Spanish Dialect Classification: A Comparative Study of Linguistically Tailored Features, Unigrams and BERT Embeddings

Laura Zeidler[◁,*] **Chris Jenkins**[*] **Filip Miletić**[*] **Sabine Schulte im Walde**[*]

[◁]CSAI Department, University of Technology Nuremberg, Germany
[*]Institute for Natural Language Processing, University of Stuttgart, Germany
laura.zeidler@utn.de {christopher.jenkins, filip.miletic, schulte}@ims.uni-stuttgart.de

Abstract

The task of automatic dialect classification is typically tackled using traditional machine-learning models with bag-of-words unigram features. We explore two alternative methods for distinguishing dialects across 20 Spanish-speaking countries: (i) Support vector machine and decision tree models were trained on dialectal features tailored to the Spanish dialects, combined with standard unigrams. (ii) A pre-trained BERT model was fine-tuned on the task. Results show that the tailored features generally did not have a positive impact on traditional model performance, but provide a salient way of representing dialects in a content-agnostic manner. The BERT model wins over traditional models but with only a tiny margin, while sacrificing explainability and interpretability.

1 Introduction

Dialects are often merely perceived as non-standard ways of expressing oneself. However, this simplistic view obscures the fact that dialects represent distinct language varieties which are clearly associated with specific geographic areas or groups of speakers (Trudgill, 2003) and therefore constitute a key part of a person's identity. Dialect use can reveal a lot about someone's background and we are constantly exposed to it in everyday life. For this reason, automatic dialect classification to improve non-standard representations and enhance performance on downstream tasks such as dialogue systems (e.g., in customer service applications) has become a vital NLP task. Differently to other NLP tasks, in automatic dialect classification simple traditional machine learning approaches like support vector machines (SVMs) remain competitive with transformer models (Chifu et al., 2024), presumably because transformers lack explicit knowledge of linguistic structures. Transformer models might therefore primarily rely on topic-related lexical cues (Zampieri et al., 2013), instead of focusing on linguistic characteristics.

Following this line of reasoning, we hypothesize that utilizing linguistic knowledge may be beneficial for dialect classification: We investigate the benefits of incorporating dialect-specific linguistically tailored features into machine learning classifiers using unigram features, and contrast them with a transformer-based model. We focus on Spanish, which exhibits strong variations in vocabulary and syntax across dialects, and has adequate resources available. We primarily leverage linguistic observations by Lipski (1994) to find potentially helpful dialect-specific characteristics in corpus data encompassing 20 Spanish dialects. Our classification task is therefore considerably more challenging than classification experiments in previous research, which only considered a handful of Spanish dialects (e.g. Zampieri et al., 2014, 2015; Chifu et al., 2024). The features are added to two unigram-based models, namely an SVM and a decision tree (DT) model, and compared to the models which only take individual feature types into account. Our contributions are as follows:[1]

1. We curate an extensive set of dialect-specific empirical features for the task of Spanish dialect classification.
2. We conduct a battery of classification experiments demonstrating that the linguistically tailored features do not enhance unigram-based models, but do provide a promising way of representing dialects in a content-agnostic manner.
3. We show that our transformer model only marginally outperforms traditional methods, raising the question whether this minor gain warrants sacrificing efficiency, interpretability, and explainability.

[1]Code and data can be found at: https://github.com/lurr98/spanish_variation

Label	Included Countries
ANT	Cuba, Dominican Rep., Panama, Puerto Rico
GC	Costa Rica, Guatemala
MCA	El Salvador, Honduras, Nicaragua
CV	Colombia, Venezuela
EP	Bolivia, Ecuador, Peru
AU	Argentina, Uruguay

Table 1: Mapping of country labels to more coarse-grained labels. CL, MX, PY and ES retain their own labels, so the total number of classes is 10.

