Party lines: Speech clustering in Spanish politics

An analysis of the differences in the speeches of members from the main parties in Spain.

Bennett Tharaeparambil, Andrew
Handt Fueyo, Miguel
Mirabent Rubinat, Guillem

Hannes Mueller

Barcelona School of Economics

March 18, 2024

1 Introduction

The political landscape of Spain used to be dominated by two main parties, the Spanish Socialist Workers' Party and the Popular Party (PP). However, since the mid-2010s, it has become more fragmented. This shift began notably around 2014-2015, when new parties emerged and started to challenge the traditional two-party system. The first ones to emerge were Podemos on the left and Ciudadanos (Citizens) on the center-right. Podemos, founded in 2014, capitalized on the discontent from the financial crisis and the corresponding austerity measures, positioning itself as a radical leftist alternative advocating for social justice and against corruption. Ciudadanos gained importance around the same time as Podemos and positioned itself as a liberal centrist party, advocating for national unity and economic liberalism. Further change came with the rise of Vox, a far-right party that entered the national parliament in the April 2019 general elections. Vox's emergence marked the first time a far-right party had won seats in the Spanish parliament since the transition to democracy. It capitalizes on issues such as national unity, opposition to Catalan independence, criticism of immigration, and a pushback against what it perceives as political correctness.

These developments have led to a more polarized and complex political environment in Spain, with coalition governments becoming more common and the traditional parties being forced to find compromises. The increased number of national players has also made regional parties more influential in national politics, as their support is often key to forming a government. In this context, we would like to explore the party lines and ideologies through the lense of NLP, using speech classification to gain insights into similarities and differences in the ways the main Spanish politicians talk about their beliefs. Through this approach, we would like to answer questions such as: "Can we train machine learning models to distinguish speeches from different parties accurately?""Are speeches from right wing parties clearly different from left-wing speeches?" "Do parties on the same side of the spectrum use similar words? (or combinations of words?)"

2 Literature Review

The state of text analysis research in Spanish parliament has focused on creating broad frameworks to give context to abstract text corpuses and on identifying argumentative strategies of the Spanish politicians. Researchers have found that by using contextual frameworks, the results of NLP and sentiment analysis are more rigorous and targeted, allowing analysis on specific topics like populism, emotion, and negativity (Torregrosa López et al. (2022)). The other identified front of text analysis in Spanish politics, is argumentative strategies. Researchers have identified general argumentative patterns that politicians use in their debate. The research details the patterns of attack of each politician, gathering insights on debate strategies. More specifically, researchers find that Spanish politicians employ a framework that defines and delimits who their enemy is, then draws that enemy as the enemy of Spain (Miljana Micovic (2020)). This identified pattern is particularly useful to our analysis, as we use it to preprocess our text with stopwords that occur in the argument pattern. These areas of text analysis in Spanish politics create frameworks that allow more specific and rigorous methodologies that are implemented with advanced machine learning models.

Political alignment classification is widely studied and most recent advancements use machine learning models to place corpuses of text within categories. The beginnings of political alignment classification began with the development of the DW Nominate political spectrum. This was the first quantitative measure of political bias from -1 to 1 based off of role call voting in the United States Congress (Poole and Rosenthal (1983)). Quantitative methodologies have been furthered since the 1980s with text comparison models which take the form of cosine similarity, sentiment analysis, and even more unique models like Wordfish (Slapin (2009)) to assign numerical values to political alignment(Sapiro-Gheiler (2018)). These comparative models, are typically used to create a network of users or documents with weighted edges based on document similarity, political alignment and sentiment scoring. More recently, researchers use labeled data and train models such as recursive autoencoders, neural networks, and BERT models to assign documents to a political spectrum (Essig and DellaPosta (2024)). One of these papers, ”Classifying party affiliation from political speech is most relevant to our project” Yu et al. (2008).

