Across the Atlantic: Distinguishing Between European and Brazilian Portuguese Dialects

David Preda[1], Tomás Freitas Osório[1,2], Henrique Lopes Cardoso[1,2]

[1]Faculdade de Engenharia da Universidade do Porto (FEUP)
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal
[2]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC)
up201904726@up.pt, tomas.s.osorio@gmail.com, hlc@fe.up.pt

Abstract

Dialect Identification is the task of determining the regional or social variety of a spoken or written language. While specific languages have received considerable attention in this regard, others, such as Portuguese, remain largely unexplored. Furthermore, previous works on the Portuguese language are often outdated in the rapidly evolving landscape of NLP, and many suffer from methodological flaws. We revisit the task of differentiating between European and Brazilian variants of Portuguese, addressing and rectifying the mistakes found in prior research. For that, we carefully select a parallel corpus and explore both feature-based traditional classifiers and state-of-the-art neural approaches. Our findings[1] demonstrate that whereas Transformer-based models provide solutions that are robust to out-of-distribution data, traditional NLP techniques are still competitive in this task.

1 Introduction

Dialect identification (DI) is crucial for enhancing language processing tasks, enabling a better understanding of regional and social variations in communication – an essential aspect in computational sociolinguistics (Nguyen et al., 2016). These variations can range from subtle grammar changes to the same word having entirely different meanings, which may imply a different appropriate social setting. Therefore, NLP applications must be aware of the regional variety of the language they work with. Several tasks have been created to encourage the development of systems capable of handling these tasks, such as the Discriminating between Similar Languages (DSL) shared task organized under the Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial) (Aepli et al., 2023) or the Nuanced Arabic Dialect Identification (NADI) shared task (Abdul-Mageed et al., 2023). These tasks cover a few languages and a limited set of dialects.

Some work has been done on Portuguese DI. Existing work in linguistics details grammatical differences between European (PT-PT) and Brazilian (PT-BR) Portuguese variants (Mattos e Silva, 2013; Rio-Torto et al., 2022). The task of distinguishing between what are arguably the most economically relevant variants of Portuguese has received some attention, including in the DSL shared tasks (Zampieri et al., 2014; Aepli et al., 2023). However, some of the past approaches to DI in Portuguese suffer from methodological flaws. For instance, Zampieri and Gebre (2012) mention how entity names influence their models, thus deviating from DI through spurious correlations in the training data, which affects model generalization. On the other hand, the corpus collection used in DSL shared tasks (Tan et al., 2014) reveals some issues with the quality of the samples, namely their size, provenance, and label quality (Zampieri et al., 2023). By revisiting this problem, we intend to refine good practices for training DI models through careful data selection.

Our research strives to explore the task of Portuguese Dialect Identification (PDI) further. To accomplish this, we assess the performance of modern NLP techniques against classical methods. This is especially pertinent in light of the rapid advancements in NLP techniques. Additionally, we explore text length variability during training and evaluation, aiming to uncover its influence on model performance. By addressing these two critical aspects, we contribute valuable insights to the field and encourage others to join us in enhancing PDI.

Our contributions can be summarized as follows:

- We define a non-exhaustive set of useful features to distinguish European Portuguese from Brazilian Portuguese.
- We explore how different approaches perform

[1]Code and results available at: https://github.com/dtpreda/ata-portuguese-di

in PDI, investigating whether traditional NLP techniques do a good job compared to state-of-the-art Transformer-based models.

- We analyze how text length variability influences PDI models' performance.
- We provide robust state-of-the-art neural approaches for PDI.

2 Related Work

Language identification has been heavily studied (Jauhiainen et al., 2019), as it is particularly significant in our multilingual digital landscape, where diverse languages and dialects coexist. Identifying the employed language (Bender, 2011) facilitates effective communication and enhances the performance of language-dependent applications. Dialect identification can be seen as a particular case of language identification (Franco-Salvador et al., 2017). An example of this closeness is the work by Ljubesic et al. (2007) in distinguishing Croatian from other Slavic languages, which contain a high degree of lexical overlap.

