Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection

Galo Castillo-López, Arij Riabi, Djamé Seddah

Analyzing Zero-Shot transfer Scenarios across Spanish variants for Hate Speech Detection

Galo Castillo-López[1,2*] and **Arij Riabi**[1,3] and **Djamé Seddah**[1]
[1]INRIA Paris, France
[2]Université Paris-Saclay, France
[3]Sorbonne Université
galo.castillo@upsaclay.fr
{firstname.lastname}@inria.fr

Abstract

Hate speech detection in online platforms has been widely studied in the past. Most of these works were conducted in English and a few rich-resource languages. Recent approaches tailored for low-resource languages have explored the interests of zero-shot cross-lingual transfer learning models in resource-scarce scenarios. However, languages variations between geolects such as American English and British English, Latin-American Spanish, and European Spanish is still a problem for NLP models that often relies on (latent) lexical information for their classification tasks. More importantly, the cultural aspect, crucial for hate speech detection, is often overlooked.

In this work, we present the results of a thorough analysis of hate speech detection models performance on different variants of Spanish, including a new hate speech toward immigrants Twitter data set we built to cover these variants. Using mBERT and Beto, a monolingual Spanish Bert-based language model, as the basis of our transfer learning architecture, our results indicate that hate speech detection models for a given Spanish variant are affected when different variations of such language are not considered. Hate speech expressions could vary from region to region where the same language is spoken.

1 Introduction

Hate speech detection is a task that has gained much attention from the NLP community due to the exponential spread of social media platforms[1]. This task aims to identify whether a piece of text contains hateful messages against a person or a group based on characteristics such as color, ethnicity, race, sexual orientation, religion, and others (John, 2000). Gender and nationalities are no exceptions to this. According to the Pew Research Center report in 2021, 33% of women under 35 report having been sexually harassed online, compared with 11% of men under 35 (Vogels, 2021). Misogyny is harm against women due to gender, which might result in psychological, reputational, professional, or even physical damage (Ging and Siapera, 2018). On the other hand, xenophobia is "*attitudes, prejudices, and behavior that reject, exclude and often vilify persons, based on the perception that they are outsiders or foreigners to the community, society or national identity*"[2]. An online manifestation of such behaviors may include hostility, social exclusion, threats of violence, and other forms of discrimination. As a result, the Internet becomes a less equal, less safe, and less inclusive environment for targetted groups.

Online hate speech detection in social medias platforms has been tackled in several studies (Pamungkas et al., 2018; García-Díaz et al., 2022; Pamungkas et al., 2020; Ahluwalia et al., 2018; Muaad et al., 2021; Shushkevich and Cardiff, 2018; Díaz-Torres et al., 2020). However, most studies have been carried out using English language data or limited Spanish data. For example, the significant morphosyntactic variations between Spanish variants (Bentivoglio and Sedano, 2011) make considering the Spanish language homogeneous challenging for language models. According to Ethnologue[3] in 2022, Spanish is currently declared as the official language in 22 countries, being the fourth language with the most significant number of countries. Due to the numerous regions where Spanish is the spoken language, expressions associated with hate speech may differ across various locations. For example, in the variation of the Spanish language from Spain, the word *"fregar"* only means *"scrub"* while at the same time, the same word in the Span-

*Work conducting during an internship at Inria Paris.

[1]Please be aware that this paper contains some examples of offensive slurs that may be considered upsetting.

[2]https://home-affairs.ec.europa.eu/pages/glossary/xenophobia_en

[3]https://www.ethnologue.com/ethnoblog/gary-simons/welcome-25th-edition

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 1–13

ish Ecuadorian variant can also mean *"annoy"* or *"having fun"*. We note the different connotations of that term across various Latin American regions provided by the Royal Spanish Academy[4] (RAE in Spanish). Thus, the phrase *"Anda a tu casa a fregar"* can only be interpreted as *"Go home to scrub (the dishes)"* by people from Spain, which can contain a misogynous connotation. On the other hand, for people speaking Spanish in Ecuador, it may be mostly interpreted as *"Go home to have some fun"* or *"Go home to annoy (other people)"*, which are not related to discriminatory connotations. Despite these scenarios, studies proposing models for hate speech detection towards women and immigrants in Spanish generally do not include information about the language or cultural variation of the text.

Due to the previously described challenge, other difficulties emerge when developing hateful content detection systems for online platforms.
Current state-of-the-art pre-trained language models (LM), such as the "multilingual" version of BERT (mBERT) (Devlin et al., 2018; Pires et al., 2019), are widely used in several NLP tasks and achieve impressive results. However, mBERT might not help to detect hate speech against women or immigrants when language-specific variants appear. It has been proven to perform worse than monolingual implementations of BERT under certain circumstances (Martin et al., 2020; Wu and Dredze, 2020). mBERT is trained on only Wikipedia data, particularly the entire Wikipedia dump for each language, excluding user and talk pages. However, this is problematic for the Spanish language as according to Wikipedia's *Spanish Wikipedia* article (Spanish Wikipedia, 2021) by September 2017, 39.2% of the Spanish Wikipedia edits come from Spain, being the country with the largest edits, while the rest come from other countries located in regions such as the Americas and others. It is important to note that Spain is the fourth country with the most prominent Spanish language native speakers, whereas Mexico is the first according to Statista (2021). Therefore, language models trained on Wikipedia data may not represent the differences between Spanish variants (Hershcovich et al., 2022). Thus, in this study, we aim to address the following research questions:

- **RQ1**: How does language-specific language models' performance differ from multilingual LM to detect online hate speech against women and immigrants in Spanish corpora?
- **RQ2**: Is zero-shot transfer effective for hate speech detection when different language variants of the same language are considered?

