Forecasting Spanish Elections by predicting the Political Orientation of YouTube comments

Miguel Cardona Polo

Master of Science
School of Informatics
University of Edinburgh
2023

Abstract

As a result of the change in pricing of the Twitter API, we suggest YouTube as the alternative platform for the study of Spanish election predictions. This research introduces a new methodology to collect and clean YouTube comments related to the Madrid, Barcelona and Valencia municipality elections. Our analysis validates the use of YouTube comments for this area of research, as it showed a large share of politically charged content and the mention of candidates and parties, by many users. We improve on the winning model of the political ideology classification of tweets competition PoliticEs, achieving a 96% macro f1 score for the right, left and non-political speech classes. However, we also highlight the challenges of adapting this model directly for YouTube comment classification.

Research Ethics Approval

This project, of ethics application number: 927459, did not receive approval from the Informatics Research ethics committee due to a late application, which could not grant retrospective approval and could not be approved before the dissertation deadline. This project was planned in accordance to GDPR, respecting the privacy of the users from which the comments were collected. The files containing the analysis and raw data collected were password protected and the usernames were encrypted and simplified to the format of *user_[random id]*, protecting them from being traced back. The collected data was not shared with anyone outside of this study, which only includes the student and the appointed supervisors.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Miguel Cardona Polo*)

Acknowledgements

I would like extend my deepest gratitude to my supervisors, Walid Magdy and Youssef Al Hariri, for their support throughout this year. Our discussions and sessions have sharpened my critical reading skills and improved my organizational aptitude. This research has reignited my passion for Spanish politics, deepening my understanding of national affairs and the nature of the Spanish political environment. Inspired by the renewed enthusiasm this study has fostered, I am committed to further exploring this topic to gain a more profound understanding of the political paradigm of my home country, Spain.

Table of Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Problem Description 3

1.3 Project Objectives 4

2 Background **5**

2.1 Spain's political situation 5

2.2 Locality affairs 6

3 Literature Review **8**

3.1 Introduction to Methodologies 8

3.2 Methodology for Data Collection from Twitter 9

3.3 Using YouTube Data for Election Prediction 9

3.4 Integration of Deep Learning Models and Political Orientation Systems 9

4 Methodology and Experimental Setup **11**

4.1 Data Collection 11

4.1.1 Collection period 11

4.1.2 Collection Method 12

4.2 Data Cleaning 15

4.2.1 Date and Comment filtering 15

4.2.2 Video Title filtering 15

4.2.3 Video Author Filtering 16

4.3 Dataset Analysis 16

4.3.1 Video and Comment Timelines 16

4.3.2 Word Frequency 17

4.3.3 Length Distribution 20

4.3.4 User Comment Frequency 21

4.3.5 Author Attribution Analysis 22
4.3.6 Language Analysis 24
4.4 Building a Political Orientation Model 26
4.4.1 Collecting Training Data 26
4.4.2 Training Data Analysis 27
4.4.3 Choosing a Model 31

5 Results and Analysis **34**
5.1 Training and Fine Tuning 34
5.2 Testing on YouTube data sample 37

6 Conclusion and Future Work **38**
6.1 Conclusions 38
6.2 Future Work 39

Bibliography **41**

A List of Parties and their Candidates **46**

B Data Analysis on Barcelona **47**

C Data Analysis on Valencia **50**

D Language Analysis of YouTube data **53**

E Political Ideology Prediction model - Training Data analysis **54**

Chapter 1

Introduction

1.1 Motivation

Spain is a unitary state with a parliamentary democracy and a constitutional monarchy, according to the European Union [1]. Its governance is organized into three tiers of authority: the State, Autonomous Communities, and Local Entities, as recognized by the United Nations [2]. To illustrate this division, we can take the example of Barcelona, which serves as a local entity within the autonomous community of Catalonia, itself being part of the broader Spanish state.

This study focuses on electoral predictions concerning local entities, known as Municipality Elections. The latest Spanish elections of this kind took place on May 28, 2023. These have employed the D'Hondt method of proportional representation, a widely used mathematical formula within proportional representation systems, recognized by the European Parliament [3]. In this method, votes are translated proportionally into whole seats. Each political party presents a closed list of candidates, ranked in order of priority, matching the number of local councillor seats to be filled. The allocation of councillors is proportional to the total votes received. For instance, in a town with 30 councillors, a party obtaining 60% of the vote would secure 18 councillor seats. Within one month of the election, the councillors convene to elect their Mayor. The Mayor typically belongs to the majority party and subsequently appoints their councillors to form an executive board to oversee council departments. In cases where no single party secures an outright majority, minority parties may form a coalition group that does not necessarily include the party with the highest vote share [4].

The parties with seats in the council can make policy decisions that directly impact the city's economy and, by extension, the country's. This is particularly true for those

cities with large influence within Spain, such as Madrid or Barcelona. The ability to predict the outcome of elections can sway public opinion, expectations, as well as reduce uncertainty regarding the future direction of the country. The traditional method that has been used to predict the outcomes has been through surveys and polls, done in person or through the phone. Usually, these surveys involve asking people which party they support, as explained by the Spanish Center for Sociological Research [5]. This approach has been critiqued by sociology academics such as Jeff Manza [6] or Andrew J Perrin [7], questioning its validity. Some of the issues they describe include low response rates and restricted answer options. These drawbacks have motivated new methodologies based on social media, which due to their increasing user base, they have become a great source of information to extract public opinion, as explain by Bachner in its study on advances in public opinion and policy attitudes research [8].

Over the years, Facebook and Twitter have been the most popular social media platforms to extract public data for election prediction. The academic survey conducted by Khan et al. [9] in 2021, found 98 studies on election predictions using Twitter. Spain was the third country under study with most published papers (7), just behind USA (27) and India (24) - in the upcoming Literature Review section, these studies will be discussed in more detail. These papers have leveraged the quick and easy access of Twitter data, making most studies predominantly *Twitter based*. However, the recent change in pricing of Twitter's API has made data collection from this platform extremely expensive for academics [10]. Pushing institutions to consider alternative platforms for political-related data collection. In this research, we propose YouTube as the alternative platform.