The DSL shared tasks started in 2014 (Zampieri et al., 2014; Tan et al., 2014), with 13 languages and varieties divided into six groups. One of the groups is composed of the PT-PT and PT-BR Portuguese variants. The task has seen four editions (Zampieri et al., 2014, 2015; Malmasi et al., 2016; Zampieri et al., 2017) using the DSL corpus collection (composed of short excerpts of newspaper texts), which has also evolved to cover other languages.

In the first edition (Zampieri et al., 2014), the best system used a two-step classification approach: first predicting the language group using a Naive Bayes classifier and then discriminating between varieties within the chosen group using an SVM classifier. Most systems used words and character n-grams as features, while some have also explored using lists of words exclusive to a particular language or variety. Although the task included an open submission track where systems were allowed to be trained using data from outside the DSL collection, those that did ended up performing worse than the closed track submissions. In the second edition (Zampieri et al., 2015), the organizers included an additional test set, where capitalized named entities were replaced by placeholders, to avoid topic bias in classification while evaluating the influence of proper names in the classifiers' performance. The best-performing system was based on an ensemble of SVM classifiers, using word unigrams and bigrams and character n-grams as features. In this edition, the organizers conjectured that it would be relevant to analyze the influence of text length on the classification performance. In the third edition of the task (Malmasi et al., 2016), the organizers created two out-of-domain test sets, based on Twitter posts, for a subset of the languages to assess further the ability of the participating models to generalize. As before, most systems used standard word and character n-gram features and standard classifiers such as SVM and logistic regression. Some participants used neural network-based approaches, which did not turn out to be competitive. The fourth edition (Zampieri et al., 2017) followed previous trends (Medvedeva et al., 2017). The winning participant used an SVM-based two-step approach for classification and relied on BM25 weighting for feature representation, which was found to work better than TF-IDF. It has also added features such as the proportion of capitalized letters, punctuation marks, and POS tags modeled as n-grams for Latin languages such as French, Portuguese, and Spanish.

Acknowledging problems with the DSL corpus collection (namely issues with sample sizes, provenance, and label quality), a 2023 edition of DSL used a human-annotated corpus (Aepli et al., 2023; Zampieri et al., 2023). However, this new dataset adds a layer of complexity, as it includes an additional "neutral" label for cases where a text excerpt does not present enough information for discriminating between two similar languages or varieties. As an outcome, most participating systems have fallen below the provided baselines.

Some shared tasks have focused on a larger number of dialects within a language, such as for Arabic (Malmasi et al., 2016), German (Zampieri et al., 2017), Italian or French Aepli et al. (2022). The NADI shared task (Abdul-Mageed et al., 2020) aimed to address the complexity of Arabic, a language with diverse dialects and language variants, some of which lack mutual intelligibility. Despite its linguistic diversity, Arabic is often erroneously treated as a single, unified language. Some works in these tasks have focused on Transformer-based models (Camposampiero et al., 2022; Martin et al., 2020; Shammary et al., 2022; Khered et al., 2022), with some of these approaches reaching the best performances on the leaderboard.

Specifically targeting Portuguese, two salient works have explored the differences between PT-

PT and PT-BR. Marujo et al. (2011) translate between the two dialects. Zampieri and Gebre (2012) use character and word n-gram models to classify texts into PT-PT or PT-BR accurately. However, potential bias was noted due to the choice of data, as the authors have used two distinct journalistic corpora, one from texts published in 2004 by the *Folha de São Paulo* newspaper for Brazilian Portuguese and the other from texts published in 2007 by *Diário de Notícias* for European Portuguese.

An important issue to consider in PDI is the coming into force in 2009 of the Portuguese Language Orthographic Agreement (Ricardo, 2009) in both Portugal and Brazil. This spelling reform has the potential to significantly impact the few prior works done for PDI, given its effect on unifying orthography in the Portuguese language.

3 Dataset

The dataset choice for dialect identification is of utmost importance – a careless choice may lead to a biased model, predicting something other than the dialect. Zampieri and Gebre (2012) kickstarted the development of PDI, but the authors mention that region-specific entity names easily influence the model. This is due to the models being trained on local newspapers from different time periods without masking any content that may flag which newspaper the text comes from.