For instance, Yu et al. (2008) apply a range of nowadays considered simple models to classify speeches held in the United States Congress and House of Representatives in 2005. They argue that while this classification task has an additional complexity as opposed to sentiment analysis due to the language being less expressive and sometimes sarcastic, it is possible to distinguish the opinions of liberals and conservatives not on expressions of sentiment but rather based on their ideology, their underlying beliefs and how they talk about them (p.34-35) They then apply the following models: SVM (Support Vector Machines) and NB (Naive Bayes). The latter, Naive Bayes, calculates the likelihood of each word appearing in speeches from each party, then classifies the speeches based on the words that appear in it. It is called ”naive” since it assumes that each word contributes to the outcome independently of the other words. SVM works on the principle of finding the optimal boundary between different categories, or in this context, party affiliations, within a multidimensional space where speeches are represented as points. For SVM to be applied, the raw text has to be converted into a numerical format. Here the authors apply BOW (Bag of Words) and TFIDF (Term Frequency-Inverse Document Frequency) in order to compare the results later. In BOW, every speech is formatted as a vector in a multidimensional space, where each dimension corresponds to a unique word from the collection of speeches. On the other hand TFIDF can show how unique a word is to certain speeches as opposed to the entire set of speeches by giving higher weight to words that differentiate speeches and lower weights to words that appear very often across all of them. It is common practice to remove words that are very infrequent and words that are too frequent and the authors decide to remove those that appear less than 3 times and the 50 most frequent. Regarding the outcomes of these models, Yu et al. (2008) display accuracies ranging from 70 to 80 percent, with SVM outperforming NB. More interestingly, the results show that some of the models trained exclusively on the House of Representatives speeches generalise well to those from the Senate in the same year. However, when trained on speeches from both institutions and applied to classify ideology in earlier years, the performance in terms of accuracy drops over time. A possible explanation is that the ”time-dependency actually is a consequence of issue-dependency” (Yu et al., 2008, p.41) therefore the vocabulary changes from time period to time period based on what the current political issue is.

3 The Data

The data comes from an open source project called ParlaMint. This project aims to aggregate and annotate parliamentary debates in Europe. Currently, the project has speech documents from 29 countries and parliament groups, fully transcribed and annotated in multiple languages. Additionally, the project team at ParlaMint has worked to create extensive metadata for each of the speeches, including speaker names, party affiliations, gender, age and position. We selected the Spanish corpus for our analysis.

The Spanish corpus consists of 32,551 speeches from the Spanish parliament. The speeches span 8 years from 2015 to 2023, and includes 5 legislative terms. Over the course of the 5 legislative terms there were 445 sessions where speakers on average talked 27 times. In total there were over 50 parties in the data including subgroups of the two major parties we are interested in, PP and PSOE. Given the complex nature of the data and numerous party classifications, we had to preprocess the data effectively.

4 The Preprocess

Besides usual preprocessing steps such as turning ”PSOE” into 0 and ”PP” into 1, fixing some typos in speakers’ names, etc. we also carried out other processes which were more specific to our parliamentary speeches because of the nature of the data .

From the beginning, our data was not directly found in a nicely formatted .csv file but rather in multiple .tsv files which contained metadata for the speeches, which were found in .txt files. There were many folders with many .tsv-.txt pairs of documents so that was the very first step of preprocessing: turning the data into a dataframe we could effectively use.

Once we had the information in a dataframe, we were able to see that the only interventions we were interested in could be easily selected from the rest by selecting only ”diputado/a”(congressman/woman) speakers. The rest of the interventions at the Spanish Parliament come from either the president of that chamber (who comes from a party, and even always from PSOE or PP, but only talks to moderate the debate so their ”speeches” would be mostly noise words like ”thank you” or ”adelante” (go on)) or from guest speakers whose affiliation to a party, if it exists, is not necessarily known. By filtering those speeches further to only members of the parties we were interested in (PSOE, PP, Podemos, Vox, ERC, JxCat, PNV, Bildu) we brought the shape of the dataset down to approximately 12.5 thousand speeches for the duo PSOE-PP and around 2.5 to 4 thousand speeches for the rest of the duos.

Afterwards, we prepared some other variables: we turned parties into 0 and 1 labels (leftist ones to 0 and rightist ones to 1), we defined a dummy for whether the party of the speaker was in power at the moment of the speech, we cleaned and turned into integers the years of birth of the speakers to turn them into ages at the time of the speech, we solved some typos in the speaker ids (which were some names that were misspelled), we cleaned and turned gender into a dummy, we turned the variable term into a label encoded variable for the ”legislatura” in which the speech took place, and we also made sure that all the rows we were using had a speech associated to them by dropping the rows that had NA values in the column of the speeches.