2 Related Work

Variation in language poses considerable challenges for many NLP tasks, sparking growing interest in the field. Concerning the dialect classification task, interesting insights were obtained from early shared tasks on discriminating between similar languages (DSL) (Zampieri et al., 2014, 2015), where documents from different language varieties were classified. Top-performing models used SVM classifiers or ensembles, a trend that was also observed in later DSL tasks (Malmasi et al., 2016; Zampieri et al., 2017), suggesting that traditional classifiers tend to outperform neural networks on this task (Zampieri et al., 2020). Results from recent iterations, however, indicate that neither approach consistently dominates (Chifu et al., 2024).

Since much of previous work is based on feature-based classifiers, the choice of features is of great importance. Best performing models in the DSL tasks used word-based representations or character n-grams of higher order (Zampieri et al., 2020). Furthermore, some studies incorporated linguistically motivated features like POS tags, resulting in conflicting results about whether these features contribute positively to the model performance (Zampieri et al., 2013; Bestgen, 2017). Demszky et al. (2021) even manually selected dialect-specific features from linguistic literature to tackle the task of dialectal feature detection. These linguistic features are *tailored* to the specific dialects at hand.

3 Data

Our experiments on Spanish dialects rely on the *Web/Dialects* portion of the Corpus del Español (Davies, 2016). It contains texts from about two million web pages from 21 Spanish-speaking countries (>2B words). Table 4 in Appendix A shows an overview of the data by country.[2] The corpus consists of documents and is tokenized, lemmatized and POS-tagged. For pre-processing, we lowercased tokens and removed punctuation and digits. Due to a significant imbalance in number of documents per class, the data was balanced by randomly selecting from each class as many documents as the smallest class contains, such that every class is represented by an equal number of documents. The data was randomly split into train, development and test sets with a ratio of 80/10/10.

[2]We did not include the data extracted from US websites.

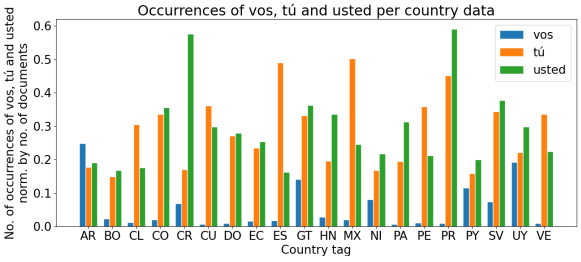

Figure 1: Distribution of *vos*, *tú* and *usted* in the corpus.

	Features	Counted Items
Frequent	CLITIC	clitics *lo*, *le* and *les*
	DIFFTENSE	14 different verbal tenses/aspects
	DIM	*-ito/a*, *-ico/a*, *-illo/a*, *-ingo/a*
	OVSUBJ	9 overtly realized subject pronouns
	SER_ESTAR	*ser* and *estar* for adjective predicates
	VOSEO	1) "familiar" pron.s (*vos*, *tú*, *usted*) 2) verbs of the *voseo* paradigm
	VOSOTROS	pronouns *vosotros* and *os*
Rare	ADA	productive nouns ending in *-ada*
	ARTPOSS	indef. article, poss. adj. and noun
	MASNEG	*más* preceding negative adjectives
	MUYISIMO	*muy* preceding *-ísimo*
	NONINV	non-inverted WH questions
	SUBJINF	subj. pronoun and infinitive/gerund

Table 2: Description of the tailored features.

4 Experimental Set-Up

We conducted three experiments: (i) We trained and tested the classifiers on the pre-processed, balanced data set. (ii) We replaced named entities (NEs) and nationalities (e.g. "peruano") with a placeholder and trained and tested the models on the altered data to reduce reliance on too obvious lexical cues, as noted for BOW models in prior research (Zampieri et al., 2013). (iii) We took a broader view on dialect classes by clustering countries belonging to a linguistic grouping of dialects according to Lipski (2012) (see Table 1), and training and testing the models with these new classes.

4.1 Models

We fine-tuned a pre-trained BERT model[3] on our data. For the feature-based models (SVM and DT)

[3]The model can be found on *huggingface* (Wolf et al., 2020): `dccuchile/bert-base-spanish-wwm-cased`.