To do so, we compare mBERT with a Spanish version of BERT, named BETO (Canete et al., 2020), for binary classification in two different hate speech domains using various datasets on xenophobia and misogyny. We analyze the effects of Spanish language variants on model performance in both domains using a xenophobia detection corpus we created for this purpose as no other corpora include language variant metadata at the tweet level. Finally, an error analysis conducted with the SHAP interpretability framework (Lundberg and Lee, 2017) highlighted the *vulnerability* to cultural-specific hateful terms of language models fine-tuned on another geolect. In an era where cross-cultural issues in NLP become of increasing and welcome importance (Hovy and Yang, 2021; Nozza, 2021; Hershcovich et al., 2022) , our work and methodology constitute an interesting step in this process. This is why we release our datasets[5], models, and guidelines to the community, hoping to enrich a burgeoning ecosystem.

Our main contributions may be summarized as follows:

- The compilation and annotation of HaSCoSVa-2022, a new corpus of tweets related to hate speech towards immigrants written in Spanish. This corpus contains information regarding the language variant. The dataset is subdivided into two subsets according to the language variant: (1) Latin American and (2) European. The dataset is released to the research community.
- Experiments on zero-shot transfer between European and Latin American Spanish language variants on hate speech detection towards women and immigrants to investigate how the performance of the models vary when used on different variants of the same language.

2 Related Work

Automatic hate speech detection in online platforms has been previously studied across different hate speech domains such as misogyny (Fersini et al., 2022; Plaza-Del-Arco et al., 2020), xenopho-

[4] https://dle.rae.es/fregar

[5] https://gitlab.inria.fr/counter/HaSCoSVa

bia (Romero-Vega et al., 2020; Benitez-Andrades et al., 2022), homophobia (Karayiğit et al., 2022; Arcila-Calderón et al., 2021) and others (Davidson et al., 2017; Lozano et al., 2017). Nevertheless, a limited number of works focus on Spanish data to develop ML-based systems for online hateful content identification.[6] Most of the research developed with Spanish corpora posted on micro-blogging platforms are related to participation in a few recent shared tasks, namely AMI 2018 (Fersini et al., 2018), HatEval 2019 (Basile et al., 2019) and others. In addition, there is a lack of studies considering different variations of the Spanish language and how state-of-the-art language models such as BERT perform in hate speech detection when used for cross variants over the same language (Zhang et al., 2021; Hershcovich et al., 2022).

In (Plaza-del Arco et al., 2021), multilingual and monolingual pre-trained language models were compared to Deep Learning architectures (CNN, LSTM, and Bi-LSTM) and traditional ML models (SVM and Logistic Regression) for detecting hate speech on tweets written in Spanish. The authors used two datasets to conduct the comparison. The first corpus, HaterNet (Pereira-Kohatsu et al., 2019), has no information about the hate speech domain or the location where the tweets were posted. The second dataset is the HatEval corpus which contains only information about the target for hate speech against women and immigrants. They used BETO (Canete et al., 2020), a Spanish language implementation of BERT trained on Wikipedia articles, movies and TED Talks subtitles, scientific documents, and others written in Spanish. Results obtained in (Plaza-del Arco et al., 2021) showed that BETO, a monolingual LM outperforms multilingual pre-trained models such as XLM and mBERT as well as the rest of the models they evaluated for hate speech detection in Spanish. Results in line with Plaza-del Arco et al. (2021) have also been achieved in other similar studies on hate speech detection (Benítez-Andrades et al., 2022; Tanase et al., 2020).

Nozza (2021) studied hate speech detection against women and immigrants across three languages: Spanish, English, and Italian. She investigated the limitations of zero-shot cross-lingual approaches using mBERT. Her results suggest that hate speech targets –i.e. different languages– should be studied separately as transfer learning in zero-shot scenarios is ineffective for misogyny detection. In addition to her findings, we aim to investigate whether such nuances can be extended to cross-variants within the same language.

[6]Many of these works can be found via the IberLEF annual shared tasks.

3 Datasets

In this section, we describe the datasets we use for training the misogyny detection models and the procedure we follow to compile and annotate the HaSCoSVa-2022 corpus, which is later used to train and evaluate our models to detect hate speech against immigrants. In Table 1, you can find a summary of the datasets we used in this work.

3.1 Misogyny existing datasets

3.1.1 MisoCorpus-2020

The MisoCorpus-2020 dataset (García-Díaz et al., 2021) compiles tweets written in Spanish, which are grouped into three categories: **VARW** (Violence Against Relevant Women), which refers to violent tweets directed to women with a significant social relevance; **SELA** (Spanish from Europe vs. Spanish from Latin America), which consists of tweets charged of misogynistic content written in Spanish from Europe – i.e., Spain – and posts with the same type of content written in a Latin America's variation of Spanish; and **DDSS** (Discredit, Dominance, Sexual harassment, and Stereotype), which comprises Twitter posts subdivided into different types of misogynistic attacks, such as derailing, rape, gender violence, and others. The dataset contains 10,244 tweet IDs in total. However, as the tweets were posted some years ago, we could find only 7,575 tweets in total (74% from the original dataset), where 49.2% is labeled as misogynistic.