YouTube comments have been widely neglected by the academic community as a source of public data for election prediction studies, however it has been found to be a great source of opinionated speech in regards with political stance, as per an early 2012 study [11]. This study by Mejova et al. also found that the YouTube's comment author's political stance and the sentiment of the comment do not always match, indicating that sentiment and political stance should be treated differently. These outcomes suggest that YouTube comments are a suitable source for election predictions. Another study has also proven similarities between YouTube and Twitter in the US elections of 2020. This study by Shevtsov et al. [12] showed that Twitter communities correlate with YouTube comment communities, possibly conveying a similar communication style and patterns.

In this study, we will test for the transferability of tweets with YouTube comments,

and leverage the existing Twitter datasets to create models for the political orientation prediction of YouTube comments. Through the political stance of each user we will make a share of expected votes, for the left and for the right wing parties.

1.2 Problem Description

Historically, the techniques used for predicting Spanish elections have remained relatively static, often relying on Volume, Sentiment Analysis, or their combination. While these methods have been successful to some degree, they've often been critiqued for their simplicity, particularly in the backdrop of the more advanced deep learning models available today, as discussed in the Literature Review section. Metaxas et al., in their paper titled *How (Not) to Predict Elections*, argue that predictions derived from such methods on Twitter data, are as accurate as random chance [13].

While the focus on electoral predictions persists, there's been a growing interest in using Twitter data to anticipate political ideologies. Within the context of Spanish politics, two parallel studies stand out. Prati and Hung's 2019 study delved into tweets from the 24M elections, which achieved 77% accuracy on the prediction of right and left-leaning tweets [14] with a Random Forest [15]. The other, more recent one, is the PoliticES 2022 competition, which set forth a challenge to extract political ideologies from texts [16]. The winning team, *Los Calis*, leveraged a deep learning model, boasting an impressive f1-score of 96% in categorizing users as left or right leaning [17].

The research discussed earlier suggests that YouTube has emerged as a substantial reservoir for politically-opinionated comments, potentially succeeding Twitter for such analyses. Given the evident success of deep learning models in related studies, it's clear there's a gap in the current methodologies applied to Spanish election predictions. This has paved the way for our research proposal targeting the Municipality elections of May 28, 2023.

Our methodology involves the adaptation of the deep learning model built by Los Calis. We aim to predict political orientations by analysing YouTube comments related to the Municipality elections of the 28 of May. We will focus on the three most populated cities of Spain, to ensure a large collection of comments. These cities, according to the World Population Review, are: Madrid, Barcelona and Valencia [18]. We will obtain the data through the YouTube API.

For training our classification layers, we will employ Twitter data from various politicians and journalists, available from different sources. Drawing from Shevtsov

et al.'s findings [12], we'll test if the data can be transferred between these platforms. Once we predict a comment's political orientation, we will categorize users based on a majority vote system. For example, a user with three right-leaning comments out of five will be classified as right-oriented. The influence of each user will be measured by the number of likes on their comments, a method validated by numerous studies which use similar metrics for tweets, such as favourites or retweets [19, 20, 21]. Finally, we'll determine the share of votes for both left and right orientations based on the resultant user influences.

1.3 Project Objectives

The success of this study can be measured by the quality in which the following objectives were embarked.

- **Devise proper methodology for YouTube data collection and cleaning** - No research has yet described a methodology for collecting YouTube comments in the realm of Spanish election predictions. Our objective is to outline a methodology analogous to those in studies that use Twitter as their source of public data.
- **Analyse data to check for politically opinionated comments** - We want to validate the study by Mejova et al. [11] that suggests that YouTube is a rich source of politically polarised comments. Therefore, testing our hypothesis that YouTube comments are adequate for election predictions.
- **Incorporate Deep Learning methods and compare our political orientation model to previous studies** - Construct a new dataset, which introduces a previously unseen class non-political speech. Using Los Calis model as a guide, build a model that uses BETO [22] and MarIA [23], and incorporate a classification layer to distinguish between the three ideologies (right, left, none). Evaluate the performance of our model by comparing it to previous research.
- **Test transfer learning from tweets to YouTube comments** - Drawing from the conclusions of Shevstsov et al. [12], test if models trained on tweets can be used for the prediction of YouTube comments. If successful, then feed the comments collected to build a share of ideologies and compare this prediction to the outcome of the elections.

Chapter 2

Background

2.1 Spain's political situation

To further contextualize this study it is important to have a basic understanding of the current and past political situation in Spain. The Spanish state transitioned to a democracy after the end of Francisco Franco's dictatorship in 1975. The first election held after the transition was in 1977, the winning party *Union de Centro Democratico* (UCD) came into power and lead the writing of the new and current Spanish Constitution (1978) [24]. The constitution stated that the senate and the deputies are elected for a four year period, after which elections must be held, as per its article 68 [25].

Since the next elections in 1982, there have been mainly governments of the right wing *Partido Popular* (People's Party) and the left wing *Partido Socialista Obrero Español* (Spanish Socialist Workers' Party). After the elections held on December 2015, new political parties and citizen candidacies emerged, which had already participated in the formation of local and regional governments. These include, the center stance *Ciudadanos* (Citizens party), the left wing party *Unidas Podemos* (Together we can), and the right wing party *Vox*. These elections also saw the emergence of parties fighting for the independence of different regions, like *Esquerra Republicana de Cataluña*, which fights for the independence of Catalonia. However, these are mostly supported in their respective regions.

Now, there are five main parties casting the majority of the votes across different regions. A public survey conducted by *SocioMétrica* for the newspaper *El Español* [26], collected the voters opinions on the ideological positioning, in the left to right realm, of each party. Figure 2.1 shows the result of this survey, where 0 represents an extremely left stance and 10 represents an extremely right stance.