Model	Features	Standard Classification		Named Entity Filter		Grouped Labels	
		Accuracy	Macro-F	Accuracy	Macro-F	Accuracy	Macro-F
SVM	Tailored	0.10	0.08	-	-	0.18	0.14
	Unigrams	0.65	0.65	0.55	0.54	**0.66**	**0.66**
	Both	0.65	0.65	0.55	0.55	**0.66**	**0.66**
DT	Tailored	0.09	0.09	-	-	0.15	0.15
	Unigrams	0.38	0.45	0.16	0.17	0.41	0.44
	Both	0.38	0.45	0.17	0.17	0.42	0.44
BERT	Embeddings	**0.67**	**0.67**	**0.59**	**0.59**	**0.66**	**0.66**

Table 3: Accuracy and Macro-F1 of all models on the test set in the initial experimental setup.

we used the machine learning library *scikit-learn* (Pedregosa et al., 2011). While transformers yield state-of-the-art performance in many NLP tasks, they are black-box methods which are computationally very expensive. In contrast, statistical models are more efficient as well as interpretable.

4.2 Features of the Statistical Models

Linguistically Tailored Features: Assuming that features that are tailored to the dialects at hand are beneficial to the models, we collected features with indicative morphological and syntactic characteristics from literature research (Lipski, 1994). For example: Pronoun usage varies across Spanish dialects, with "vos" replacing "tú" in some dialects (*voseo*), while others prefer the formal "usted" in familiar settings. Corresponding counts in our corpus capture these characteristics well (see Figure 1 for the above example), thus confirming linguistic assumptions from prior research and suggesting the usefulness of these features. The tailored features can be grouped into two categories: (i) features that model distributions of frequently occurring phenomena and (ii) features that count the occurrences of rare phenomena. In total, 13 features were extracted, they are listed in Table 2.

Unigram-based Features: Here, we pursued a simple BOW approach, using term frequencies (tf) by means of *scikit-learn*'s TfidfVectorizer class:

$$\text{tf}(t, D) = \frac{\#t_D}{\sum_{t' \in D} \#t'_D} \quad (1)$$

where $\#t_D$ is the frequency of a token t in a document D, divided by the total amount of tokens in the document (Manning et al., 2008). Only tokens that occur at least twice in the training data were considered. We ignored tokens corresponding to tailored features in order to clearly distinguish the informativeness of the two approaches.

Merged Features: We joined unigram-based and tailored features by normalizing the tailored feature vectors by the number of tokens in the document to match the tf scale and concatenating them with the corresponding unigram-based vectors.

4.3 Hyperparameter Choice

Hyperparameters for the traditional models were selected using *scikit-learn*'s GridSearchCV; results and best values are shown in Tables 5 and 7 in Appendix A. For the transformer, we limited epochs to 5 to keep runtime reasonable, and set batch size to 16 to avoid memory issues (Table 6 in Appendix A).

5 Results

Table 3 shows the results of the classification experiments, which are further discussed below.

5.1 Standard Classification

The BERT model achieves the best performance with an accuracy score of 0.67, closely followed by the SVM models (0.65) using purely unigram-based or merged features. The corresponding DT models lag behind with an accuracy of 0.38 in both settings. The tailored features perform much worse with scores around 0.1. While the confusion matrices of most models exhibit a typical diagonal, Figure 3 shows that the SVM model using tailored features mainly resorts to class ES (Spain), thus implying that this class exhibits characteristics that are distinct from all other dialects, which is supported by linguistic literature (Lipski, 1994). The DT model using solely BOW or merged features behaves similarly (Figure 4 in Appendix A).

