

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

Laura Zeidler Chris Jenkins Filip Miletic Sabine Schulte im Walde

## Motivation

- *La nena juega muy poquísimo.* → Peru 🇵🇪
- *Qué tú quieres?* → Cuba 🇨🇺, Puerto Rico 🇵🇷, Dominican Republic 🇩🇴

## Data – Corpus del Español

- Texts from about 2M web pages from 21 Spanish-speaking countries
- We use data from 20 of them, leaving out the data from the US



## Research Questions

1. Do **tailored, dialect-specific features** improve a unigram model for automatic Spanish dialect classification?
2. Can a **simple ML model beat BERT** for Spanish dialects?

## Models

### Traditional ML Models

- Support Vector Machine
- Decision Tree

### Transformer Model

- Spanish BERT model

## Features

### Unigram-based Features

Simple BOW approach using term frequencies.

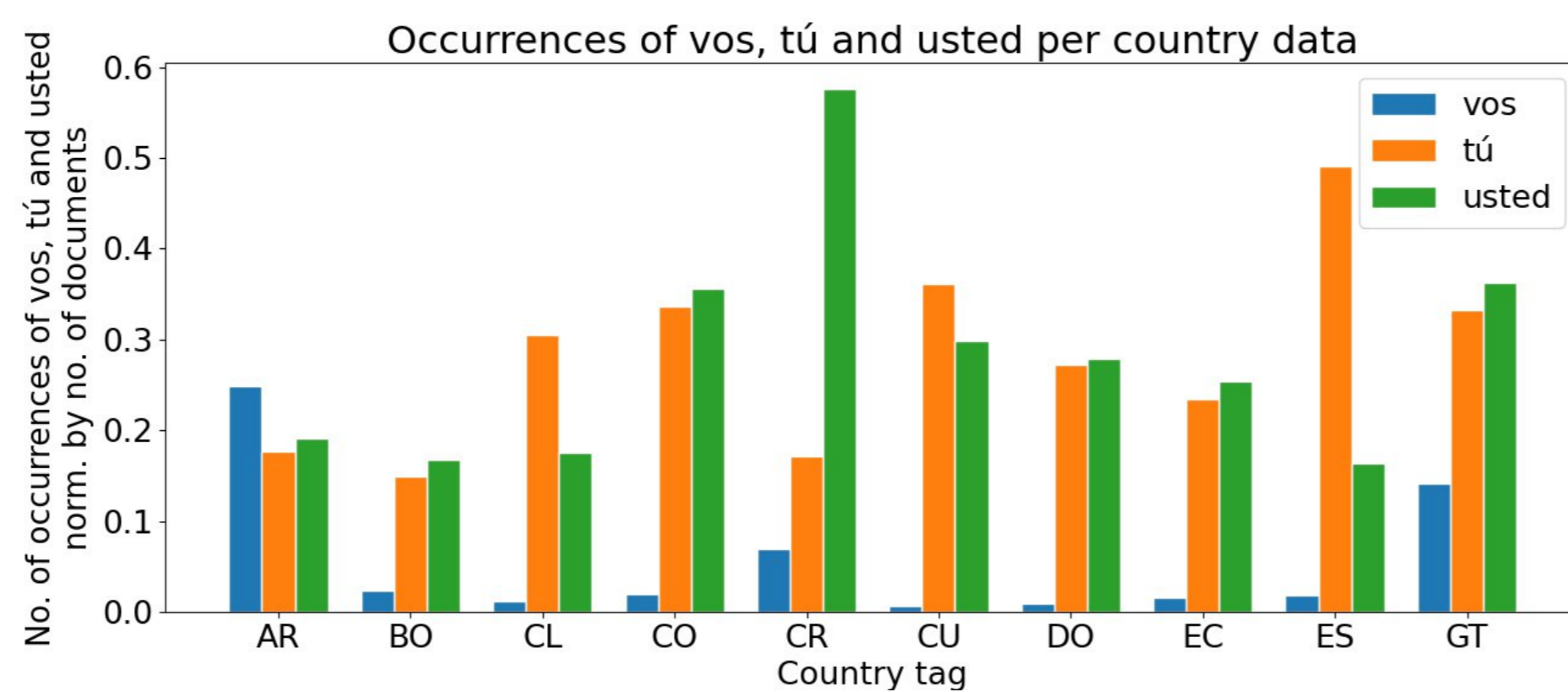
### Merged Features

Concatenate unigram-based and tailored features.