Furthermore, in the same way a model may tie dialects with entity names, it can also associate writing styles, genres or topics with each class. For example, if one of the dialects is represented by a set of medical texts while others focus on sports news, the model may deviate from its intended purpose and distinguish between themes instead.

To avoid these issues, one should rely on comparable corpora (Zanettin, 2014) containing documents that share some thematic or topical similarity while being produced in different languages. However, obtaining such corpora for different language variants is hard, as ensuring that documents within comparable corpora share thematic or topical similarity requires careful curation to create a meaningful and coherent collection. To circumvent this problem, we rely instead on a parallel corpus containing the same text translated into various languages and dialects. Note that a parallel corpus can also be seen as a comparable one, even though different versions of the same text are actually translations instead of being natively created in different languages. Tiedemann and Thottingal (2020) collect and maintain parallel corpora with several different topics, genres, and formats. In particular, we focus on the *Ted Talks 2020* (TED2020) dataset (Reimers and Gurevych, 2020), which contains a crawl of nearly 4,000 TED and TED-X transcripts both in European (PT-PT) and Brazilian Portuguese (PT-BR). This allows us to focus solely on the differences between dialects instead of getting other aspects of the text mixed up during training.

3.1 Data Preparation

We gathered the first 2,000 samples from the original TED2020 dataset. However, to investigate the impact of varying text length on model performance, we created three different versions of the dataset: (1S) transcripts are split at a sentence level; (4S) transcripts are split into groups of 4 sentences; (FT) original unsplit form (full transcripts). While allowing us to increase the amount of data, this multi-faceted approach will enable us to draw meaningful conclusions about the effectiveness of our models under various text length conditions.

If the samples are too short (particularly at a sentence level), insufficient information will be available to distinguish between the dialects. Therefore, for each version, we group the instances into bins according to their size, and a threshold is set so that most instances with lengths smaller than that of the most common bin (the mode) are removed. Ultimately, we filter out samples with less than 10, 40, and 500 characters for the 1S, 4S, and FT versions, respectively. Afterwards, a quality filter is passed through the data, removing entries containing special characters. Furthermore, identical entry pairs from different dialects were removed (these are likely to occur in sentence-level splits, given the high similarity between PT-PT and PT-BR).

We split each dataset into a 60:20:20 train/dev/test split. Table 1 shows the final composition of all three dataset versions – the number of samples per class slightly differs due to the quality filters.

3.2 Morphosyntactic Features

Finally, we run Part-Of-Speech (POS) tagging on all samples, to incorporate POS tags as features during training. We use a POS tagger[2] trained on the Mac-Morpho (Fonseca et al., 2015) corpus. We default to a single tagger for two different reasons.

[2] Available at `https://github.com/inoueMashuu/POS-tagger-portuguese-nltk`

Version	Train		Dev		Test	
	PT-PT	PT-BR	PT-PT	PT-BR	PT-PT	PT-BR
Single Sentence (1S)	84719	85759	22219	22324	24129	23952
4 Sentences (4S)	26523	26793	6913	6856	7454	7324
Full Transcript (FT)	914	905	337	297	355	304

Table 1: Dataset composition (number of samples) after data preparation.

Firstly, using a tagger per language may imply the usage of two different tagsets, which would introduce unwanted bias into the data. Secondly, even if the tagset was the same for all taggers, we have no way of knowing which dialect we are dealing with at test time, and we would be unable to decide on one tagger over the other.

4 Feature-Based Approaches

We begin exploring PDI through feature-based models. Based on previous works on Portuguese variant conversion (Marujo et al., 2011) and on a compilation of representative linguistic aspects that characterize the differences between PT-PT and PT-BR (Rio-Torto et al., 2022), we developed a set of handcrafted features, which we present in Table 2. It is worth noting that vocabulary-based features are non-exhaustive, as there are many vocabulary differences between the dialects, and, to the best of our knowledge, a readily available list with corresponding word pairs does not exist.