To exploit the interpretability of the models, we calculate feature weights to get insights into the behavior of the models. Figure 2 shows the most

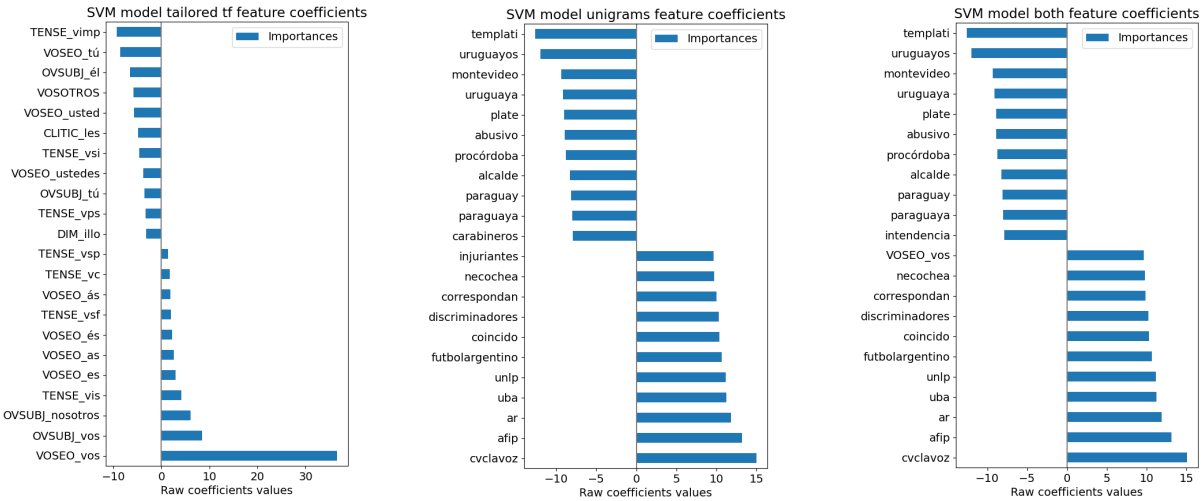

Figure 2: Feature relevance in SVM models: tailored, BOW and merged features

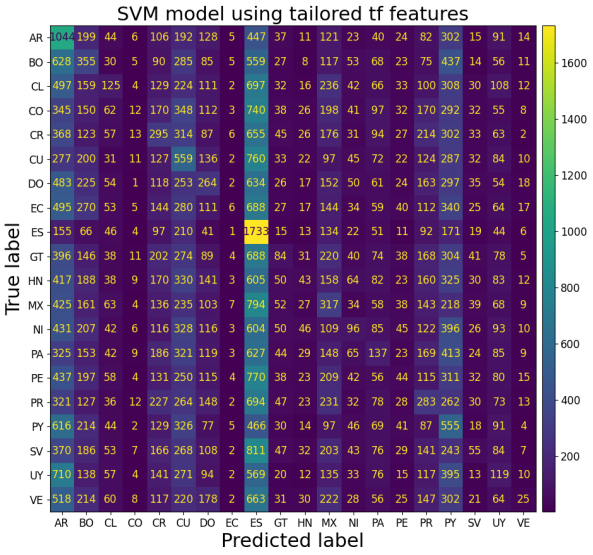

Figure 3: Confusion matrix (SVM, tailored features) over the predicted vs. true country labels.

important features of the SVM models using the three feature types, based on their coefficients. The weights indicate that the most important features of the SVM model only using tailored features display a high focus on tenses and VOSEO and OVSUBJ features. Generally, the most frequent features are also the most relevant ones, which is also true for the DT model. In unigram-based models, topic-related tokens (e.g. nationalities, places) dominate the importance rankings, which is consistent with prior research (Zampieri et al., 2013). The merged models exhibit similar rankings, while some tailored features like VOSEO$_{vos}$ appear among the most important ones (Figure 2). Given that these tokens would anyway occur as unigram features, the tailored features provide little extra benefit.

5.2 Effect of Named Entity Features

Table 3 shows that the overall performance drops significantly compared to the standard setup when NEs and nationalities are removed from the features. Again, the transformer model outperforms the other models with a score of 0.59. The accuracy of the SVM is the same for merged and unigram-based features (0.55). The DT results are again low, showing a slightly but significantly stronger performance (0.17>0.16) with merged features[4]. The fact that all models deteriorate on this task shows that they heavily rely on content-related textual cues. Now tailored features play a bigger role for the models using the merged feature set: More tailored features are among the most important ones in SVM and DT models (Figure 6 and 7 in Appendix A), such as indicative simple preterite tense. This confirms that the tailored features add explicit information to the models that can only be found implicitly in unigrams.