## Linguistically Tailored Features

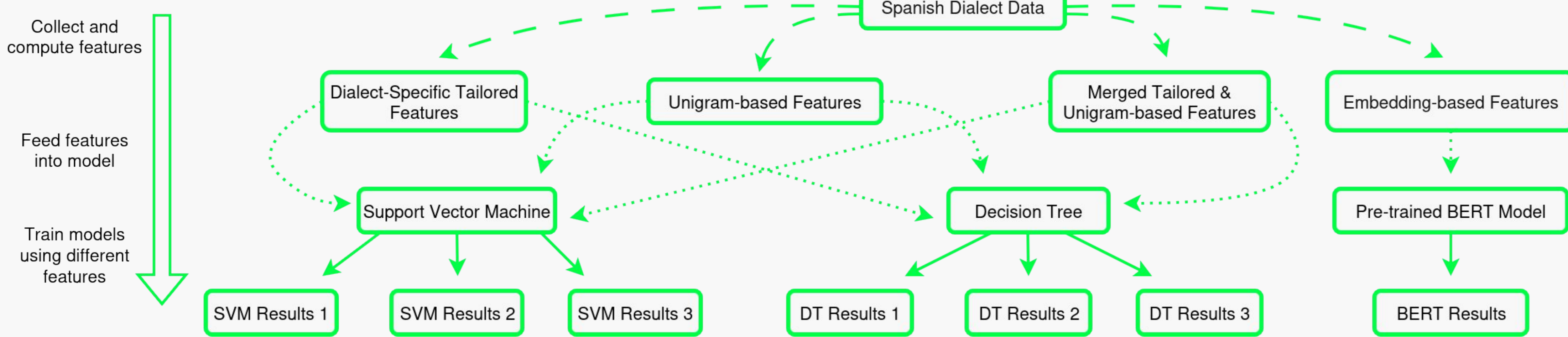
Features with **morphological and syntactic characteristics** that are **potentially indicative for specific dialects**.

Example: 2nd person singular pronoun use



Category	Feature name	Counted Items
Frequent	CLITIC	The clitics <i>lo, le</i> and <i>les</i>
	DIFFTENSE	14 different verbal tenses or aspects
	DIM	The diminutives <i>-ito/a, -ico/a, -illo/a</i> and <i>-ingo/a</i> resp.
	OVSUBJ	9 overtly realized subject pronouns
	SER_ESTAR	Verbs <i>ser</i> and <i>estar</i> in the context of adjective predicates
Rare	VOSEO	1) pronouns to address someone familiarly ( <i>vos, tú, usted</i> ) 2) verbs of the different conjugations associated with <i>vos</i> Pronoun <i>vosotros</i> and subsequent <i>os</i>
	VOSOTROS	
	ADA	Non-standard nouns formed with the suffix <i>-ada</i>
	ARTPOSS	Combination of indef. article, possessive adjective and noun
	MASNEG	<i>más</i> preceding negative adjectives
	MUYISIMO	<i>muy</i> preceding a token ending in <i>-ísimo</i>
	NONINV	Non-inverted WH questions
	SUBJINF	Subject pronoun preceding an infinitive or gerund

## Overview



## Experimental Set-Ups

- Standard set-up.
- **Replace named entities (NEs) and nationalities.**
- **Cluster countries into 10 aggregated classes.**

## Results

Large gap between tailored and other features

Tailored features reflect lower inter-class similarity

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	<b>0.66</b>	<b>0.66</b>
	Both	0.65	0.65	0.55	0.55	<b>0.66</b>	<b>0.66</b>
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	<b>0.67</b>	<b>0.67</b>	<b>0.59</b>	<b>0.59</b>	<b>0.66</b>	<b>0.66</b>

