[8] https://shorturl.at/kANY4

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

[10] https://shorturl.at/moDHN

[11] https://www.linguateca.pt/cetenfolha/index_info.html

[12] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Domain	Metric	Result
Literature	Fleiss Kappa	0.23
	Fleiss Kappa W/o Und.	0.51
	Und. Rate	36%
Legal	Fleiss Kappa	0.46
	Fleiss Kappa W/o Und.	0.73
	Und. Rate	34%
Politics	Fleiss Kappa	0.78
	Fleiss Kappa W/o Und.	0.87
	Und. Rate	10%
Web	Fleiss Kappa	0.67
	Fleiss Kappa W/o Und.	0.84
	Und. Rate	20%
Social Media	Fleiss Kappa	0.53
	Fleiss Kappa W/o Und.	0.94
	Und. Rate	42%
Journalistic	Fleiss Kappa	0.72
	Fleiss Kappa W/o Und.	0.90
	Und. Rate	4%

Table 6: Extended per-domain analysis of the agreement between annotators. Fleiss Kappa W/o Und. measures Fleiss Kappa excluding undetermined documents.

hours in such conditions, and for BERT it takes approximately 52 hours to finish. The training in the all scenario, which took three hours for n-grams and approximately ten hours for BERT.

F Usage of AI Assistants

The authors have previously installed GitHub Copilot in its IDE. It was used to perform minor data manipulation operations when needed.

Domain	Variety	Split	Set	# Doc.	# Tokens
Literature	PT-PT	Train	-	20k	16M
		Test	Validation Set	2.5k	187k
			Platinum Set	21	1.4k
	PT-BR	Train	-	49k	31M
		Test	Validation Set	2.5k	161k
			Platinum Set	15	953
Legal	PT-PT	Train	-	29M	133M
		Test	Validation Set	500	24k
			Platinum Set	21	1k
	PT-BR	Train	-	4k	168k
		Test	Validation Set	500	22k
			Platinum Set	16	963
Politics	PT-PT	Train	-	25k	5M
		Test	Validation Set	500	98k
			Platinum Set	19	3.7k
	PT-BR	Train	-	626k	3k
		Test	Validation Set	500	103k
			Platinum Set	29	6.3k
Web	PT-PT	Train	-	41k	12M
		Test	Validation Set	5k	1.5M
			Platinum Set	17	5k
	PT-BR	Train	-	40k	12M
		Test	Validation Set	5k	1.4M
			Platinum Set	17	4.5k
Social Media	PT-PT	Train	-	18M	32M
		Test	Validation Set	500	9.3k
			Platinum Set	15	685
	PT-BR	Train	-	4k	65k
		Test	Validation Set	500	8k
			Platinum Set	13	231
Journalistic	PT-PT	Train	-	1.4M	177M
		Test	Validation Set	5k	655k
			Platinum Set	16	2.3k
	PT-BR	Train	-	307k	23M
		Test	Validation Set	5k	365k
			Platinum Set	20	2.7k
DSL-TL	PT-PT	Test	-	269	10k
	PT-BR	Test	-	588	23k
FRMT	PT-PT	Test	-	985	24k
	PT-BR	Test	-	985	24k

Table 7: Datasets split stats.