5.3 Effect of Grouped Dialects

When grouping dialects into larger classes, all statistical models show an increase in performance (Table 3), as expected due to the label reduction of 50%, which renders the task easier. The transformer model, however, deteriorates and is now on par with the unigram-based SVM model (accuracy score: 0.66). Although the performance is still comparably low, the models using tailored features almost double their accuracy from 0.10 to 0.18 (SVM), and from 0.09 to 0.15 (DT), while the unigram-based and merged features models only

[4]We measured statistical significance using the McNemar test (Seabold and Perktold, 2010) with a threshold of 0.05.

slightly increase their performances. These observations show that the change in inter-class similarity is clearly reflected by the models using tailored features, whereas it has little effect on the others, suggesting that the tailored features represent the dialectal differences in the language better than the standard BOW features.

5.4 Summary of Observations

Our results show that the traditional classifiers did not outperform the fine-tuned transformer model. Yet, it is important to note that the performance gap to the SVM models, while statistically significant, was marginal (at most 0.04 points) and in the case of the grouped dialects non-existent. Considering that SVMs have significantly shorter runtime than transformer models and are typically more interpretable and transparent, it is valid to question whether substituting slightly better performance for a more efficient, explainable and interpretable statistical model is reasonable.

The study of the features has revealed that the tailored features perform much worse than the other features and, with one exception, do not improve performance of the unigram-based features. However, the high scores produced by the other features and also the BERT model reflect a rather content-dependent classification, which is not necessarily desirable. In contrast, the tailored features by design model the dialects in a content agnostic manner and the grouping of the classes has revealed that they indeed reflect the inter-class similarity much better than the other methods. In this light, we argue that the use of tailored features is a promising approach that deserves to be explored further.

6 Conclusion

In this work, we tackled the task of automatic dialect classification for dialects from 20 Spanish-speaking countries. We compared two traditional machine learning models, an SVM and a DT model, to a fine-tuned BERT model and experimented with three types of features for the feature-based models: linguistically motivated dialect-specific features, BOW unigram features and a merged version. The traditional models could not outperform the transformer model. However, the margin to the best-performing SVM model was at most 0.04 points, which raises the question of whether this slight improvement in performance is worth sacrificing the efficiency, explainability and interpretability of traditional machine learning models. Regarding the features, the current tailored feature set generally did not contribute positively to the performance of the traditional models. Still, we demonstrated that they represent the dialects in a salient, content-agnostic manner, and thus carry an inherent potential to go beyond obvious lexical cues like BOW features and BERT embeddings, and to capture inter-class similarity for broader linguistic areas. Investigating the use of dialect-specific features therefore constitutes a promising approach.

7 Limitations

A current limitation which regards the tailored features is that – even after exhaustive literature search – they constitute a comparatively small feature set which moreover includes features that occur very rarely. For future work, finding more dialectal characteristics that occur with a relatively high frequency and thus building a larger feature set could improve the performance of the models using such a feature set. Also, some of the literature that was consulted for feature collection dates back to 1994 (Lipski, 1994) and, although very well-established, may not be fully representative of the current varieties that are spoken and written in Latin America. This issue may have contributed to the generally poor performance of the tailored features.

The focus of our paper is on comparing statistical vs. transformer-based classifiers, rather than identifying the single best transformer model. Nevertheless, it is worth noting that we do not know whether the Spanish BERT model we used was pre-trained on an appropriate amount of Latin American Spanish data. While we expect our fine-tuning procedure to compensate for any such shortcomings, it may still be relevant to experiment with other Spanish BERT models to better assess the effect of pre-training with different data mixes. Furthermore, implementing models from different families (e.g. GPT) could yield different results and presents an interesting direction for future work.

Finally, we observed that *spacy*'s built in NER model did not consistently recognize all NEs in the data. While we expect any effects to be roughly the same for all classes, future work could benefit from applying a more sophisticated NER model for Spanish. Also, it would be reasonable to remove other cues like country tags that are not directly targeted by NER tools.