3.1.2 Detección Misoginia (DetMis)

The Detección Misoginia (DetMis) dataset (Vera Lagos et al., 2021) contains 35K tweets geo-located in Mexico. The corpus is based on keywords related to sexism, stereotyping, and discrimination towards women from (Fisher et al., 2013). The authors used such keywords to search and filter tweets geo-located in each of the 32 states of Mexico. Since tweets were filtered based on keywords, a maximum of 5 tweets per keyword and label (misogynous and non-misogynous) were selected for annotation. Finally, 1K tweets were obtained per label after annotation. It is important to note that only one annotator participated in the annotation process.

Domain	Dataset	Nb tweets	% Hate speech	Variation
women	MisCorpus-2020	7575	49.2%	Europe, LatAm
women	DetMis	2000	50%	LatAm
women	IberEval 2018	3307	49.9%	Europe, LatAm*
immigrants	HaSCoSVa-2022	4000	13.9%	Europe, LatAm

* This dataset does not distinguish between both variations of Spanish — i.e. we cannot identify which tweets correspond to Europe or LatAm variations.

Table 1: Description of Spanish language corpora used for training the binary classification models.

3.1.3 IberEval 2018

The dataset is from the Automatic Misogyny Identification shared task at IberEval 2018 (Fersini et al., 2018). The corpus contains misogynous tweets in English and Spanish, and we only use the Spanish data. There are two main steps in the annotation process: First, part of the dataset was labeled by two annotators to define a gold standard. Next, the rest of the tweets were labeled through a majority voting approach on the CrowdFlower[7] platform based on the standard defined in the first step.

3.2 New Dataset: HaSCoSVa-2022

We reviewed the publicly available data for hate speech against immigrants in Spanish. However, to the best of our knowledge, there are no tweets corpora containing information about different language variations. Therefore, we create the HaSCoSVa-2022 corpus (**Ha**te **S**peech **Co**rpus with **S**panish **Va**riations) to conduct our experiments in the immigration domain. We focus on two immigration cases: immigration from Latin America and certain African countries to Spain and immigration from Venezuela to its surrounding countries where Spanish is their official language. Both cases carry a strong discriminatory online discourse due to religion, stereotypes, and other factors that concern a fraction of the local population.

3.2.1 Data Extraction

We define two geographical coordinates and radius to obtain geo-located tweets from Spain and Latin American regions. Tweets from Spain were extracted from a 520 Km radius surrounding latitude: 40.416705, longitude: -3.703583. The area from where we extracted geo-tagged tweets about immigration coming from Venezuela is centered on latitude: -3.976015, longitude: -79.225102, considering a radius of 1,200 Km. Note that the defined region for obtaining the European tweets is the same as the one defined by García-Díaz et al. (2021). However, since the Latin American region the authors proposed includes Venezuelan territory, we slightly changed it to exclude tweets produced in Venezuela as we need tweets from neighboring countries[8]. The regions we determine to extract the posts can be visualized in Figure 2 in Appendix A. We define three lists of keywords related to immigration and hate speech towards immigrants. Two sets of keywords contain 72 and 18 terms regarding European and Latin American immigration, respectively. In addition, the third set of keywords comprises 26 generic terms related to immigration – i.e., such terms are not region-specific. The terms are mainly demonyms, country names, and nicknames (offensive or not) related to such regions. The tweets were collected in two-time frames: from June 6th to June 28th and July 21st to August 4th. As a result, 75,834 tweets were obtained in total.

3.2.2 Data Annotation

To perform the data annotation, we randomly sampled 2,500 and 1,500 tweets produced in Europe and Latin America. We describe in detail the sampling strategy we follow in Appendix A.4. Two annotators, native Spanish speakers from Latin America, carry out the manual annotation. Both annotators tag each tweet into one of the three labels: xenophobic, non-xenophobic, or ambiguous. Whether a tweet is difficult to manually classify by an annotator, then the label provided by the annotator is "*ambiguous*". Otherwise, a tweet is classified as "*xenophobic*" if it matches **all** following conditions:

1. The content of the tweet primarily targets immigrants as a group, or even a single individual, if they are considered to be a member of that group (and NOT because of their individual characteristics).
2. The content of the tweet propagates, incites, promotes, or justifies hatred or violence to-

[7] https://figure-eight.com/

[8] Note that our aim is to analyze xenophobia against Venezuelan immigrants in regions surrounding Venezuela.

wards the target or a message that aims to dehumanize, hurt or intimidate the target.

We used the guidelines proposed by Basile et al. (2019) with minor modifications. A third annotator participated in the annotation campaign to provide a final label for tweets labeled as "*ambiguous*" by both previous annotators and posts previously tagged with different labels (i.e. one annotator tagged as "*xenophobic*" and the other as "*non-xenophobic*"). This annotator did not have access to the other annotations. Finally, 554 tweets were tagged as xenophobic, while 3,446 were labeled non-hateful towards immigrants. Thus, 13.9% tweets belong to the label of interest. The resulting corpus contains the tweet ID, the full text of the post, its label, and the language variation (LatAm or Europe). The inter-rater agreement reliability between both initial annotators according to Cohen's Kappa (Cohen, 1960) is 0.443 (88% agreement), which can be interpreted as a moderate agreement according to its author. The resulting HaSCoSVa-2022 dataset, keywords used for tweets extraction, and annotation guidelines are freely available to the research community[9].

4 Experimental Settings

Language Models. For the multilingual language model, we use mBERT (Devlin et al., 2019), the multilingual version of BERT, trained on Wikipedia data from 104 languages. We also experiment with BETO (Canete et al., 2020), a monolingual Spanish Bert, trained on the whole Spanish Wikipedia dump combined with the Spanish language texts of the OPUS Project (Tiedemann, 2012) without any differentiation between the Spanish variants. Others models for Spanish exist and are posterior to BETO (Gutiérrez-Fandiño et al., 2021; la Rosa et al., 2022), we decided to focus on BETO because of its pretraining data that makes it more comparable to mBERT. It would be of course interesting to conduct a large-scale Spanish monolingual models study on that topic but we leave it for future work.