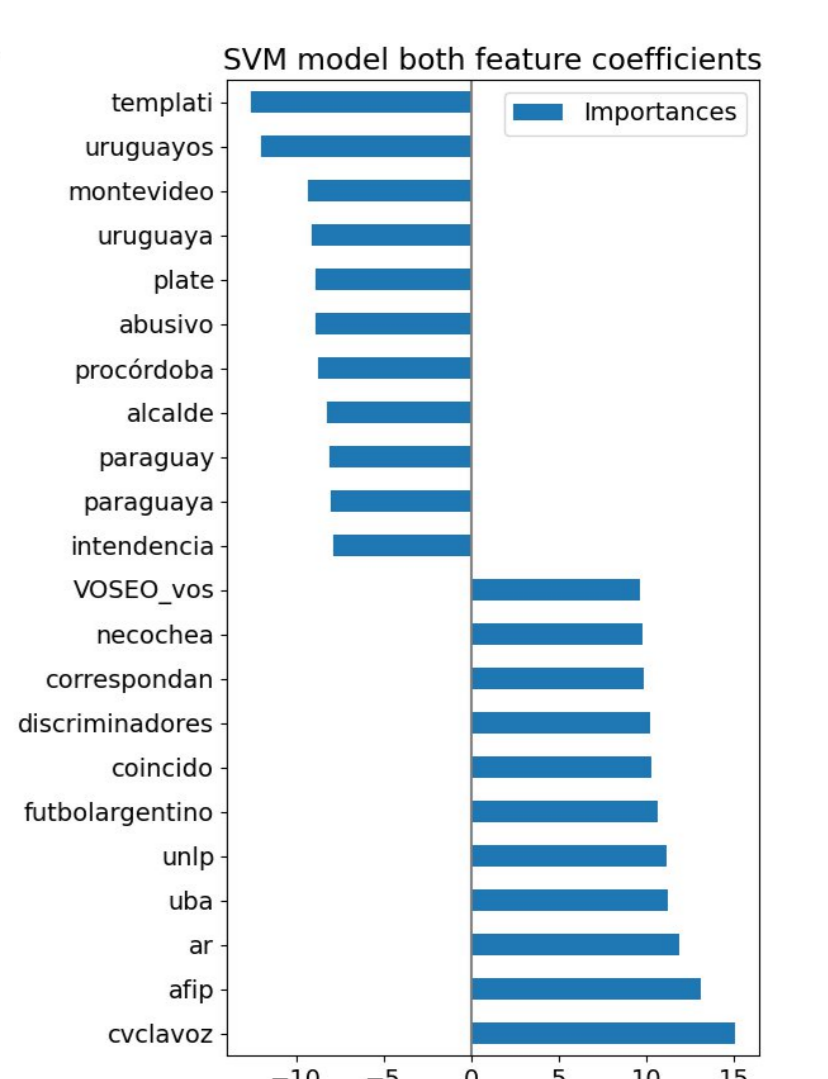
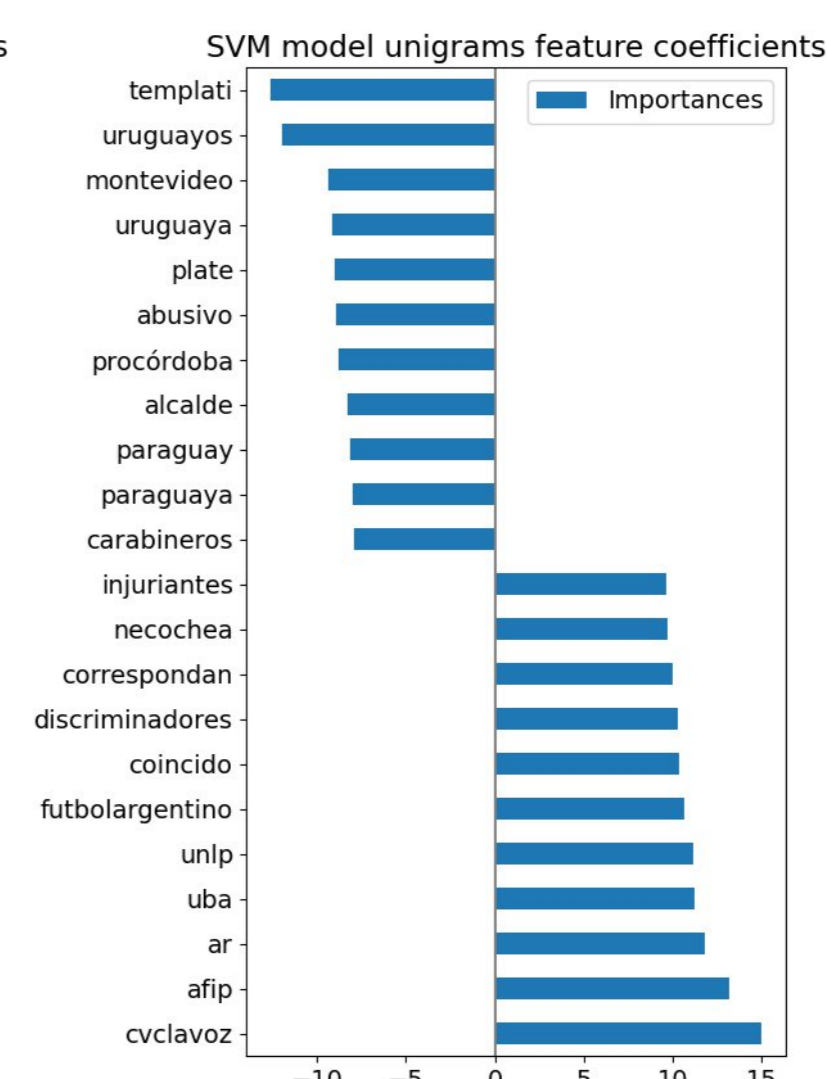