References

Yves Bestgen. 2017. Improving the character ngram model for the DSL task with BM25 weighting and less frequently used feature sets. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 115–123, Valencia, Spain. Association for Computational Linguistics.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

Mark Davies. 2016. Corpus del español: Web/dialects.

Dorottya Demszky, Devyani Sharma, Jonathan H Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 19th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

John M. Lipski. 1994. *Latin American Spanish / John M. Lipski.* Longman linguistics library. Longman, London.

John M. Lipski. 2012. *Geographical and Social Varieties of Spanish: An Overview*, chapter 1. John Wiley & Sons, Ltd.

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.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Peter Trudgill. 2003. *A Glossary of Sociolinguistics*. Edinburgh University Press, Edinburgh.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Binyam Gebrekidan Gebre, and Sascha Diwersy. 2013. N-gram language models and POS distribution for the identification of Spanish varieties (Ngrammes et traits morphosyntaxiques pour la identification de variétés de l'espagnol) [in French]. In *Proceedings of TALN 2013 (Volume 2: Short Papers)*, pages 580–587, Les Sables d'Olonne, France. ATALA.

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, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

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.

A Appendix

Country	Country tag	# of Documents
Argentina	AR	177,920
Bolivia	BO	43,293
Chile	CL	71,620
Colombia	CO	184,970
Costa Rica	CR	33,255
Cuba	CU	51,708
Rep Dom	DO	47,065
Ecuador	EC	63,160
España	ES	421,520
Guatemala	GT	61,434
Honduras	HN	43,227
México	MX	286,275
Nicaragua	NI	35,696
Panamá	PA	29,312
Perú	PE	121,814
Puerto Rico	PR	33,879
Paraguay	PY	33,301
El Salvador	SV	38,217
Uruguay	UY	36,154
Venezuela	VE	112,571

Table 4: Overview of the number of documents in the Corpus del Español per country (Davies, 2016).

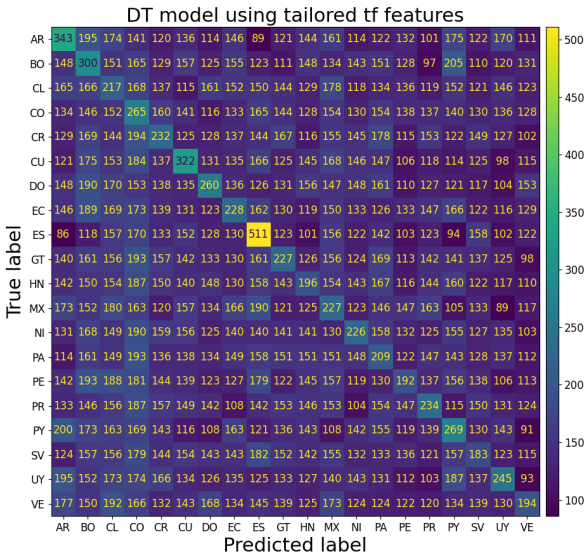

Figure 4: Confusion matrix of the DT model using tailored features.

C	Acc.	std
10	0.104	0.0010
0.1	0.094	0.0009
0.01	0.087	0.0009
0.001	0.080	0.0006

C	Acc.	std
10	0.637	0.0018
0.1	0.580	0.0019
0.01	0.496	0.0017
0.001	0.323	0.0015

Table 5: Accuracy and standard deviation results produced by SVM models using a different parameter value for C using GridSearchCV. The tables show the results for tailored (left) and unigram features (right).

Hyperparameter Name	Value
Number of epochs	5
Batch size per device during training	16
Number of warm-up steps for LR scheduler	500
Weight decay	0.01

Table 6: Hyperparameters of transformer models.

max_depth & max_features	Acc.	std
30_None	0.085	0.0002
50_None	0.085	0.0006
30_log2	0.083	0.0009
30_sqrt	0.083	0.0012
50_sqrt	0.083	0.0010
50_log2	0.082	0.0006

max_depth & max_features	Acc.	std
50_None	0.382	0.001
30_None	0.366	0.0018
50_sqrt	0.124	0.0105
30_sqrt	0.096	0.0056
50_log2	0.058	0.0012
30_log2	0.054	0.0009

Table 7: Accuracy and standard deviation results produced by DT models using different parameter combinations for `max_depth` & `max_features` using GridSearchCV. Left table uses tailored and right table unigram-based features.