We present the results for our first models in Table 3. The macro-F1 score is used as it gives a better picture of how the model is handling both classes, and it is used extensively in DI literature (Jauhiainen et al., 2022a, 2021; Bayrak and Issifu, 2022). We opt for exploring Naive-Bayes (NB) as it is reported to have a good performance on dialect identification shared tasks for other languages, in particular, European Romance languages (Jauhiainen et al., 2022a, 2021), more similar to Portuguese. Furthermore, we also train Logistic Regression (LR) classifiers, which have also been reported as suitable for DI (Camposampiero et al., 2022). Albeit the crudeness of the features and the simplicity of the models, the results are promising, especially for longer samples, where the repetitive occurrence of the crafted features allows the models to learn the distinction between classes despite having a smaller number of examples. As these models are feature-based, with most features relying on grammar (thus being context-agnostic), we believe them to be good baselines for later models.

A question that might arise when looking at Table 3 is whether better results for longer texts are due to the model's performance or the nature of the dev set being evaluated. In other words, how differently will the models perform when provided with texts of varying lengths? To investigate this, we evaluate models for each combination of train and dev sets. The results obtained are shown in Table 4. The differences are, in fact, primarily due to the text length in the validation set. It is interesting to observe that training on longer text leads to only marginally better results.

The results of feature-based approaches in the TED2020 test sets are included in Table 9 of the Appendix.

5 N-Gram-based Models

As done in works for DI in other languages (Camposampiero et al., 2022; Jauhiainen et al., 2022a), we explore word-level n-grams in conjunction with shallow NLP techniques. We conducted an investigation into how increasing the n-gram count influences the results while reanalyzing the impact of variations in text length. At the same time, we also explore how POS tags can help these classifiers achieve better performance.

Our experimentsrevealed that, for most cases, bigrams report better performance than any other n-gram count. In Table 5, we report the results for all models trained on bigrams. It is worth noting that the features passed to each classifier are simple word counts with a limit of 10,000 features.

As in Camposampiero et al. (2022), Logistic Regression reports the best results, especially with the help of POS tags. However, contrary to feature-based model results in Table 4, training with shorter text obtains slightly better results. It is, therefore, uncertain which option is more suitable as a general rule. Still, similar to Table 4, longer texts in the validation set lead to better results.

The results of bigram-based approaches in the TED2020 test sets are included in Table 10 of the Appendix.

Name	Description	Pearson Correlation with Label (Training set)		
		1S	4S	FT
pt_pt_pronoun_ position_hints_bool	PT-PT pronoun-based hints, in the format *verb-personal_pronoun*	0.191	0.338	0.352
pt_pt_pronoun_position_hints		0.185	0.321	0.641
a_plus_infinitive_count_bool	PT-PT verb-based hints: preposition *a* followed by an infinitive verb	0.174	0.281	0.175
a_plus_infinitive_count		0.171	0.280	0.586
count_article_before _possessive_pronoun_bool	PT-PT article based hints, verifying the presence of an article before a possessive pronoun	0.125	0.213	0.453
count_article_before _possessive_pronoun		0.122	0.204	0.523
count_portuguese_words	PT-PT vocabulary-based hints, detecting PT-PT specific words	0.060	0.099	0.358
pt_pt_second_ person_hints_bool	PT-PT vocabulary-based hints, verifying the use of typical PT-PT personal and possessive pronouns	0.039	0.060	0.036
pt_pt_second_ person_hints		0.038	0.057	0.098
count_acute_accent	Count of acute accents, typically more frequent in PT-PT	0.018	0.026	0.020
count_uncontracted_ words_bool	Count of uncontracted prepositions, typically more frequent in PT-BR	-0.017	-0.028	-0.044
count_uncontracted_words		-0.016	-0.074	-0.106
count_brazilian_words	PT-BR vocabulary-based hints, detecting PT-BR specific words	-0.043	-0.074	-0.269
count_circumflex_accent	Count of acute accents, typically more frequent in PT-BR	-0.148	-0.234	-0.400
pt_br_pronoun_position_ hints_bool	PT-BR pronoun-based hints, in the format *personal_pronoun verb*	-0.164	-0.203	–*
pt_br_pronoun_position_hints		-0.175	-0.286	-0.423
pt_br_second_ person_hints_bool	PT-BR vocabulary-based hints, verifying the use of typical PT-BR personal and possessive pronouns	-0.172	-0.260	-0.174
pt_br_second_ person_hints		-0.170	-0.264	-0.488
gerund_count_bool	PT-BR verb-based hints, gerund verbs, detected by *ndo* end of word	-0.207	-0.343	-0.229
gerund_count		-0.195	-0.321	-0.643