Data Preprocessing. We replace all URLs and mentions with the same tokens, *url* and *@user*, respectively. In addition, since hashtags' segmentation has been shown to improve the results for certain tasks (Rosa et al., 2011; Declerck and Lendvai, 2016; Gromann and Declerck, 2017), we segment all hashtags into words to enrich tweets' messages with actual words. To develop such hashtags segmentation, we use Python's package *wordsegment*[10]. We randomly split the dataset into 70% for training and 30% for testing to ensure that each set's class distribution remains balanced. Also, we randomly pick 20% of the previously selected training set as the development set.

Evaluation. All fine-tuned models are trained over 5 different seeds, and all reported performance metrics are averaged over such runs to ensure evaluation robustness. Moreover, we select the best model out of 5 epochs after each training process according to the macro-F1 score on the development sets.

4.1 Multilingual vs. Language-specific

We use all the data described in Section 3 to compare the performance of the two models, mBERT and BETO. We aim to evaluate the differences between mBERT and BETO to detect hate speech in Twitter posts written in Spanish.

4.2 Spanish Language Variations

We use BETO to evaluate the performance of a monolingual model across Spanish variants. For this set of experiments, the Spanish variant of the tweet is relevant. Then, we exclude tweets that do not contain information about the region of origin. As a result, we keep 6,082 tweets for the misogyny experiments, where 3,596 posts correspond to the LatAm variant and 2,486 to the European. More details on the misogyny dataset used for this set of experiments can be found in Table 6 in Appendix A. All tweets on the immigration corpus are kept for this set of experiments.

The Latin American and European variation datasets sizes are not comparable according to the hate speech target. Therefore, we randomly undersample the largest variation dataset depending on the hate speech domain to set both variations to the same size and ensure the comparability of the transfer setting. As a result, the misogyny corpus for this set of experiments ends up with two sets of 2,486 tweets each –i.e., one set per variation. Therefore, each variation contains 1,392 tweets for training, 348 for development, and 746 for testing the models. Similarly, each variation subset in the immigration dataset includes 840, 210, and 450 records for training, development, and testing. An

[9] https://gitlab.inria.fr/counter/HaSCoSVa

[10] https://pypi.org/project/wordsegment/

overview of the train-dev-test splits can be found in Table 7 in Appendix A.

5 Results

Results obtained from the comparison between mBERT and BETO over the whole corpora are shown in Table 2. Results suggest that BETO outperforms mBERT in both hate speech domains. Specifically, BETO macro-F1 score is 11 points higher than mBERT on misogyny detection, whereas 4 points higher on xenophobia-related tweets classification. High standard deviations in both mBERT models compared to BETO suggests that BETO shows more stable and consistent performance across different runs. In line with previous works (Martin et al., 2020; Plaza-del Arco et al., 2021; Benítez-Andrades et al., 2022; Tanase et al., 2020), we find that using a language-specific LM where much more Spanish data is used for training and no other languages are considered, results in a better performance for detecting hateful posts written in Spanish.

Model	women	immigration
mBERT	74.4 ($\pm$ 7.0)	69.6 ($\pm$ 2.8)
BETO	**84.9** ($\pm$ 0.3)	**73.1** ($\pm$ 0.8)

Table 2: Models' average macro-F1 scores obtained on the test split over five runs. We select the best model out of 5 epochs for each run according to the macro-F1 score on the development set. The standard deviation computed over the 5 runs is inside parenthesis.

The second set of experiments aims to compare mono-lingual and cross-lingual settings across Spanish variants. Table 3 shows that the BETO model performance is significantly higher when trained and tested on the same language variant in both hate speech domains. For instance, the score of the misogyny model trained on European Spanish is 18 points higher when tested on European Spanish than on Latin American Spanish. On the other hand, the difference is 8 points for the xenophobia model, when the model is trained on Latin American Spanish and tested on tweets from Europe. We can also note that in all cases, macro-F1 scores present a higher standard deviation when the source data comes from Latin America.

6 Error Analysis

In this section, we analyze and compare errors in cross variants evaluation. We briefly examine the reasons that might lead to poor performance when the model is trained and tested on different language variants. Part of our analyses is inspired by the error analysis carried out in (Plaza-del Arco et al., 2021). First, we analyze the errors obtained by BETO. Such analysis is detailed in Table 4. Regarding the misogyny models, we can observe that models tend to wrongly classify non-harmful tweets from LatAm as misogynous, as 59.5% errors in common by both models are false positives. Moreover, in the xenophobia-related errors, we can see that 81.5% of the errors obtained in common by both models on European tweets correspond to false negatives. Similarly, a higher rate of false negatives is obtained by both models on the LatAm target since 65% of errors obtained in common are actual xenophobic tweets tagged as non-hateful by both models. We can attribute these results to the class imbalance in the immigration dataset (13.9% of the tweets are xenophobic), which might result in a difficult task for models to detect the minority class.