This step left us with a pretty clean dataframe that was nearly ready to be put into the TFIDF vectorizer, but we decided to take a last twist to it. After a first round we saw that TFIDF had a hard time getting rid of some words that were just formalities, and we also saw that there were speeches that were nearly only formalities (for example from short interactions of a ”diputado/a” with the president of the parliament. We also noticed that the length of the speeches was very inconsistent and that could affect our results. To tackle these issues at once, we decided to use the following steps:

a) We started by tokenizing all of the speeches (we had already applied a little bit of lemmatization and stopwords removal previously when preparing the dataset).

b) Next we eliminated the speeches that were shorter than 20-100 tokens so that the short interactions that had little interest were already out of the equation (I say it as a range because we fine-tuned it quite similarly to how we tuned the hyper-parameters in the models).

c) Then we proceeded to slice the speeches that were longer than 200-500 tokens into pieces of at most 200-500 (we also treated this maximum as a sort of hyper-parameter). The result was that a text of 900 tokens, when the max tokens were set to 300 would end up being 3 rows with speeches of 300 tokens each. This solution, though, brought up another problem, the same row with a speech of 900 tokens would end up being cut into 4 rows if the max tokens was set to 250: 3 rows with speeches of 250 tokens of length and 1 row with a speech of 150 tokens. In this case this would not be too much of a problem, but if the leftover list was only 10 tokens long, it could introduce unwanted noise to the model so we came up with the next step.

d) We applied the first filter again after the split to take out all the leftover speeches that were not long enough to go over even this basic threshold.

e) We experimented with the minimum tokens and maximum tokens thresholds and found out that 100 and 300 were quite good for getting good predicting abilities of the model in both the PSOE-PP dataset and also at least in the Podemos-Vox one (as we will comment on later, the ability to discern between regional parties is more difficult to extract from speeches of the national parties).

f) As previously identified by Miljana Micovic (2020), we found many names and proper nouns that were being used in politicians’ debate strategies and had to remove those. We prepared a custom stopword list which contained names and surnames of important figures of both PSOE and PP and also some other words which could act as giveaways to the model such as the names of the parties themselves. We used this list so that the model would not be trained using these words and we found out that these giveaways were actually relevant for the model because the performance went down by more than 5% after we removed them.

5 The Model(s)

Our intent is to train the model in such a way that it can classify speeches from different parties with respect to the training but being the training on just the two main parties in Spain (PSOE-PP) and then using this same trained model on other speeches to see if performance remains similar or goes down drastically. To do that, we split our data into 4 initial datasets:

a) PSOE - PP (main center-left, center-right national parties): approx. 12.5k speeches.

b) UP - Vox (more clearly left and right national parties): approx. 4k speeches.

c) ERC - JxCat (main center-left and center-right Catalan parties): approx. 3k speeches.

d) Bildu - PNV (main left and right parties from the Basque Country): approx. 2.6k speeches.

Please notice that we always mention the couples of parties with the more leftist one first. This has a practical relevance in our model because we set those parties to the same label, 0. On the other hand, the more rightist parties have the same label as well, 1. As a result, we have PSOE, UP, ERC, Bildu with label 0 and PP, Vox, JxCat, PNV with label 1. With this we are trying to put them in comparable relative positions with respect to their couple to see if the classification between PSOE and PP is then useful as well to classify speeches on the rest of the datasets.

From there we further split the main dataset (PSOE-PP) in training and testing but also with a twist. To avoid overfitting the model to detect specific speakers instead of general trends in speeches as a party, we split the dataset by speakers, so that speakers in the test dataset were always new for the model. There are 70% of speakers in the train dataset and 30% of speakers in the test dataset, which turns in the end to approximately a test dataset with 26% of the speeches.

We also used the rest of the datasets as test datasets. When we use the model to classify speeches from UP and Vox, we are using a model trained on speeches from PSOE-PP exclusively, so it has never seen a speech from UP or Vox before. This was also the reason why we decided to include such a thorough list of giveaways. With giveaways we refer to words that make it very clear what party the politician is representing. These are names, since it happens often that a politician from one party addresses another from the opposed party in his/her speech ("Señor Pedro Sánchez,..."), and party names. When the giveaways were included in the training, the model used them to fit better to the training speeches and then the classification for the rest of the parties under-performed compared to its potential. In short, taking the giveaways out of the training made the model more generalizable.