Table 2: Full list of features for distinguishing PT-PT (positive class, label=1) from PT-BR (negative class, label=0). Suffix *_bool* refers to a flag that signals the presence of the feature. *Missing due to an unknown error during calculation.

Dataset	NB	LR
1S	0.650	0.671
4S	0.772	0.778
FT	0.965	0.976

Table 3: Macro-F1 scores for feature-based models on dev sets. NB = Naive Bayes, LR = Logistic Regression

Train Set	Dev Set	NB	LR
1S	1S	0.650	0.671
1S	4S	0.775	0.769
1S	FT	0.964	0.972
4S	1S	0.649	0.690
4S	4S	0.772	**0.778**
4S	FT	**0.971**	0.972
FT	1S	**0.683**	**0.692**
FT	4S	**0.778**	0.762
FT	FT	0.965	**0.976**

Table 4: Macro-F1 scores for all combinations for feature-based models on the dev sets. Values in bold are the best for each dev set and classifier type.

5.1 Adaptive Naive-Bayes

Jauhianien et al. (Jauhiainen et al., 2021, 2022b,a) have shown promising results with European Languages using an adaptive version of Naive-Bayes (ANB). Instead of starting with a new model and train it with the available data, this method begins with a pre-trained model. In Jauhiainen et al. (2021), the authors start with another of their NB approaches as the base model. The training data is divided into n fractions. Then, for each fraction, the top k samples for which the model is more confident are used to continue training the model. In this context, confidence is the difference between the probabilities of the sample belonging to one class or the other. A simple threshold α defines whether the model is confident about an example. This process is repeated for all fractions until one of two conditions is met: all samples within the fraction have been processed, or a maximum number i of iterations has been reached. In this approach, α, n, k, and i are hyper-parameters of the model.

We adapt this method to our needs and resources – we start from simpler models trained on a subset of the data, and we do not fine-tune the algorithm parameters (such as the number of iterations or the fixed size fraction of lines with the highest score). We restrict our experiments to only the 4S and FT versions of the datasets due to the computational demand in running this algorithm for the 1S versions. For all models, we set n to one-tenth of the size of each split and experiment with i equal to 4 and 10. We report our top 3 results for each dataset version and split size combination in Table 6. Once again, we focus on bigrams, which perform better than other n-gram counts.

Although the difference in performance is notable when varying the number of iterations for the 4S version, we observe no significant improvement compared to the results in Table 5.

The results of ANB-based approaches in the TED2020 test sets are included in Table 11 of the Appendix.

6 Transformer-Based Models

Following recent trends in efficiently fine-tuning Transformer-based models, we perform low-ranked adaptations (Hu et al., 2022) on Albertina (Rodrigues et al., 2023), a DeBERTa V2 base model (He et al., 2021) pre-trained on Brazilian or European Portuguese text. A linear layer is stacked on top of the model, converting it to a binary classifier that is then fine-tuned for PDI.

Low-ranked adaptations (LoRA) is a method to enhance the efficiency of language models customized for specific tasks by reducing the number of training parameters while surpassing the performance of other fine-tuning techniques. This is achieved by freezing pre-trained model weights and incorporating two additional weight matrices for task-specific adaptation. After training, these weights can be combined with the frozen weights, eliminating latency during inference and providing a significant advantage over alternative low-rank adapters (Houlsby et al., 2019; mahabadi et al., 2021; He et al., 2022).