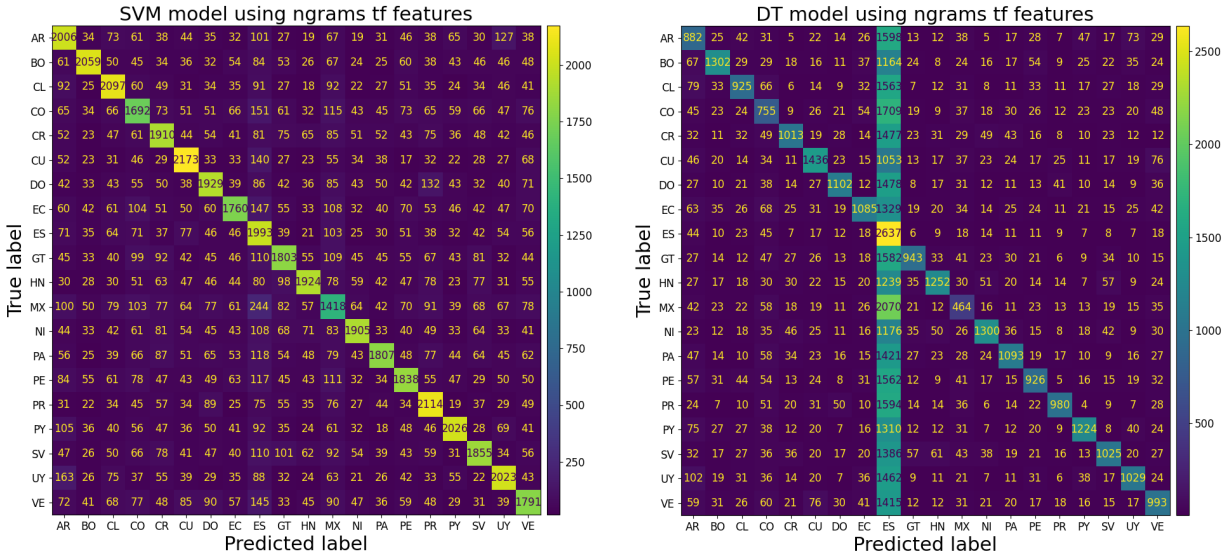

Figure 5: Confusion matrices of the SVM (left) and DT model (right) using BOW features.