We ended up deciding to go for a bi-modal approach to the matter. We trained both a Random Forest and an XGBoost model on the speeches from the training set to see how each performed. Our goal was to train the model on labels 0 and 1 so that we could use the probability prediction capabilities to produce a similarity score. If a speech is more similar to PSOE, it should get a probability closer to 0, while if a speech is closer to PP, it should get a probability closer to 1. In the end, the scores for both speeches were similar but the predictions of the Random Forest were more centered around 0.5 while the XGBoost gave distributed predictions. This is great because it means that both models are using the 0-1 line differently but they are both arriving at similar results so our model is a little bit more robust than if only one of the models was able to classify them with good performance.

6 Results

Having applied the pre-processing as described, we are now in a position where we can show the results that the three models get. Overall, we see that the best results both in-sample and out of sample are obtained using XGBoost, followed by Random Forest and finally unsurprisingly the Random classifier. The ROC-AUC curves below were obtained on out-of-sample PSOE and PP speeches, made by politicians that were not seen during training. If we compare these results with others in the literature, we are seeing better results than other simple approaches such as the one with TFIDF and SVM in Yu et al. (2008). They do not mention in their paper whether the speeches in the test set are made by politicians that do not appear in the training set. Recall that an AUC score of 0.9 means that there is a 90% chance that the model can distinguish whether a speech was made by PSOE or PP, so we are quite satisfied with these results.

Comparison of Classification Models (out-of-sample Speeches)

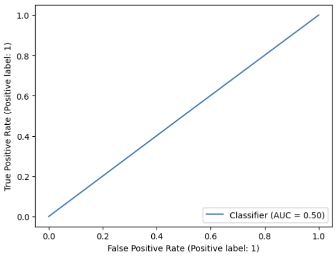

Figure 1: Random

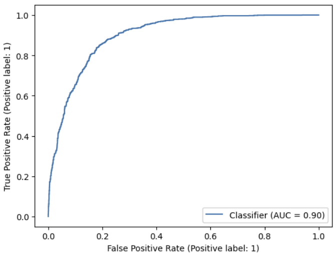

Figure 2: RFC

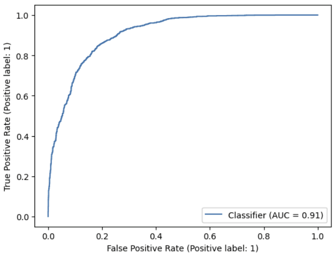

Figure 3: XGBoost

6.1 Feature Importances

We can analyse the feature importances to learn more about how XGBoost is classifying the speeches (See Table 1). We see that the most important n-grams are *derecha, ciudadanía, tranformación, socio* and *crear empleo*. It is hard to decipher whether they are indicative of PSOE or PP unfortunately but it is likely that *derecha* and *ultraderecha* would be used by the left wing party to criticize the right wing parties, and that the right wing parties would use *izquierda* in the same manner. Since most speeches in our dataset are held during periods of time where PSOE is the party in power, it would make sense that the n-grams with positive connotations are mostly used by them, for instance: *crear empleo, extraordinario, digno, mayoría absoluta, crecer*, while the negative ones would indicate the right wing's discontent: *recorte, propaganda, destruir, pagar*. There are also some terms that are very indicative about some recently important topics such as the independence of Catalonia (*independentista*).

Table 1: Top 50 Feature Importances

	Values		Values		Values
derecha	0.038142	izquierda	0.004678	término	0.002546
ciudadanía	0.024661	asunto	0.004543	comentar	0.002514
transformación	0.011111	elemento	0.004511	proyecto presupuesto	0.002503
socio	0.010764	2016	0.004418	millón español	0.002464
crear empleo	0.008942	recortar	0.004065	convalidación real	0.002435
cuestión	0.008694	ultraderecha	0.004033	mundo	0.002373
evidentemente	0.008617	extraordinario	0.003891	español española	0.002241
diputado bien	0.007627	digno	0.003856	pagar	0.002240
transición	0.007456	destruir	0.003501	español tener	0.002175
respecto	0.007144	indicar	0.003105	ciudadano ciudadana	0.002162
creación empleo	0.006475	colectivo	0.003102	historia	0.002135
pandemia	0.006063	inversión	0.003018	periodo	0.002118
empleo	0.005920	ciudadana	0.002989	formula	0.002117
recorte	0.005070	amnistía fiscal	0.002922	constitución	0.002065
mayoría absoluto	0.004873	2015	0.002893	tejido productivo	0.002044
independentista	0.004827	propaganda	0.002889	trabajador	0.002038
necesidad	0.004817	crecer	0.002835		