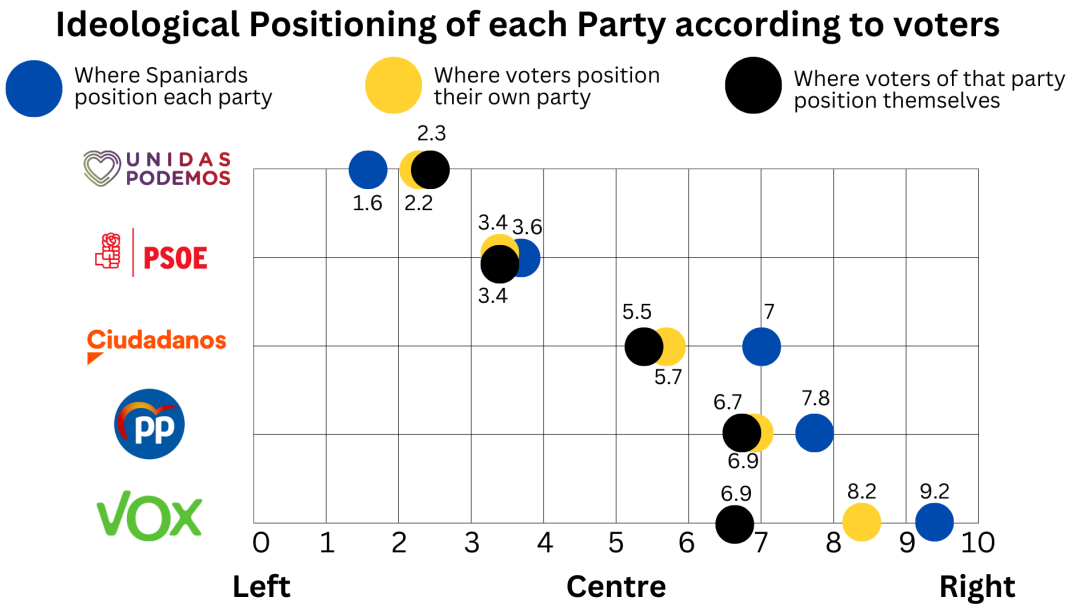

Figure 2.1: Visual representation of the results from *SocioMétrica*'s survey for the newspaper *El Español* [26]

These parties are represented across many municipalities, but each locality also holds many different parties specific to their area. The list of parties used for the collection of data are available in the appendix.

2.2 Locality affairs

It is important to take into account that each city has their own history and political affairs. One of the key differences between the cities of study is that Valencia and Barcelona have their own co-official languages, Valencian and Catalan alongside Spanish. On the other hand, Madrid's only official language is Spanish. In terms of the political situation of these cities, we can consider Madrid and Valencia to be similar in the aspect of being divided between right and left parties. This is not the case of Barcelona, due to the Catalonia independence movement. A detailed and historical explanation for this movement is out of this research's scope, a concise summary of the current situation, from the elections perspective, will be attempted through a summary of The Guardians' article titled *Is Catalonia still dreaming of independence from Spain?* [27].

In October 1, 2017 the president of the Community of Catalonia, Carles Puigdemont defied Spain's government and courts by calling all citizens of Catalonia for a unilateral

referendum for the independence of Catalonia. The response of the government was to send thousands of Spanish police officers, whose violent attempts to stop the referendum ended up on newspapers around the world.

Days later, secessionist Catalan MPs voted to establish an independent republic. The government of the Spanish president, Mariano Rajoy, decided to sack Puigdemont and his cabinet, assuming direct control of Catalonia and order a new regional election. During the crisis of October 2017 a survey by the Catalan government's Centre for Opinion Studies, found that 48.7% of Catalans supported independence, while 43.6% did not. According to a survey conducted in 2022 by the same centre, 52% of Catalans now oppose independence, while 41% are in favour. Thus, being a polarising topic in the community of Catalonia.

In last regional elections, pro-secessionism parties won an overall majority of the popular vote for the first time – 51% – but the party that took the biggest share of the vote was the unionist Catalan branch of PSOE - Catalan Socialist party (PSC). Eventually two independentist parties, ERC and Junts, formed coalition to govern on the community of Catalonia [27].

Today, Catalonia's indpendentist movement representation is divided into more parties than ever, as per *Radio Televisión Española* [28], being represented by five different groups ERC, Junts, CUP, Espai CiU and PDeCAT. This comes to show that the elections of the municipality of Barcelona, falls in the political spectrum of left and right, but also in that of independence and unionism.

Chapter 3

Literature Review

3.1 Introduction to Methodologies

Numerous studies have been conducted to predict elections in Spain, employing various approaches such as Volumetric and Volumetric with Sentiment Analysis [29, 30, 31]. The former relies on quantifying candidate support through the counting of party and name mentions, while the latter considers sentiment analysis to gauge positive and negative support for each candidate. Additionally, certain investigations take into account the popularity of tweets, indicated by metrics like retweets, likes, and favourites, to emphasize comments with greater community impact.

Among the different methodologies, the Volumetric with Sentiment Analysis approach has demonstrated superior performance, as highlighted in Grimaldi et al.'s study [32]. The effectiveness of prediction systems are commonly evaluated by computing the Mean Absolute Error (MAE) between predicted share of votes for each party and the actual election results. Grimaldi et al. achieved a remarkable MAE of 2%, representing the lowest error score observed in Spanish elections. Alternatively, some studies adopt a more conservative approach, focusing solely on predicting the winning candidate, deeming the prediction successful if it accurately identifies the election winner. In this study, we will use a three-class political orientation prediction - right, left or none. With this method it is not appropriate to use MAE for the parties, as we are not concerned with specific parties, but rather with ideologies. Therefore, the metric we will use is the winning ideology and a MAE for the two ideologies.