Moreover, in Table 5, we summarize the vocabulary coverage by the training sets on the test sets. In other words, we display the proportion of terms from the test sets included in each training set. We use a Spanish POS tagger to only consider nouns and adjectives for this analysis. For instance, in the case of the xenophobia dataset, we found 1,095 terms appearing in the LatAm test set and excluded in the Europe train set. As expected, for a given test set, a more significant proportion of terms found in the training set of the same variation than the other one. For instance, in the case of misogyny data, 50.3% of terms from Europe's test set can be found in Europe's training set, while only 39.6% is found in LatAm's training set. On average, test sets include 9.2% more terms in the training sets of the same variation than the others for both hate speech domains. Although we do not only consider hate-speech-related terms for this analysis, we found that various of the most frequently excluded terms correspond to derogatory words associated with a particular variant. For instance, the word "cerda" (which means *pig*) is found in the misogyny Europe set of tweets, but it does not appear in the Latin America tweets. Such a term is more used in Spain as an insult than in Latin America. The same happens with the term "vieja" (which might mean *old woman*), appearing in LatAm tweets but not in the European dataset. This term is mainly used in

		women		immigrants	
	Target	Europe	LatAm	Europe	LatAm
Source	Europe	**89.6 (± 0.6)**	70.5 (± 0.5)	**69.6 (± 0.9)**	64.9 (± 1.7)
	LatAm	71.4 (± 5.0)	**81.8 (± 0.5)**	62.8 (± 5.6)	**73.3 (± 2.7)**

Table 3: BETO's average macro-F1 scores obtained on the test splits over 5 runs. We select the best model out of 5 epochs for each run according to the macro-F1 score on the development set. The standard deviation computed over the 5 runs is inside parenthesis. Scores in **bold** indicate which source outperforms the other for a given target.

		women		immigrants	
Target	Source	False Pos.	False Neg.	False Pos.	False Neg.
Europe	Europe	38 (50.7%)	37 (49.3%)	31 (53.4%)	27 (46.6%)
	LatAm	108 (45.2%)	131 (54.8%)	13 (27.7%)	34 (72.3%)
	Common	15 (50.0%)	15 (50.0%)	5 (18.5%)	22 (81.5%)
LatAm	Europe	117 (55.7%)	93 (44.3%)	32 (43.2%)	42 (56.8%)
	LatAm	68 (55.7%)	54 (44.3%)	23 (43.4.%)	30 (56.6%)
	Common	44 (59.5%)	30 (40.5%)	12 (34.3%)	23 (65.7%)

Table 4: Number of tweets mislabeled per setting for each hate speech domain. In parenthesis, we show the percentage of mislabels on each type of error (False Pos. and False Neg.) from all the mislabels of a given domain and setting. Common mislabels correspond to errors obtained by both models (sources) on the same target.

Mexico for referring to women and can contain a derogatory connotation.

Finally, we use SHAP (Lundberg and Lee, 2017) to study the behavior of BETO in terms of explainability. SHapley Additive exPlanations, also known as SHAP, is a well-known model explainability technique used to interpret the models' decisions. SHAP is based on Game Theory and assigns importance scores to features for a given example classification. Such scores indicate how much a feature influences the model toward its final output. In NLP tasks, it can assign importance scores to terms. Thus, we use SHAP to examine how the models behave when the word *tonta* (which means *idiot*, female gendered, in English) appears in a text. Such an insult is an example of how the same term can be interpreted differently in two variations of Spanish. In Spain, that insult is much more aggressive than how it may be interpreted in Latin America. We take one misogynous tweet containing the word *tonta* from our corpus and classify such text by the misogyny Europe and LatAm models. A colored representation of the scores computed by SHAP on both classifications is shown in Figure 1. We can observe both models provide different classifications, where the model trained on European data performs correctly. SHAP finds the word "*tonta*" highly influences the model trained on European tweets to classify the tweet as misogynous, as shown in Figure 1b. In contrast, the same term provides almost no influence on the LatAm model's final decision according to SHAP in Figure 1b. We can note the analyzed term slightly contributes towards the wrong (non-misogynous) class when the LatAm model is used.

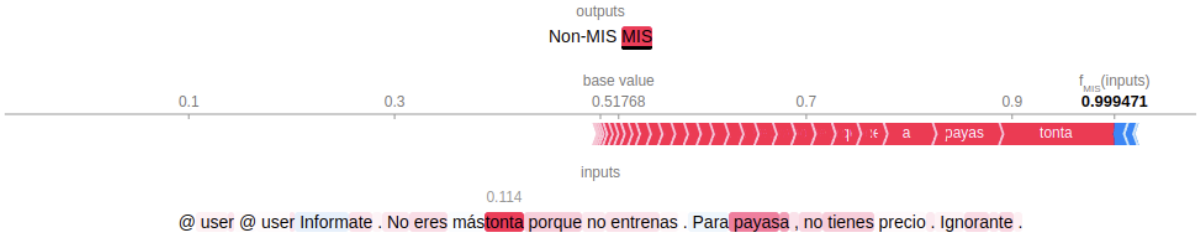
7 Conclusions

In this study, we showed how BETO, a Spanish version of BERT, as expected, performs significantly better than Multilingual BERT for classifying tweets as hateful for two hate speech domains: misogyny and xenophobia. Our outcomes align with previous studies mostly conducted with corpora proposed in popular shared tasks on hate speech detection. This does not mean that Multilingual BERT is not useful since findings in (Wu and Dredze, 2020) suggested that mBERT is remarkably useful on low-resource language tasks, in contrast to monolingual BERT implementations that use a significant amount of data.