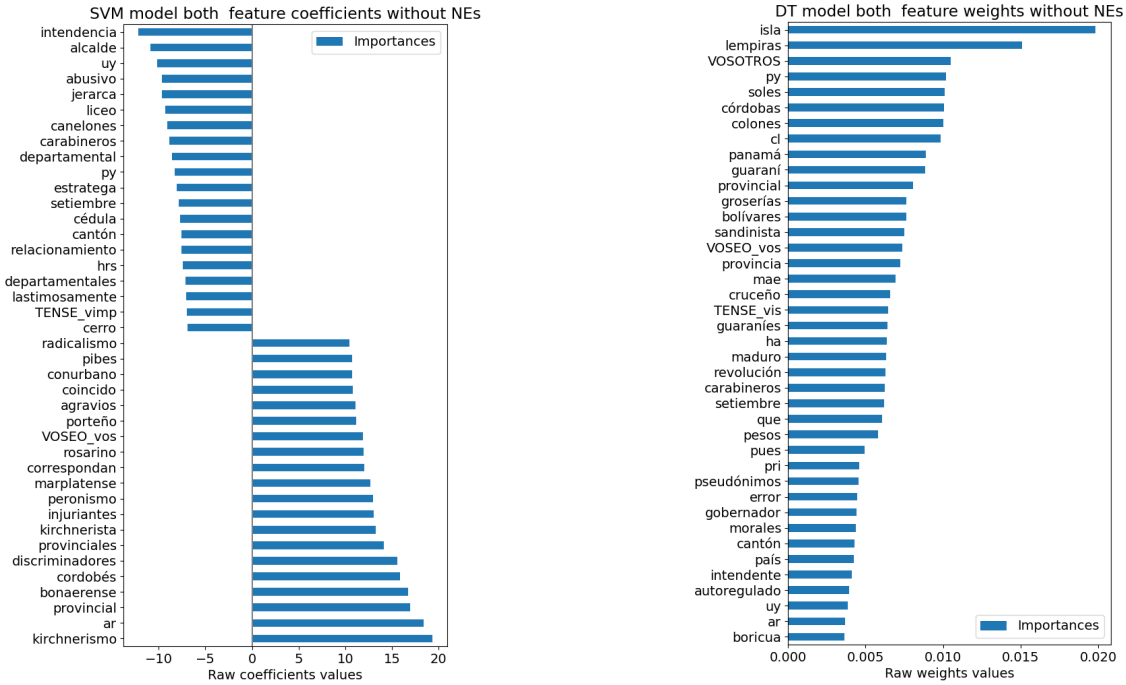

Figure 6: Feature relevance in SVM (left) and DT (right) models using merged features when NEs are filtered out.

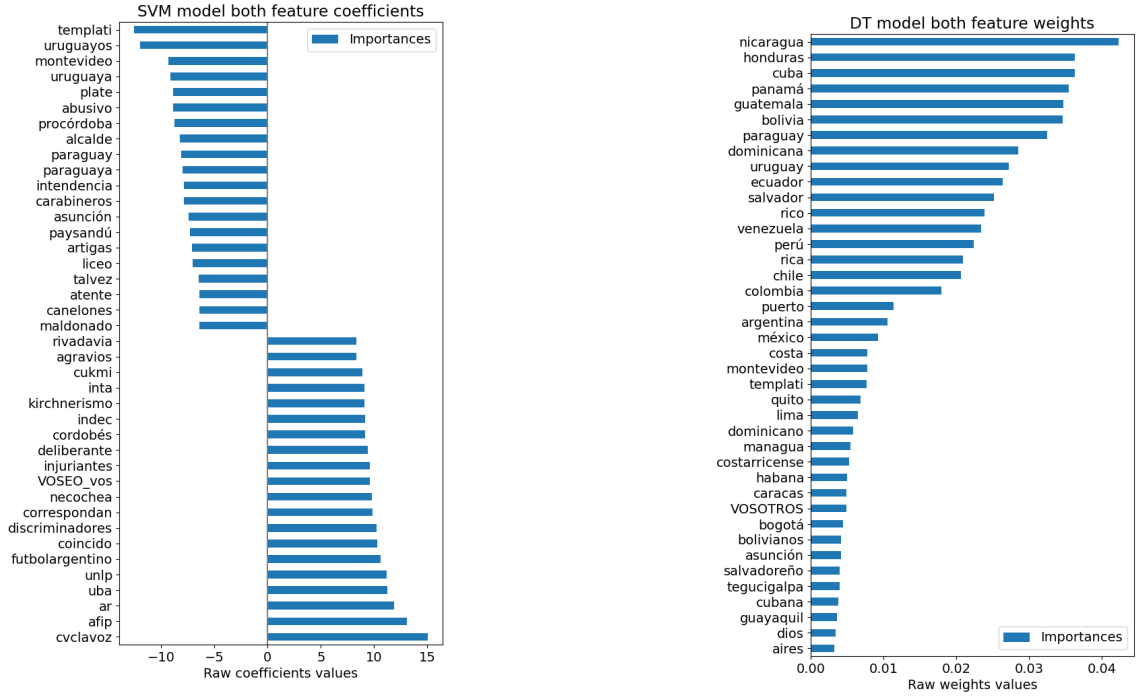

Figure 7: Feature relevance in SVM (left) and DT (right) models using merged features for comparison with Fig. 6.