3.2 Methodology for Data Collection from Twitter

In the data collection process, most studies, irrespective of the country under study, use a similar methodology for Twitter data extraction. Specific sets of keywords associated with each candidate or party are used to identify relevant comments. These keywords are sought within the tweet text, mentions (in the format of @*mention*), and hashtags. The keywords may have variations of candidate names and party designations, including surnames alone. Moreover, other studies also consider the popularity of tweets as an indicator of impact. In such cases, tweets are collected 24 hours after publication to standardize the time frame for popularity assessment, capturing metrics like favourites, likes, and retweets [19]. We understand that simply relying on YouTube comments with mentions of political parties or candidates' names, may not produce sufficient comments to generalize over the large populations of the city. We can assume this, as per the preliminary tests we conducted by emulating the tweet collection procedure on the YouTube API. Instead, as we will explain in detail in the section 4.1.2, we will search for keyword matches in YouTube video titles.

3.3 Using YouTube Data for Election Prediction

While YouTube has been employed as a data source for election prediction in prior research, it has primarily been used to estimate candidates' popularity based on the number of views of their campaign videos [33]. However, our project focuses on analysing YouTube comments. Recent research has explored the relationship between users engaging with political and news content on both Twitter and YouTube [12]. This study discovered correlations between the two platforms, including user preferences, sentiments, and interactions, suggesting YouTube's potential as an economical alternative to Twitter data for election predictions in Spain.

3.4 Integration of Deep Learning Models and Political Orientation Systems

Despite the growing interest in election prediction, a general survey [9] pointed out the scarcity of deep learning models in this domain. Furthermore, studies concerning Spanish elections have not fully leveraged political orientation systems for predictive purposes. However, a recent competition in Spain focusing on author profiling for

political ideology demonstrated promising results with deep learning methods achieving a 96% macro f1 score for binary (left, right wing) classification. We intend to improve this current state of the art method, by adding the *none* ideology class, and exploit the suspected relationship between tweets and YouTube comments to test the resulting model on comments.

Chapter 4

Methodology and Experimental Setup

The data collection process for many studies have involved the use of an API to extract information from Twitter. However, this research introduces a novel approach by utilizing YouTube as an alternative to Twitter. The shift towards YouTube was prompted by recent changes in Twitter's API pricing, leading research to explore other platforms.

To collect comments, we define a specific collection period and create queries to search for relevant videos on YouTube. Once all relevant videos are identified, the comments for each video are collected. The retrieved comments, alongside the data from each video and user, forms the dataset for each city.

4.1 Data Collection

4.1.1 Collection period

Setting a collection period is the initial step in data collection. Previous studies faced limitations in this aspect, Gayo Avello [34] pointed out that using the Search API on Twitter restricted tweet searches to a few days before the search date. This requires the academics to perform searches daily throughout the election period. The YouTube API, however, allows searches within a chosen time window, alleviating this constraint.

The duration for comment extraction is arbitrary, with no established consensus. Some studies, like the one predicting the outcome of the 2016 Spanish general elections [35], used a 20-day period, while others, such as a study by Alonso González [20], collected data for a month.

In order to decide the period in which to collect comments, we looked at a study showing the common characteristics of Tweets and YouTube comments in the context

of the 2020 US presidential elections [12]. This study shows the number of tweets and YouTube comments in a timeline leading to the day of the election.

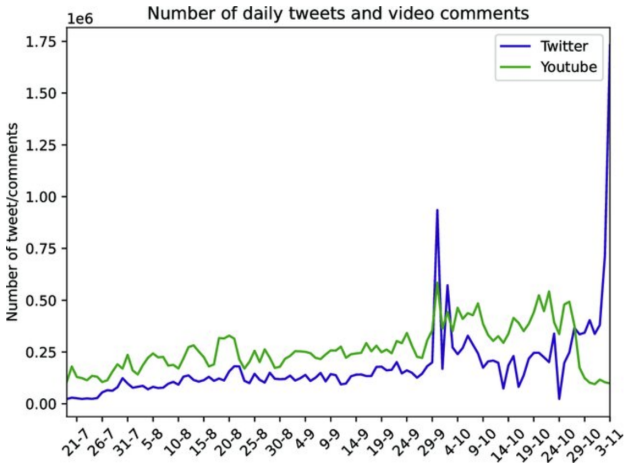

Figure 4.1: Number of daily tweets and comments on YouTube videos. From Shevtsov et al. study [12].

The graph in Figure 4.1 shows a large peak in tweets the days before the election. Whereas the number of comments decreases as it reaches the election day. Based on this analysis, we opted for a two-month search window (28th of March to 28th of May). This gives enough time to capture the start of the campaign for all parties and it also allows us to extract a number of comments capable of generalizing over the city of study. The day of the election is also included because the outcome of the election is not known until the day after. Such large window would not be appropriate for a study of general elections, but in the case of municipalities, data is scarce, so a longer window is required. The downside of this time period is that it will involve a thorough cleaning of the datasets as it opens for more noise to be extracted alongside relevant data.

4.1.2 Collection Method

The most common collection method involves using candidates' or parties' names as keywords to filter through the Search API. This approach is equally applicable in this study, as the YouTube API relies on queries to search for videos. However, there is a slight difference in that we will not look for mentions to candidates or parties in comments, but rather, in the video titles and descriptions.

Gayo Avello [34] critized in their analysis of election prediction that many studies limited the candidate monitoring to those that are popular and therefore more likely

to cast votes. For instance, in the study for the Spanish General Elections of 2016 [35] only four candidates were included in the research, whereas a total of 11 parties participated. The way in which some studies have decided which candidates to search is based on those who belong to parties that previously had a seat in parliament. In that study, the effect was not major, however this would have affected studies that were done on the 2019 Spanish general elections, where a previously not seated party, Vox, was largely voted and achieved 10.3% of seats in parliament [36]. In order to avoid this scenario, we assumed no previous knowledge and included all candidates.