We use the 4S version dataset (taking the FT version would surpass the model's max input length while the 1S version would contain too little information). We train the models for ten epochs with a batch size of 8, a maximum context length of 128, and the following hyper-parameters for low-rank adaptation: r = 8, alpha = 32, dropout = 0.05, learning rate = 2×10^{-5}, weight decay = 0.05.

The scores shown in Table 7 are from the checkpoint with the highest macro-F1 score on the validation set. Despite beating all other models for identical data setups (that is, compared with the models for the 4S train / test sets in Table 10), the edge provided by these models is negligible if we

Train Set	Dev Set	NB	LR	NB-POS	LR-POS
1S	1S	0.784	0.794	0.801	**0.818**
1S	4S	0.908	0.926	0.924	**0.945**
1S	FT	0.996	**1.0**	0.996	**1.0**
4S	1S	0.785	0.774	**0.801**	0.790
4S	4S	0.907	0.907	0.923	**0.927**
4S	FT	0.994	**1.0**	0.996	**1.0**
FT	1S	0.783	0.701	**0.797**	0.690
FT	4S	0.903	0.800	**0.921**	0.806
FT	FT	0.994	0.988	**0.996**	0.988

Table 5: Macro-F1 scores for bigram-based models on the dev set. Values in bold are the best for each train-dev pair.

Dataset	#Splits	#Iter	ANB	ANB-POS
4S	2	4	0.854	**0.887**
4S	4	4	0.813	0.857
4S	8	4	0.792	0.835
4S	2	10	0.907	**0.923**
4S	4	10	0.907	**0.923**
4S	8	10	0.908	**0.923**
FT	2	4	0.991	**0.996**
FT	4	4	0.991	0.993
FT	8	4	0.991	0.994
FT	2	10	0.991	**0.996**
FT	4	10	**0.996**	**0.996**
FT	8	10	0.993	**0.996**

Table 6: Macro-F1 scores for the bigram-based ANB models on the dev set. Values in bold represent the best score for each train/dev set and number of iterations.

take into account their computational requirements.

Model	Train/Test Set	Macro-F1
Albertina PT-PT	4S	0.936
Albertina PT-BR	4S	0.938

Table 7: Macro-F1 scores for the fine-tuned Albertina with LoRA models on the test set.

7 Cross-Dataset Analysis

Despite our satisfactory results, we have only worked within the closed domain of a parallel corpus on TED talks. A good PDI model should be able to exhibit equally good cross-dataset performance. To assess that, we evaluate our best-performing models against out-of-distribution corpora. We pick two distinct datasets whose examples we feed to any of our models as full transcripts.

7.1 Folha de São Paulo

We test our models against a *Folha de São Paulo* (FSP) dataset[3], which contains PT-BR news articles from between 2015 and 2017. After filtering out samples with less than 200 characters, we ended up with 2256 samples.

7.2 FEUP news corpus

To obtain a similar out-of-distribution corpus for PT-PT, we sampled articles from the *FEUP news corpus*[4], which contains articles from several Portuguese media channels, namely newspapers, from 2016. Again, we filtered out samples with less than 200 characters and sampled 2256 news articles.

7.3 Results

We show cross-dataset results for our models in Table 8. For feature-based models, we pick those trained on the FT data versions (Table 9 shows a best overall performance in this setup). As for bigram-based models (see Table 10), those trained on the 1S data versions seem to have a slight edge.

Feature-based models exhibit a considerable drop in performance, comparing the results for feature-based approaches using FT for both train and test sets (last line in Table 9) with those obtained here. This is also the case for bigram models for the FSP dataset, comparing the excellent results relying on 1S train and FT test datasets (third line in Table 10) with those for FSP using these models. For the FEUP News Corpus, on the other hand, the classifiers remain very competent. In fact, the LR bigrams model stands out as the one with the highest Macro-F1 score in cross-dataset results. We