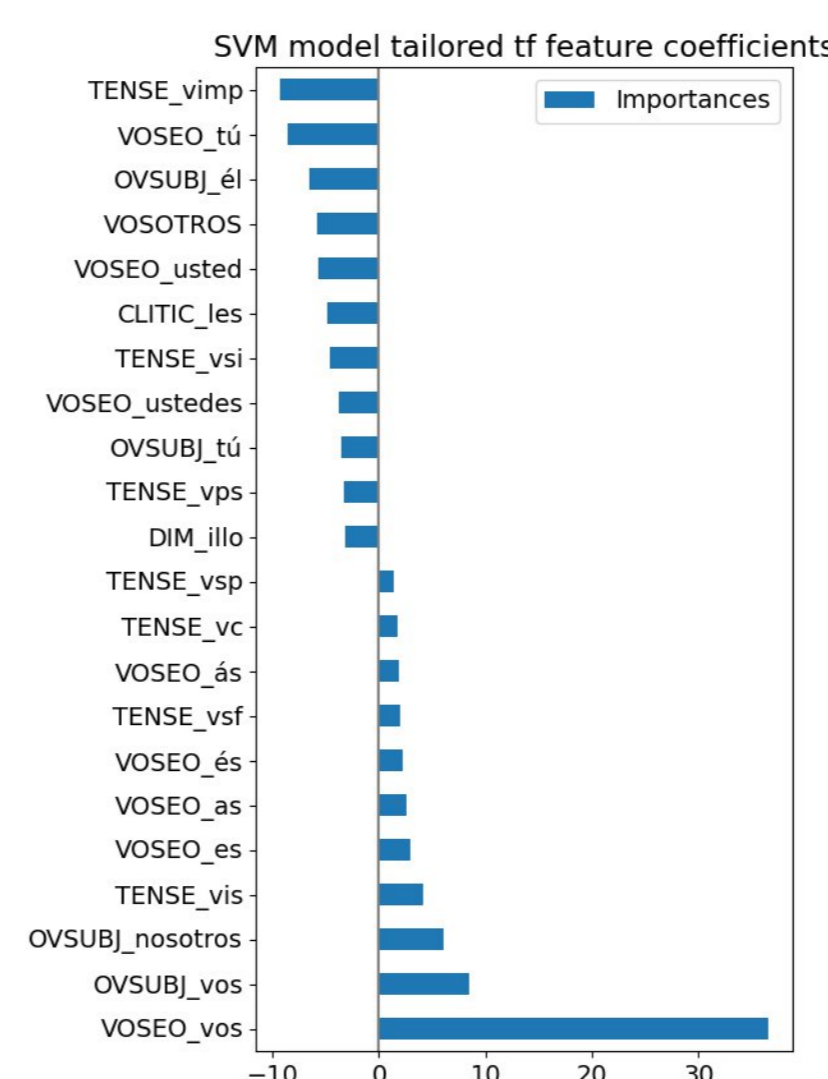
BERT outperforms other models, but only slightly