One way in which YouTube API works is through queries. By searching with a query you can build complex data requirements for the API to look through its database. Initially, we followed the API guidelines and built a query that concatenated the candidate names and parties with OR operations. This method retrieved a small number of comments and through experimentation we found an alternative way to operate with the API. For each city, we built a separate query for each candidate and party, both having the same format. This was the chosen format:

"CANDIDATE_NAME or PARTY_NAME Elecciones CITY_NAME 28M"

This format is arbitrary, and we can not be certain it is the best to extract the maximum number of videos. Testing different formats would be an expensive task, as there exists many combinations in which a search query can be written, hence we decided to leave it out of the scope of this study.

The described approach resulted in numerous queries for each city. Many of the retrieved videos were duplicated due to the similarity of the queries. To avoid repetition of comments in later extraction, we kept track of the processed videos and deleted the duplicates, leaving only unique copies.

In addition to the search through queries, YouTube's API provides other filtering methods to aid search. Table 4.1 shows these settings.

Filter	**Setting**
part	id, snippet
maxResults	50
type	video
videoDuration	any
publishedAfter	28/03/2023
publishedBefore	28/05/2023
relevanceLanguage	es (Spanish)

Table 4.1: Table showing the filters and their respective settings for the retrieval of videos through the YouTube API.

The *part* and *video* settings are related to our method of collecting comments, first collecting videos, storing their ids and then looking for comments in those videos. By setting it to *"id, snippet"*, we can get the trivial information of that video such as the title, author and id. *maxResult* was set to its maximum: 50. *videoDuration* was set to *any*, as it allowed us to manipulate conventional YouTube videos as well as finished live transmissions and YouTube Shorts, which have an upper limit of 60 seconds of duration [37], and are getting increasingly popular, according to the Digital Information World [38].

During collection we noticed the quota reached its 10,000 requests per day limit several times. To solve this we kept a log of the processed queries, to continue search starting from the last processed query on the next day. This was then optimised by obtaining several keys from different accounts. Alternatively, a paid plan to use the API can also be used. It is also worth noting that the reproducibility of the collection will not be exact due to users possibly deleting their videos and comments.

As described earlier the *type* filter was used to extract the video ids. Now that we have them, we can retrieve the comments for that video. This involves making *commentThreads* request through the API. This type of request requires a different set of filters. The main differences are listed in Table 4.2.

Filter	**Setting**
part	snippet
videoId	videoId
typetextFormat	plainText
order	relevance
pageToken	next_page_token

Table 4.2: Table showing the filters and their respective settings for the retrieval of comments through the YouTube API.

In Table 4.2 *videoId* and *next_page_token* are variables representing the unique identifier of a specific video and the identifier for the next page of comments, respectively. The *pageToken* filter indicates in which page of the comment section we are, to keep track of comments.

After collecting the data, the resulting dataset for each city had the following columns: Video ID, Video Title, User ID, Comment, Likes,Comment Date, Video Author, Video Date. Some of these attributes will only be used for data cleaning.

4.2 Data Cleaning

Due to the extended collection period, a meticulous cleaning process is vital to eliminate unwanted comments from the dataset.

4.2.1 Date and Comment filtering

The first filtering step focuses on the properties of comments themselves. While the video collection period was predetermined, the comments' timeline remained open-ended, resulting in some comments being published after the election day. Additionally, we encountered repeated comments by the same author at the same timestamp, suggesting the processing of certain videos multiple times. Furthermore, we identified empty comments without text content. All three groups were removed from the datasets.

4.2.2 Video Title filtering

The title of the video serves as a concise summary of the content of the video. There are certain keywords that we can find in video titles that provide key information in terms of how relevant it is to our study. For example, if a title contains the name of a city that is out of our study, then we can expect the comments from that video to not be relevant.

The filtering through video title starts by keeping the relevant keywords, such as the name of the city, the name of candidates and the name of the political parties. This removes a large share of comments, however it still leaves many unwanted ones because the keywords are used across different contexts. For example, for the Barcelona elections, one of the political parties is *PP*, so videos containing the name of this party remain. However, this party also participates in the general elections which occur in late July. This is a title from a video collected:

"¡ABASCAL ESTARÁ POR DELANTE DE FEIJÓO EN LAS GENERALES SI EL PP CUMPLE EL "PLAN SECRETO" DESVELADO!"

The party *PP* appears but it appears in the context of the general elections (*"LAS GENERALES"*). Therefore, the comments of this video are not relevant to our study of the Barcelona municipality elections. A selection of keywords are made to remove the comments from the titles that contain them. These included names of other cities (*Murcia, Sevilla*), words related to football (*Vinicius*, *Gol*), and others (*Generales*). These keywords were selected by choosing the largest cities, political actors and parties outside of our study. We also included some by analysing the video titles through a

word cloud, which allows us to see the most salient words.

4.2.3 Video Author Filtering

The number of video authors for the Madrid dataset, which is the biggest out of the three studied, had a approximately 90 authors. Most of them were well known news outlets, this left around 20 authors that were unknown. These could be researched one by one to understand the nature of their videos. The comments that belong to videos created by unwanted authors were removed. For instance, in the case of the Madrid dataset, there was a make-up artists that made a video with the following title: *"La première de La Sirenita en Madrid"*.

This title remained in the dataset as it contains the keyword *Madrid*, however the title describes the premiere of *La Sirenita* film (the little mermaid). These types of videos are removed through video author inspection.

Given there is no previous study suggesting data cleaning techniques for the creation of a political orientation dataset based on YouTube videos, this is the process we propose. It is largely based on the selection of keywords for keeping and removing videos, so it does not guarantee a completely clean dataset, but it removes a large amount of noise.