6.2 Main Output(s): PSOE vs. PP

It is already interesting to see that our model generalises well to politicians from the parties used during training, since this would indicate that there is a certain consistency in the semantics that each party is using, something we will look into more later on when looking at feature importances. In terms of visualization, we present a plot that displays speeches as horizontal lines above their predicted probability between 0 and 1 and are coloured by party (PSOE is red and PP is blue). We like this visualization since it resembles a political spectrum; however, it has the downside that it does not show well where along the line the speeches are clustered. That is why we have added a density plot on top, with the same colour scheme. The figure below shows this distribution of the out-of-sample predictions made by the XGBoost model. The Random Forest classifier, even though it almost performs similarly, clusters predictions around 0.5, so in terms of visualization is not as apt and we decided against including it.

We see that for the most part, PSOE speeches are receiving a probability close to zero, while PP speeches are closer to one. There are some cases where the classification is erroneous, and we can analyse those speeches for some insights.

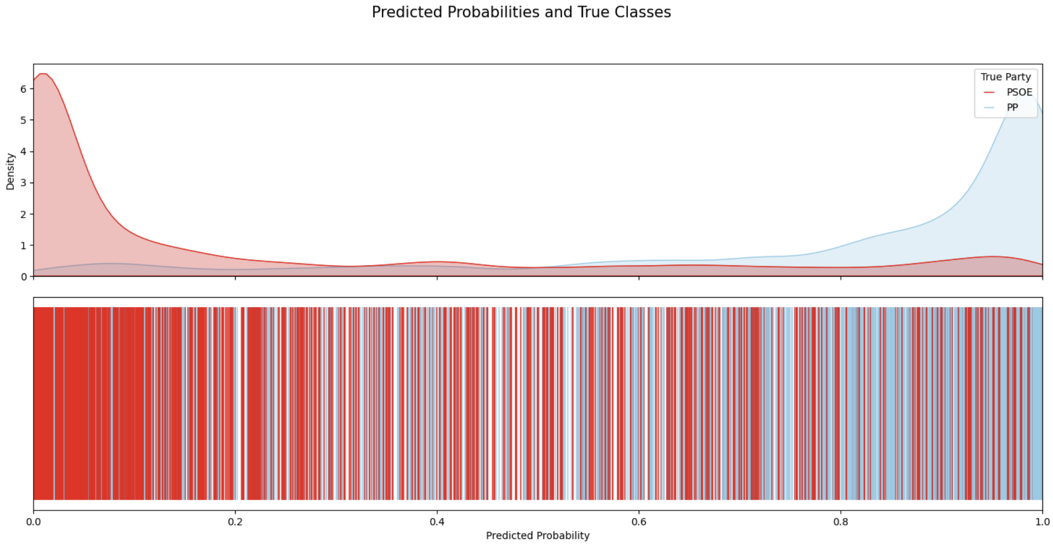

Figure 4: XGBoost out-of-sample classification

The following is an excerpt of a speech by Gabriel Elorriaga Pisarik (PP) on the 5th of October 2022 which was wrongly classified (prob. of 0.0008):

"*En fin, ya le hemos escuchado otras veces en esta Cámara decir que los medios de derechas tergiversan lo que dice. Tal vez sea su chispeante inteligencia la que va más rápido que sus palabras, pero lo cierto es que las citas que he hecho son literales y están grabadas.*

...

Ni pobre ni rico, sino todo lo contrario. Aquello que parecía una broma del absurdo se ha convertido hoy en realidad. El Gobierno ve en los españoles una especie de magma realmente incomprensible. Para ustedes, diez millones de contribuyentes con ingresos inferiores a 14 000 euros no son pobres y no merecen beneficiarse de las tímidas rebajas fiscales, ni en renta ni en el IVA, pero quienes ganan más de 21 000 son despreciables ricos, merecedores de la mayor carga fiscal en tiempos de crisis."

Taking into account the most important features in Table 1, we can speculate that the reason why this speech was misclassified lies in its usage of *derecha* and other words, but possibly also due to the fact that referring to the rich and the poor is typically more of a left wing rhetoric.

6.3 Other Outputs: Podemos-Vox, ERC-JxCat and Bildu-PNV

It is perhaps much more interesting to see that the model even generalises well to certain parties that were not represented in the training dataset at all. This is the case for the parties Podemos and Vox, which are typically situated at the extremes of the Spanish political spectrum. The XGBoost model is able to separate the two efficiently (See Figure 5) with an AUC of 0.86.