BERT and unigram-based models rely heavily on content-dependent cues

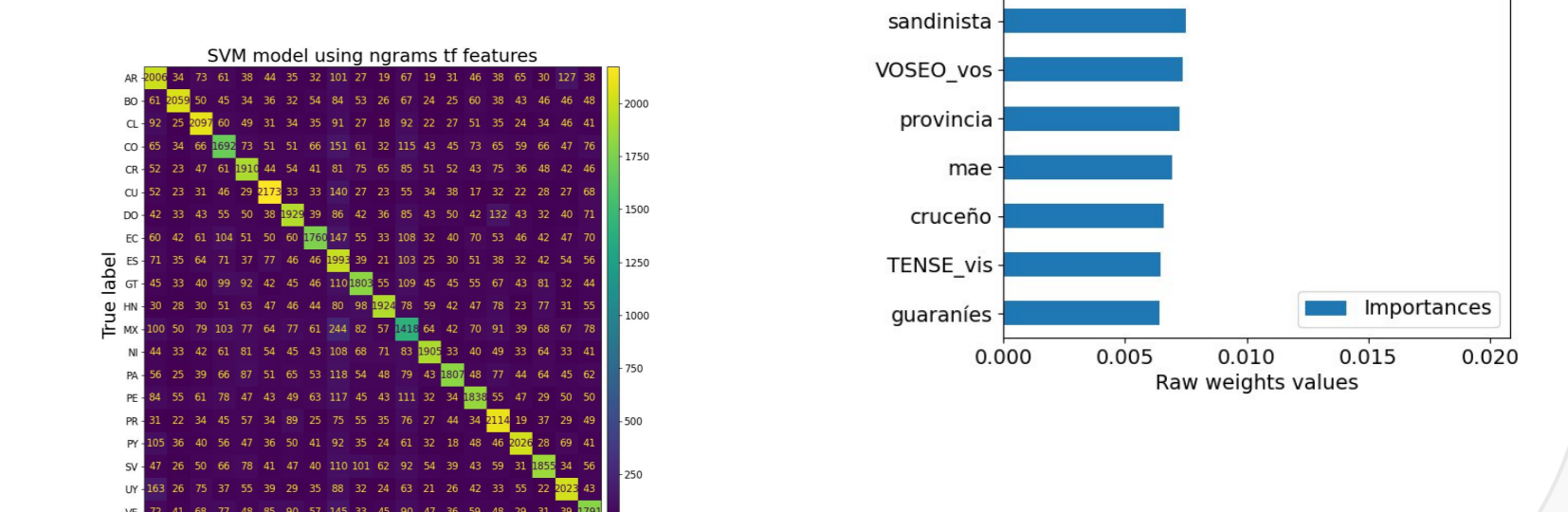
SVM is now on par with BERT

Features that occur frequently and across all dialects are most important

Obvious lexical cues dominate



When NEs are removed, more tailored features become important



## Conclusion

- BERT stayed on top — but only slightly, and **at the cost of efficiency, interpretability, and explainability.**
- Tailored features did not boost traditional models, but are promising for **dialect representation in a content-agnostic manner**

Focus on Peninsular Spanish which is very distinct from other dialects