4.3 Dataset Analysis

For the readers convenience this section will only show images and graphs from the Madrid dataset, to avoid overloading this section with figures. The discussion will be general to all cities under study, specific comments for each city will be addressed with reference to the appendix, where the same figures shown here, for Madrid, will be shown for Barcelona and Valencia.

4.3.1 Video and Comment Timelines

In the collection period section we saw trend of comments published the days leading to the voting day. To compare our dataset to the trend found in the Shevtsov et al. study [12], we constructed a timeline showing the number of videos and comments published in our collection window.

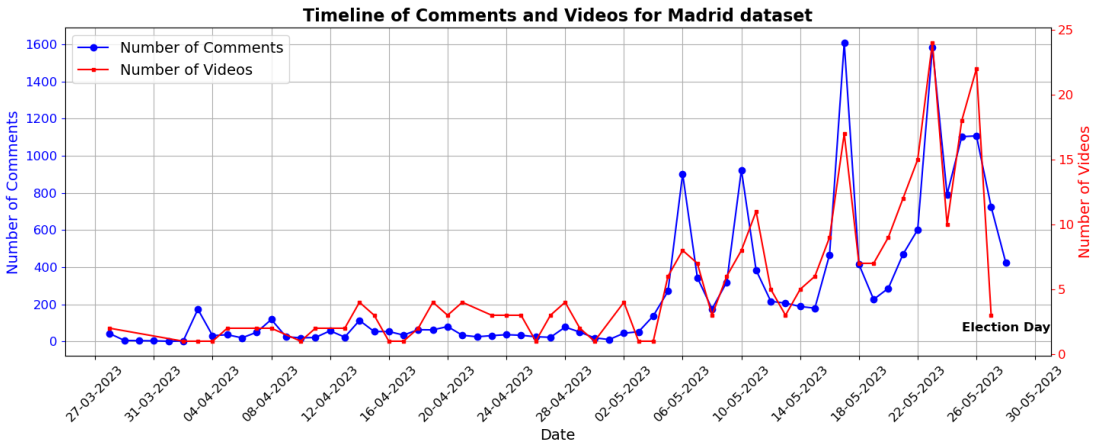

Figure 4.2: Graph showing the timeline of retrieved comments and videos for the municipality elections of Madrid

Our graph in Figure 4.2 shows similarities to that of the mentioned study, as we approach the month of the election there is a clear increase in the number of content generated in YouTube. There is also a clear relationship between the number of videos published and the number of comments found. These are shared across the three datasets - Madrid, Barcelona, and Valencia. There exists a minimal discrepancy between the study and our results, the falling in comments published by the study seems to begin a week before the election, whereas in our datasets the decrease appears to be only a few days before. Inspection of these videos suggested they were still relevant to the elections and not outliers. For example, this is the title of a video on the Vox candidate for Madrid, Ortega Smith, published one day before the election: *"Ortega Smith y Rocío Monasterio aprovechan la jornada de reflexión para estar en familia"*.

4.3.2 Word Frequency

The word frequency of comments is of special importance in studies related to election prediction, as the frequency of mentions for each political party was one of the first methods described. The percentage of mentions for each party represented the share of seats each would obtain. Usually, these included hashtags (more common in Twitter), candidate names and party names. Despite the renowned issues of this method, such as, how to deal when a tweet names more than one candidate or party, or the unknown sentiment the tweet expresses on the named entity, this technique has shown good Mean Absolute Error (MAE) by researchers.

We constructed two word clouds, one for the video titles to understand the topic

of the videos, and another one for the comments to identify the topics debated in their comment sections. These do not include stop words. A word cloud visually represents text, with word frequency determining word size. Common words appear larger, allowing for quick topic discernment.

Figure 4.3: World cloud showing the frequency of different words in the video titles in the Madrid dataset

The name of several candidates appear with large size, such as, Rita Maestre for the party Podemos, Almeida for the party PP, and Ortega Smith for the party Vox. It is interesting to see how the candidate for *Ciudadanos*, Begoña Villacís, does not appear in large or small size. The results in this municipality election of Madrid, for this candidate, were much lower than expected, only achieving 1.65% of the total votes, and 0 seats [39]. In the previous municipality elections the same candidate obtained 19.13% of the votes, obtaining 11 seats [40].

There are also other names that do not belong to the candidate list, such as Ayuso, Alejandra Jacinto and Monica García, these are politicians strongly linked to their political parties. We can argue that the videos are more focused on candidates than on parties, as the party names appear much smaller than their respective candidates. Other words give a hint as to how the videos are formatted, these are the words *"entrevista"* (interview), *"debate"*, *"elecciones"* (elections) and *"campaña"* (campaign).

We can compare this word cloud to that of the comments.

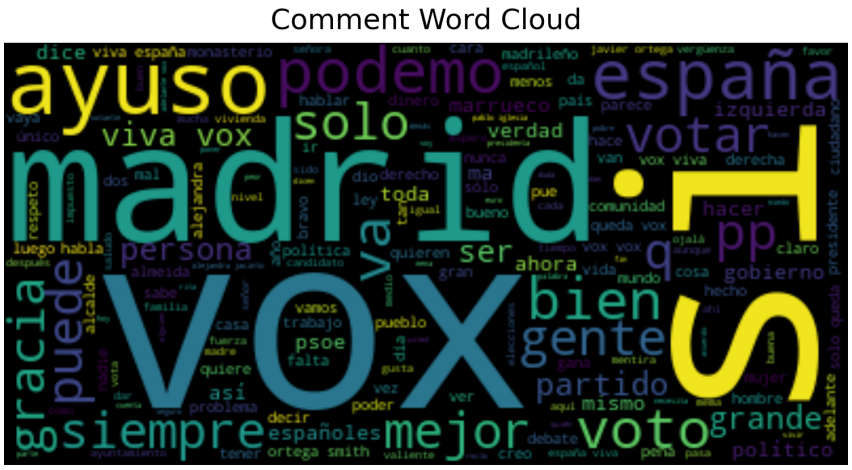

Figure 4.4: Word cloud showing the frequency of different words in the comments in the Madrid dataset

In this case, we can see in Figure 4.4 how the far right party *Vox* is much more salient in the comments than any other party. Even the phrase *viva Vox* is quite visible in the word cloud. Contrary to earlier observations where limited visibility for *Ciudadanos* candidate, Begoña Villacís, suggested fewer votes, the most mentioned party, *Vox*, did not get the most votes. Instead, it ranked fourth after *PP*, *Más Madrid*, and *PSOE* [39]. This finding supports the study of Metaxas et al., in which they argue that predictions derived from volumetric methods are as accurate as random chance [13].