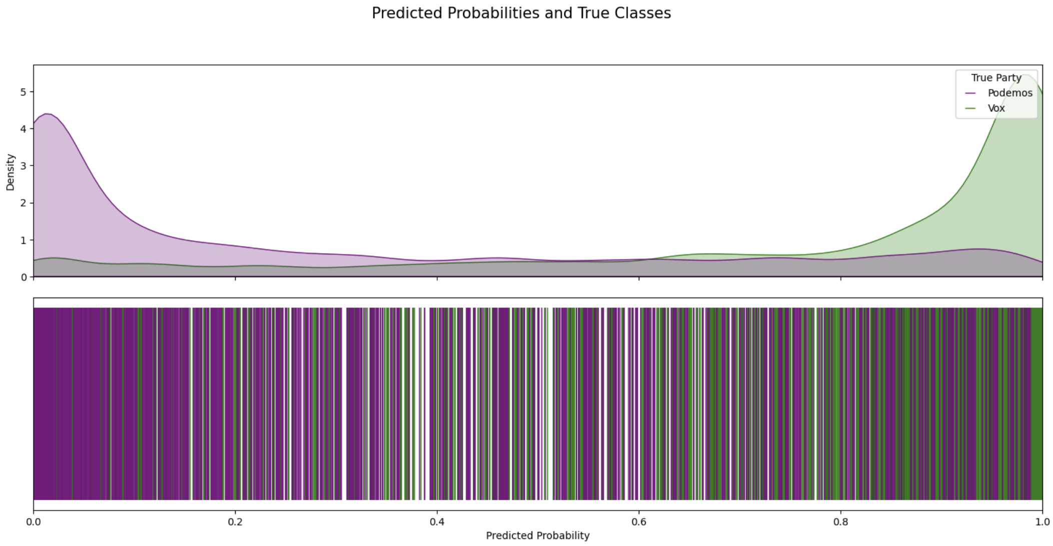

Figure 5: XGBoost out-of-sample classification

We interpret this as signifying that Podemos and PSOE share a sufficiently large part of their rhetoric and that Vox and PP do so too. This might not be surprising for someone who is familiar with Spanish politics but since the model does not have the understanding of ideology that humans do, this is purely down to semantics. To the model, some of these speeches from Podemos and VOX are so easily classifiable as belonging to PSOE or PP that they receive probabilities of nearly exactly 0 or 1. We can look at an example to try to understand why.

The following is an excerpt of a speech by Lucía Muñoz Dalda (Podemos) on the 16th of February 2023 which was correctly classified (prob. of 0.001):

"Gracias, presidenta. Ministra, señorías, hoy es uno de esos días en los que una es consciente del privilegio de ser testigo de un momento histórico desde esta Cámara. Como feminista, es para mí un honor defender una ley que blinda los derechos sexuales y reproductivos de las mujeres del país, una ley que avanza en la lucha histórica por la autonomía y el derecho a decidir sobre nuestros cuerpos, una ley feminista. Hoy damos un paso más en el largo camino que ha recorrido el movimiento de las mujeres, del que somos herederas; miles de mujeres que vinieron antes y que pusieron sus cuerpos para que estemos hoy aquí.

...

Todas ellas pavimentaron parte de este camino que hoy continuamos, y lo hacemos para que las que vengan después puedan seguir tejiendo este hilo morado del que todas somos parte. Hoy rendimos homenaje a todas ellas, a las mujeres anónimas, que se vieron obligadas a tener hijos cuando no los querían tener o a jugarse la vida para no tenerlos porque no podían pagarse un billete a Londres; a quienes establecieron redes clandestinas para acompañarlas *y a todas las que ocuparon plazas y calles, que hicieron suyo el espacio público para que nunca más una mujer de este país tuviese que morir en una camilla o cargar con la culpa y el estigma impuesto por el fundamentalismo religioso de la ultraderecha."*

We see some words out of the top 50: *histórico, ultraderecha* and some others with high importances: *mujer (pos. 366), defender (pos.278), etc.* And in general certain words that typically associated with leftist ideologies: *lucha, feminista, derechos sexuales, etc.*.