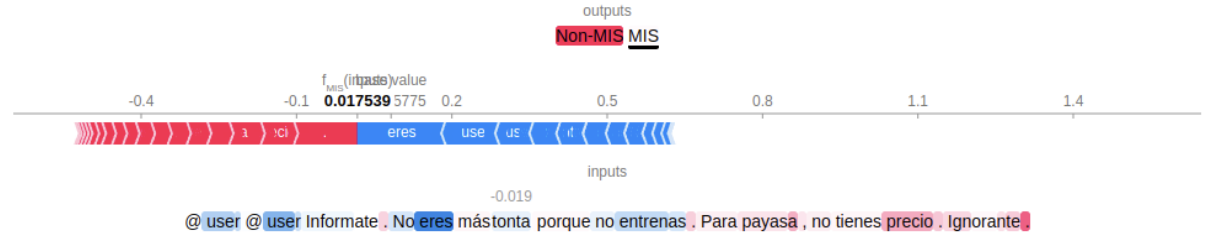
Moreover, we demonstrated that variants of a language, for instance, due to its use in different countries or cultures, affect the performance of hate speech detection models. In other words, we found that whether we train a model using data derived from only one variant of Spanish, the model's per-

	women				immigrants			
	Europe Test		LatAm Test		Europe Test		LatAm Test	
Train	Included	Excluded	Included	Excluded	Included	Excluded	Included	Excluded
Europe	50.3%	49.7%	38.5%	61.5%	45.3	54.7%	40.3%	59.7%
LatAm	39.6%	60.4%	47.7%	52.3%	36.4%	63.6%	48.1%	51.9%

Table 5: Proportion of terms on the testing sets included and excluded on each training set.



(a) SHAPE values obtained on the misogyny BETO model trained on Europe data. The output of the model for the positive label is 0.999471, classifying the tweet as misogynous. The term "*tonta*" (translated to English as *idiot*) is strongly colored in red, which means it strongly impacts the model to provide its final output towards the misogyny class.



(b) SHAP values obtained on the misogyny BETO trained on LatAm data. The model's output for the positive label is 0.017539, classifying the tweet as non-misogynous. The term "*tonta*" (translated to English as *idiot*) is almost not colored, which means it does not provide any relevant impact on the model to provide its final output.

Figure 1: SHAP values obtained from the misogyny BETO trained on European tweets 1a and LatAm data 1b classifying the same misogynous tweet from our corpus. The model trained on LatAm data detects no misogyny, whereas the European model is capable of identifying hateful content. The final output of the models towards the misogyny class is written in **bold**. Red colored terms influence the final decision towards the misogyny label, while blue colored terms provide influence the model classification towards the non-misogyny class. The tweet can be translated to English as "*@user @user Get informed, you can't be more of an idiot because you don't train, for a clown, you're priceless, ignorant.*"

formance may decay if it is used on data derived from another variant of the same language. An explanation for this may be the usage of terms, which in some regions where Spanish is spoken as a native language may denote hate, could be unrelated to hate speech in other regions where Spanish is also an official language. Thus, the terms used for denoting misogyny in countries where the same language is spoken might differ from one place to another. In our work, we used data produced in Spain, compared to data produced in Latin America, considering various countries such as Mexico (North America), Colombia, Ecuador (South America), and others. Our results extend the findings obtained by Nozza (2021) to transfer cross variants within the same language, demonstrating that different language variants from the same language for a given hate speech domain might also need to be studied separately to develop hate speech detection systems. Additionally, if different variants in the same language are not treated as separate cases but as one single scenario, we should consider using examples from as many variants as possible during the training phase to obtain models capable of dealing with data collected from different regions where hateful expressions may vary from each other. Finally, we followed a structured data extraction and annotation scheme to build a new hate speech towards immigrants corpus in Spanish, considering different language variants. Our dataset will help advance the state-of-the-art in hate speech detection for language variation and con-

tribute to a better understanding of the dynamics of hate speech towards immigrants in online environments. We release this corpus for use by the scientific community.

8 Limitations

In order to perform this work, we had to use simplified assumptions regarding the Spanish variants we worked on. We considered both variants as homogeneous geolects by themselves, whereas, of course, those geographical differences may constitute different dialects (cf. Wikipedia's world map of Spanish dialects, reproduced in Figure 3 in Appendix A.5).

The other limitation of our work is tied to the annotation biases eventually found in our dataset. Indeed, three annotators worked on the annotation of tweets forming the HaSCoSVa-2022 dataset, a new corpus we introduced for hate speech detection in two Spanish variants. Nevertheless, all annotators are from Latin America. Thus, some interpretations of tweets from the European Spanish variant might be questionable, given a potential lack of knowledge of certain hate-speech-related expressions used in Spain. To mitigate this issue, we included extensive observations regarding potentially confusing expressions from the European variant in the guidelines we provided. Additionally, the adjudicator (i.e. the annotator resolving the conflicts) in our annotation campaign has an academic background in political science and discrimination towards minorities and has lived in Spain for a significant amount of time. We thus believe that this problem has been properly handled. Nevertheless, as we will publicly release this dataset, including the guidelines and the seed words we used, within an open-source license, we will welcome any concurrent annotation and bug reports.

9 Ethical Considerations

This paper is part of a line of work aiming to investigate the effect of language variation on hate speech detection, fight the spread of offensive and hateful speech online, and have a positive global impact on the world. It has been approved by our institutional review board (IRB), and follows the national and European General Data Protection Regulation (GDPR). All our experiments were executed on clusters whose energy mix is made of nuclear (65–75%), 20% renewable, and the remaining with gas (or, more rarely, coal when imported from abroad).

Acknowledgments

We warmly thank the reviewers for their very valuable feedback. This work received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021607.

References

Resham Ahluwalia, Himani Soni, Edward Callow, Anderson Nascimento, and Martine De Cock. 2018. Detecting hate speech against women in english tweets. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:194.