The word *"voto"*, is the first person present tense for the verb *vote* in Spanish, and its also the noun *vote*. Having this word with high frequency in the comments could mean that many users are talking about their intention of vote, which is a good signal as this would make their discourse more explicit, and in turn make its political orientation prediction simpler.

Earlier, we commented that the first studies conducted for election predictions used the frequency of names as an indicator. In this dataset of comments for the municipality election of Madrid, our prediction would be skewed towards the right winning, as the name frequency for right wing parties is higher than those for the left. The highly salient parties and candidates do not directly indicate the intention of vote as the comments might be positive or negative towards them.

The comment word clouds for the other cities, Barcelona and Valencia, where really similar, having the party *Vox* as the most salient word, even higher than the city names. Again, this party did not cast the majority of votes in those cities. For Barcelona, the word *España* (Spain) appeared much more frequently than in the other cities, possibly

due to the polarizing topic of the Catalan independence movement [27].

4.3.3 Length Distribution

To better understand the nature of comments, analysing their length can be informative. Typically, shorter comments may pose classification challenges due to their limited information, while longer comments might be easier to classify because they provide more context. The bar chart in Figure 4.5 displays the frequency of comments based on their word count.

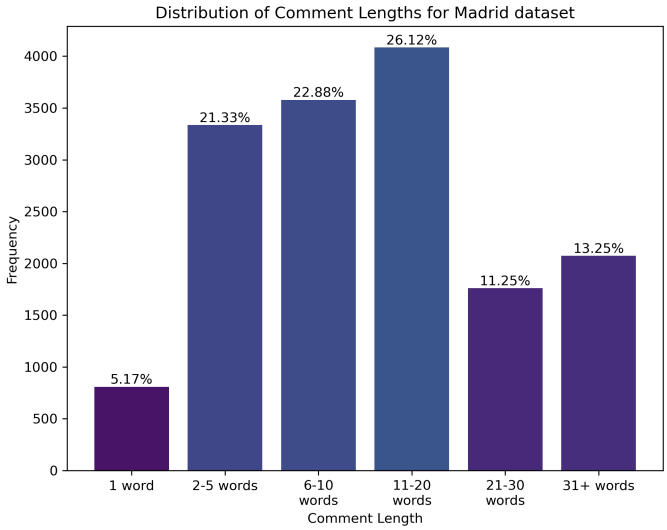

Figure 4.5: Bar graph showing the distribution of comment lengths for the Madrid dataset

This word distribution is shared between all datasets, with very similar percentages in each sentence length. It is interesting to see that the comments are spread in half between less than 10 words and more than 10. Approximately 30% of comments have a low count of words, less than 5. These are probably comments that assume the context of the video. For example, take these comments from the Madrid dataset:

Comment 1: *Buenísimo!* , Comment 2: *Ayuso presidenta.....*

The first comment can be translated to: *Great!*. It is impossible to know what the user thinks is great, and therefore what the orientation of the user is, without knowing more about the video. On the other hand, comment number 2 does give more information in a short comment, by naming a politician in a right wing party (*Ayuso*) and calling her *president*, in this case of the community of Madrid. It is still possible to interpret comment number 2 as sarcasm, if for example, the video the comment belongs to, is

about the politician being involved in a scandal. In both cases, the more information we have about the context, the better estimate we can make for their ideology.

By solely analysing the comment, we may miss fundamental information that was shared in the video, and that the user is referring to. A transcription of the video would not be ideal as there are long videos of 2 hours or more which would make it extremely difficult to locate the part the user mentions. A less resource intensive, but still effective option, is providing the title of the video. For the mentioned two comments the video title is:
Video Title: *La PARODIA de BOLAÑOS colándose en la fiesta de AYUSO — 2 de Mayo, día de la Comunidad de Madrid*
This title translates to: *The parody of Bolaños* (a politician belonging to the left wing party PSOE) *trying to get into the party of Ayuso* (a politician of the right wing party PP) *— 2 of May, day of the Community of Madrid*. With this new information we can better understand the position of each comment. The first one doesn't position on any wing, but rather simply enjoys the parody. On the other hand the second comment could be positioned on the left wing, due to a sarcastic comment of Ayuso being president.

For this reason, when processing the comments we will include the Title and Author, to provide as much information as possible to the model.

4.3.4 User Comment Frequency

We calculated the comment frequency by collecting the usernames of the authors of each comment. This will allow us to identify if a small group of people is publishing the majority comments, making the dataset less likely to generalize. As we can see in Figure 4.6, this is not the case for the comments found, most of the comments are being published by different users, and only a small proportion are commenting more than 5 times.