[3] https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol

[4] https://hdl.handle.net/21.11129/0000-000D-F8C2-0

Model	FSP	FEUP News Corpus	Macro-F1
NB feature-based	0.661	0.792	0.720
LR feature-based	0.829	0.700	0.766
NB bigrams	0.747	0.968	0.847
LR bigrams	0.894	0.952	**0.920**
NB-POS bigrams	0.634	0.982	0.789
LR-POS bigrams	0.840	0.968	0.898
Albertina PT-PT	0.723	**0.990**	0.854
Albertina PT-BR	**0.938**	0.712	0.823

Table 8: Cross-dataset results (accuracy for each corpus, Macro-F1 for the joint corpus). Feature-based models were trained on the TED2020 FT dataset, bigram-based ones on the 1S, and Albertina-based ones on the 4S version.

leave a further analysis of the different accuracy scores in both datasets for future work.

By comparing the results of Albertina-based models (Table 7) with cross-dataset results, we observe they generalize well to out-of-domain data for a corpus in the same language variant: Albertina PT-PT generalizes well to the FEUP News Corpus, while Albertina PT-BR generalizes well to FSP.

8 Conclusion

We revisit the problem of dialect identification and attempt to bring attention to this task for the Portuguese language, which has been underexplored in this regard. We address the issue by following good practices when choosing the training data for PDI models. Differences between the European and Brazilian dialects of Portuguese were compiled into a non-exhaustive, comprehensive list of features, which is one of this work's contributions.

In line with previous works for Romance languages (Camposampiero et al., 2022), we find traditional techniques to work reasonably well for PDI. Transformer-based models seem to be robust for out-of-domain data. However, the best performance was obtained using simple representation techniques and a traditional classifier.

Lastly, we would like to encourage others to work on PDI. According to the Community of Portuguese-speaking Countries[5], nine countries have Portuguese as (one of) their official language: Angola, Brazil, Cape Verde, East Timor, Equatorial Guinea, Guinea Bissau, Mozambique, Portugal, and São Tomé and Príncipe. As such, PDI goes well beyond distinguishing between the variants addressed in this paper.

[5] https://www.cplp.org/

Acknowledgements

This work was financially supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC).

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, Chiyu Zhang, El Moatez Billah Nagoudi, Houda Bouamor, and Nizar Habash. 2023. NADI 2023: The Fourth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of ArabicNLP 2023*, pages 600–613, Singapore (Hybrid). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Giyaseddin Bayrak and Abdul Majeed Issifu. 2022. Domain-adapted BERT-based models for nuanced Arabic dialect identification and tweet sentiment analysis. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 425–430, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Emily M. Bender. 2011. On Achieving and Evaluating Language-Independence in NLP. *Linguistic Issues in Language Technology*, 6.

Giacomo Camposampiero, Quynh Anh Nguyen, and Francesco Di Stefano. 2022. The curious case of logistic regression for Italian languages and dialects identification. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 86–98, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Erick R Fonseca, João Luís G Rosa, and Sandra Maria Aluísio. 2015. Evaluating word embeddings and a revised corpus for part-of-speech tagging in Portuguese. *Journal of the Brazilian Computer Society*, 21:1–14.

Marc Franco-Salvador, Greg Kondrak, and Paolo Rosso. 2017. Bridging the Native Language and Language Variety Identification Tasks. *Procedia Computer Science*, 112:1554–1561. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a Unified View of Parameter-Efficient Transfer Learning. In *International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2021. Naive Bayes-based experiments in Romanian dialect identification. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 76–83, Kiyv, Ukraine. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022a. 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.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022b. Optimizing naive Bayes for Arabic dialect identification. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 409–414, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic Language Identification in Texts: A Survey. *J. Artif. Int. Res.*, 65(1):675–682.

Abdullah Salem Khered, Ingy Yasser Hassan Abdou Abdelhalim, and Riza Batista-Navarro. 2022. Building an ensemble of transformer models for Arabic dialect classification and sentiment analysis. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 479–484, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Nikola Ljubesic, Nives Mikelic, and Damir Boras. 2007. Language Identification: How to Distinguish Similar Languages? In *2007 29th International Conference on Information Technology Interfaces*, pages 541–546.