Carlos Arcila-Calderón, Javier J Amores, Patricia Sánchez-Holgado, and David Blanco-Herrero. 2021. Using shallow and deep learning to automatically detect hate motivated by gender and sexual orientation on twitter in spanish. *Multimodal Technologies and Interaction*, 5(10):63.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

José Alberto Benítez-Andrades, Álvaro González-Jiménez, Álvaro López-Brea, Jose Aveleira-Mata, José-Manuel Alija-Pérez, and María Teresa García-Ordás. 2022. Detecting racism and xenophobia using deep learning models on twitter data: Cnn, lstm and bert. *PeerJ Computer Science*, 8:e906.

José Alberto Benitez-Andrades, Álvaro González-Jiménez, Álvaro López-Brea, Carmen Benavides, Jose Aveleira-Mata, José-Manuel Alija-Pérez, and María Teresa García-Ordás. 2022. Bert model-based approach for detecting racism and xenophobia on twitter data. In *Research Conference on Metadata and Semantics Research*, pages 148–158. Springer.

Paola Bentivoglio and Mercedes Sedano. 2011. Morphosyntactic variation in spanish-speaking latin america. *The handbook of Hispanic sociolinguistics*, pages 168–186.

José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In

Proceedings of the international AAAI conference on web and social media, 1, pages 512–515.

Thierry Declerck and Piroska Lendvai. 2016. Towards the harmonization and segmentation of german hashtags. *Bochumer Linguistische Arbeitsberichte*, page 10.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-Pineda, Manuel Montes, Juan Aguilera, and Luis Meneses-Lerín. 2020. Automatic detection of offensive language in social media: Defining linguistic criteria to build a mexican spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 132–136.

Elisabetta Fersini, Giulia Rizzi, Aurora Saibene, and Francesca Gasparini. 2022. Misogynous meme recognition: A preliminary study. In *International Conference of the Italian Association for Artificial Intelligence*, pages 279–293. Springer.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Terri D Fisher, Clive M Davis, and William L Yarber. 2013. *Handbook of sexuality-related measures*. Routledge.

José Antonio García-Díaz, Mar Cánovas-García, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021. Detecting misogyny in spanish tweets. an approach based on linguistics features and word embeddings. *Future Generation Computer Systems*, 114:506–518.

José Antonio García-Díaz, Salud María Jiménez-Zafra, Miguel Angel García-Cumbreras, and Rafael Valencia-García. 2022. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, pages 1–22.

Debbie Ging and Eugenia Siapera. 2018. Special issue on online misogyny.

Dagmar Gromann and Thierry Declerck. 2017. Hashtag processing for enhanced clustering of tweets. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 277–283.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2021. Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

T Nockleby John. 2000. Hate speech. *Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan, 2000)*, pages 1277–1279.

Habibe Karayiğit, Ali Akdagli, and Çiğdem İnan Aci. 2022. Homophobic and hate speech detection using multilingual-bert model on turkish social media. *Information Technology and Control*, 51(2):356–375.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

Estefanía Lozano, Jorge Cedeño, Galo Castillo, Fabricio Layedra, Henry Lasso, and Carmen Vaca. 2017. Requiem for online harassers: Identifying racism from political tweets. In *2017 Fourth International Conference on eDemocracy & eGovernment (ICEDEG)*, pages 154–160. IEEE.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

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.

ABDULLAH Muaad, Channabasava Chola, Bibal Benifa JV, J Hanumanthappa, et al. 2021. Detection of misogyny from arabic levantine twitter tweets using machine learning techniques. -.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Misogyny detection in twitter: a multilingual and cross-domain study. *Information Processing & Management*, 57(6):102360.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon. In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Flor-Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Urena-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Raúl R Romero-Vega, Oscar M Cumbicus-Pineda, Ruperto A López-Lapo, and Lisset A Neyra-Romero. 2020. Detecting xenophobic hate speech in spanish tweets against venezuelan immigrants in ecuador using natural language processing. In *International Conference on Applied Technologies*, pages 312–326. Springer.

Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2011. Topical clustering of tweets. *Proceedings of the ACM SIGIR: SWSM*, 63.

Elena Shushkevich and John Cardiff. 2018. Misogyny detection and classification in english tweets: The experience of the itt team. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:182.

Spanish Wikipedia. 2021. Spanish wikipedia — Wikipedia, the free encyclopedia. [Online; accessed 8-March-2022].

Statista. 2021. Countries with the largest number of native spanish speakers worldwide in 2021.

Mircea-Adrian Tanase, George-Eduard Zaharia, Dumitru-Clementin Cercel, and Mihai Dascalu. 2020. Detecting aggressiveness in mexican spanish social media content by fine-tuning transformer-based models. In *IberLEF@ SEPLN*, pages 236–245.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218. Citeseer.

Valeria Vera Lagos et al. 2021. Detección de misoginia en textos cortos mediante clasificadores supervisados. B.S. thesis.

Emily A Vogels. 2021. The state of online harassment. *Pew Research Center*, 13.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. 2021. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Datasets Details

A.1 Misogyny Dataset Description

	Europe			LatAm		
Dataset	Nb MIS	Nb non-MIS	% MIS	Nb MIS	Nb non-MIS	% MIS
MisCorpus-2020	1289	1197	51.9%	1218	378	76.3%
DetMis	-	-	-	1000	1000	50.0%
All	1289	1197	51.9%	2218	1378	61.7%

Table 6: Misogyny corpora descriptions after removing tweets without a variation tag (i.e. no information about the Spanish variation). Information about classes MIS (Misogyny) and non-MIS (non-misogyny) is disaggregated, as well as the percentage of misogyny instances per dataset and variation. The IberEval 2018 dataset is not included because it does not provide information about language variations.