Where the model does not perform as well is with the regional parties (See Figure 6). We think that the model does not generalise well in this case because the regional parties hold speeches that specifically refer to the issues relevant to their regions. It seems natural that it would be hard to separate Esquerra Republicana and Junts Per Catalunya since they are united in their demands regarding the independence of Catalonia more than the issues and semantics typical of the right and the left.

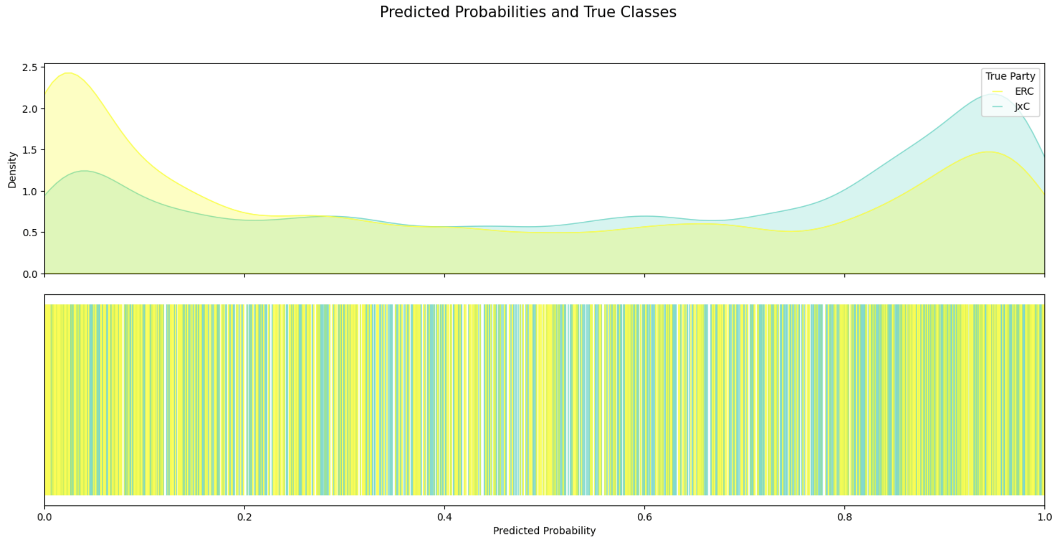

Figure 6: XGBoost out-of-sample classification

7 Conclusion

Overall the results have been able to answer the questions mentioned in the introduction to a degree that we consider satisfying. Indeed, a machine learning model can easily differentiate between right-wing and left-wing speeches, which implies that the speeches are sufficiently different. We can only imagine that a more advanced model, such as BERT would be able to do so even better. Additionally, we have found evidence of a certain semantic similarity in the speeches of parties that are somewhat aligned in their ideology, although in hindsight we have to acknowledge that it would have been better to remove words such as *derecha, izquierda*, since they could be considered "giveaways". Another caveat of our approach is that we have cut long speeches into smaller ones, potentially giving them a more significant weight in our dataset. Certainly, this is an imperfect approach but we could not afford to discard long speeches as this would have reduced the available data significantly. In that sense, with enough data it certainly would be interesting to show the evolution of the similarity between PSOE and PP in terms of their speeches over time, something we have not been able to investigate since our data is very imbalanced across time.

References

Essig, L. and DellaPosta, D. (2024). Partisan styles of self-presentation in u.s. twitter bios. *Scientific Reports*, 14.

González, J. and Novo, A. (2011). The role of the media agenda in a context of political polarization. *Comunicacion y Sociedad*, 24:131–147.

Miljana Micovic, Adrià Alsina-Leal, I. A.-R. (2020). Argumentative analysis of the electoral debates in the campaign 28-a: The construction of the enemy.

Poole, K. and Rosenthal, H. (1983). The polarization of american politics. *The Journal of Politics*, 46.

Sapiro-Gheiler, E. (2018). “read my lips”: Using automatic text analysis to classify politicians by party and ideology. [Accessed 09-03-2024].

Slapin, S.-O. P. . J. B. (2009). Wordfish manual.

Torregrosa López, F. J., D’Antonio Maceiras, S., Villar-Rodríguez, G., Hussain, A., Cambria, E., and Camacho, D. (2022). A mixed approach for aggressive political discourse analysis on twitter. *Cognitive Computation*, 15.

Yu, B., Kaufmann, S., and Diermeier, D. (2008). Classifying party affiliation from political speech. *Journal of Information Technology & Politics*, 5(1):33–48.