Rabeeh Karimi mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers. In *Advances in Neural Information Processing Systems*.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Luis Marujo, Nuno Grazina, Tiago Luis, Wang Ling, Luisa Coheur, and Isabel Trancoso. 2011. BP2EP - adaptation of Brazilian Portuguese texts to European Portuguese. In *Proceedings of the 15th Annual Conference of the European Association for Machine Translation*, Leuven, Belgium. European Association for Machine Translation.

Rosa Virgínia Mattos e Silva. 2013. O Português do Brasil. In Eduardo B. Paiva Raposo, M. Fernanda Bacelar do Nascimento, M. Antónia Mota, M. Luisa Segura, and Amália Mendes, editors, *Gramática do Português*, volume I, pages 145–154. Fundação Calouste Gulbenkian, Lisboa.

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.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Survey: Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Maria Manuel Calvet Ricardo. 2009. Breve História do Acordo Ortográfico. *Revista Lusófona de Educação*, 13.

Graça Rio-Torto, Tânia Ferreira, Ana Guerra, Zuzana Greksakova, and Zhang Yunfeng. 2022. *Português brasileiro e português europeu: um diálogo de séculos*. Universidade Politécnica de Macau.

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-*. In *Progress in Artificial Intelligence*, volume 14115 of *LNAI*, pages 441–453, Cham. Springer Nature Switzerland.

Fouad Shammary, Yiyi Chen, Zsolt T Kardkovacs, Mehwish Alam, and Haithem Afli. 2022. TF-IDF or transformers for Arabic dialect identification? ITFLOWS participation in the NADI 2022 shared task. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 420–424, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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, Reykjavik, Iceland.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Marcos Zampieri and Binyam Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. In *Proceedings of KONVENS 2012*, pages 233–237. ÖGAI. Main track: poster presentations.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language Variety Identification with True Labels.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

Federico Zanettin. 2014. Corpora in Translation. In Juliane House, editor, *Translation: A Multidisciplinary Approach*, pages 178–199. Palgrave Macmillan UK, London.

A Appendix

We include the results for feature-based and n-gram-based models in the test sets.

Train Set	Test Set	NB	LR
1S	1S	0.648	0.668
1S	4S	0.764	0.758
1S	FT	0.947	0.959
4S	1S	0.645	0.686
4S	4S	0.762	0.765
4S	FT	0.953	0.960
FT	1S	**0.687**	**0.687**
FT	4S	**0.767**	0.758
FT	FT	0.944	**0.965**

Table 9: Macro-F1 scores for all combinations for feature-based models on the TED2020 test sets. Values in bold represent the best score for each test set.

Train Set	Test Set	NB	LR	NB-POS	LR-POS
1S	1S	0.781	0.790	0.799	**0.815**
1S	4S	0.905	0.925	0.920	**0.940**
1S	FT	0.999	0.999	0.996	**1.0**
4S	1S	0.781	0.772	0.798	0.787
4S	4S	0.904	0.904	0.920	0.923
4S	FT	0.997	0.999	0.996	0.996
FT	1S	0.779	0.694	0.795	0.685
FT	4S	0.900	0.789	0.914	0.685
FT	FT	0.997	0.987	0.994	0.990

Table 10: Macro-F1 scores for bigram-based models on the TED2020 test sets. Values in bold represent the best score for each test set.

Train/Test Set	# Splits	# Iterations	NB	NB-POS
4S	2	4	0.848	**0.882**
4S	4	4	0.809	0.845
4S	8	4	0.789	0.826
4S	2	10	0.902	0.919
4S	4	10	0.902	0.919
4S	8	10	0.904	**0.920**
FT	2	4	0.990	0.992
FT	4	4	**0.997**	0.986
FT	8	4	0.987	0.989
FT	2	10	**0.997**	0.994
FT	4	10	**0.997**	0.994
FT	8	10	**0.997**	0.994

Table 11: Macro-F1 scores for the bigram-based ANB models on the TED2020 test sets. Values in bold represent the best score for each test set and number of iterations.