A.2 Subset Splits

	women			immigrants		
Variant	train	dev	test	train	dev	test
Europe	1392	348	746	1400	350	750
LatAm	2014	503	1079	840	210	450
Comparable size	1392	348	746	840	210	450

Table 7: Number of tweets per dataset split on each hate speech domain with comparable data size. The comparable data size is obtained on each hate speech domain by randomly undersampling observations to ensure the comparability of the transfer settings among language variants.

A.3 HaSCoSVa-2022 Tweets Geolocation

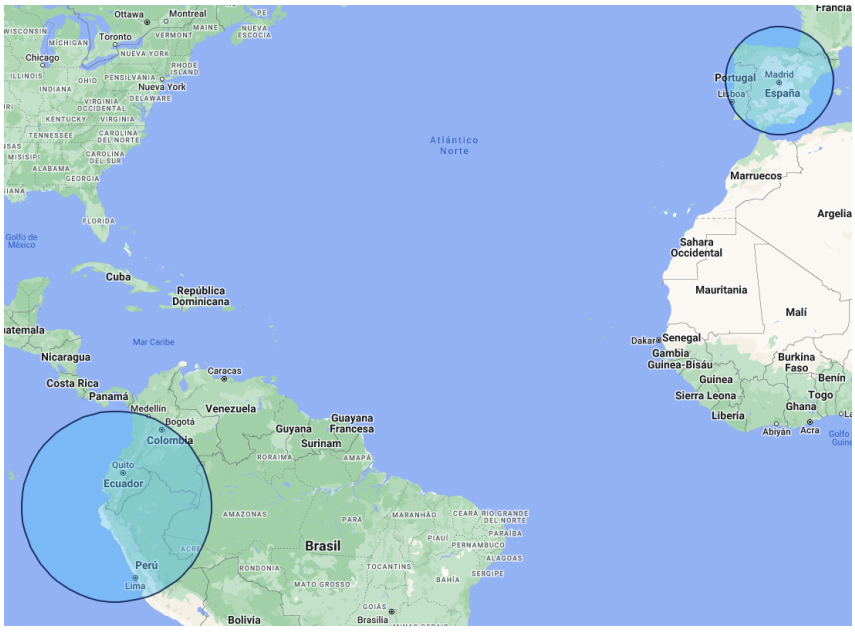


Figure 2: Boundings used to create the HaSCoSVa-2022 dataset by geo-locating European and Latin American tweets.

A.4 HaSCoSVa-2022 Sampling Strategy

In order to collect the data, we used keywords related to hate speech to extract subsets of tweets from Europe and Latin America (LatAm). For each keyword, we randomly sampled up to 50 tweets from Europe and 200 tweets from LatAm. We use a higher maximum number of tweets for LatAm due to the lower number of keywords related to hate speech we used for this region. This initial sampling strategy aims to avoid missing tweets containing non-frequent keywords. We also set a maximum number of tweets per keyword to avoid overrepresenting or underrepresenting some keywords in our final dataset.

After the initial sampling, we obtain 11,298 tweets in total. We then randomly sampled 2,500 tweets for Europe and 1,500 for LatAm from this subset. The decision to use different numbers of tweets for the two regions was based on a review of the datasets, which revealed a higher rate of hate speech in the European dataset. Therefore, we choose to annotate more European tweets to ensure an adequate number of hate speech-related tweets. This selection resulted in 231 negative examples for LatAm out of 1,500 tweets and 323 for Europe out of 2,500 tweets.

A.5 World Map of Spanish Dialects

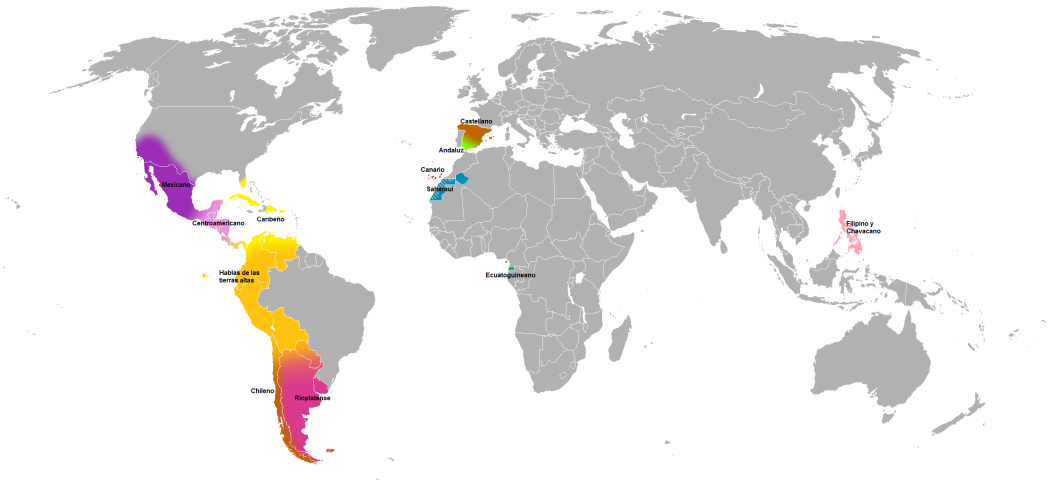


Figure 3: World map of Spanish Dialects (source Wikipedia).