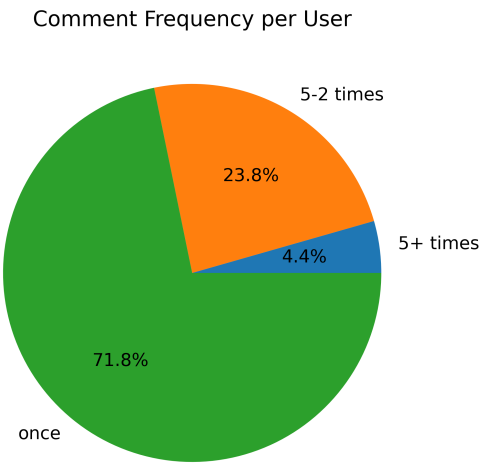

Figure 4.6: Pie chart describing the comment frequency per user for the Madrid dataset

We can consider this distribution good for generalization, as there are many users being represented in the dataset. However, the more a user comments the more likely we are to understand their political ideology. The dataset stands in a compromise between these two factors, where the ideal scenario involves many users commenting many times. This pie chart is similar to those of Velencia and Barcelona, however, these have a lower percentage of users commenting more than 5 times, making prediction per user slightly more difficult for that group of people.

4.3.5 Author Attribution Analysis

To understanding the nature of the comments we can research the video authors. An ideological investigation of the video authors could help us understand which political stances are creating videos which users respond to. We have classified these authors as: news outlets, political parties, politicians, other political video creators, non-political video creators and independent journalists.

The class of *other political video creators* constitutes creators which main topic of discussion is politics. For example, the YouTube channel *Noticiero Pijoprogre* uploads videos discussing the arguments of politicians in their interviews with other journalists. The *non-political video creators* class, takes authors whose main theme is not politics, but have however uploaded a video on this matter. Independent journalists are individuals known for talking about politics, but are not linked to a particular party or newspaper.

These have been classified manually, as the highest number of authors for a particular

dataset was only 90, allowing us to research them individually. Figure 4.7 shows two pie charts. The first one is the share of video creators, so if there are 100 creators and 25 belong to the class of politicians, then this category would have 25% representation in this pie chart. The next pie chart is the distribution of comments by author type, so if the same politicians class, have produced videos that collected 100 comments, and the total number of comments in the Madrid dataset is 1000, then this class would be represented with a 10% share.

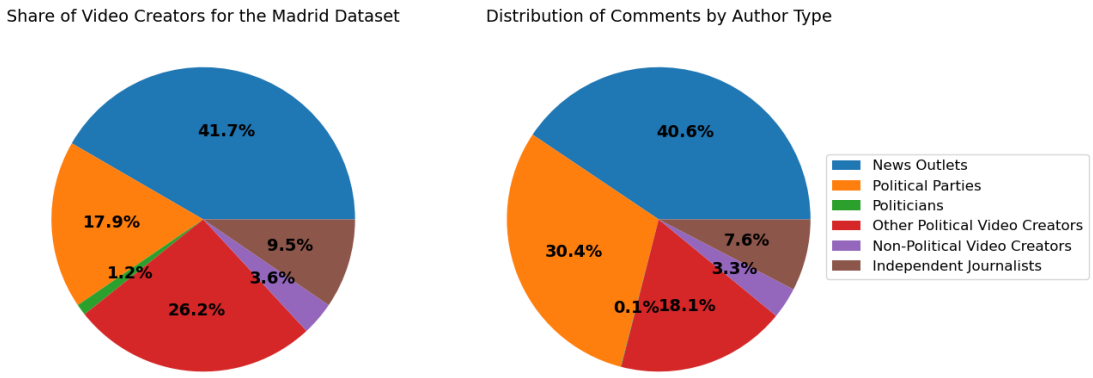

Figure 4.7: Pie charts showing the share of video creators and the distribution of comments by author type for the Madrid dataset

It is clear that most author types have a total number of comments proportional to their number of creators, with the exception of the political parties. These, despite forming only 17.9% of the number of authors, have produced 30% of the comments in the Madrid dataset.

To build an ideological chart showing the political stance of each video author, we used the results from the study conducted by Guerrero Solé in 2022 on *Measureing the political leaning of Spanish news media through Twitter users' interactions* [41]. This study uses a combination of the method of Retweet Overlap Network and a sociological study by CIS on how the Spanish population perceives the political parties, to propose an ideological thermometer of the Spanish news outlets. This study covers the classes of news outlets, political parties, and politicians. The other categories are classified manually based on their publications. Some, may not appear evident, to avoid classifying with uncertainty we will leave this as unknown. The resulting distribution is shown in Figure 4.8.

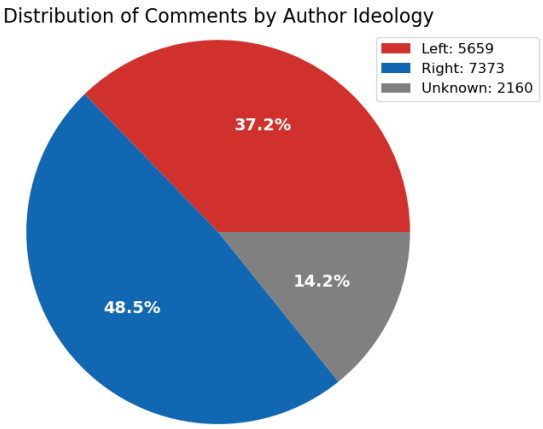

Figure 4.8: Pie chart showing the distribution of comments by author ideology on the Madrid dataset

The distribution shows there are 1714 more comments attributed to authors that are right leaning, constituting 11.3% more of all comments compared to the left. This does not entail that the comments will have this share of ideology. There remains to classify the ideology of video authors to which 14.2% of the comments are attributed to. We are unable to classify these authors due to not showing partiality in their videos, or having only very few videos on politics.

4.3.6 Language Analysis

As we explained in section 2.2, Valencia and Barcelona have the co-official languages of Valencian and Catalan, respectively. For this reason, during the collection process we filtered out comments by Spanish and by the co-offical languages of the city for which we are collecting. However, the filtering is done by relevance and is therefore not completely thorough. The YouTube API *relevanceLanguage* filter is described as follows: *"return search results that are most relevant to the specified language"* [42].

Therefore, to conduct a strict analysis of our collected data, and to inform future studies on what to expect when collecting political-related comments through the YouTube API, we used the language detection model, FastText [43] to count the distribution of languages in our dataset.

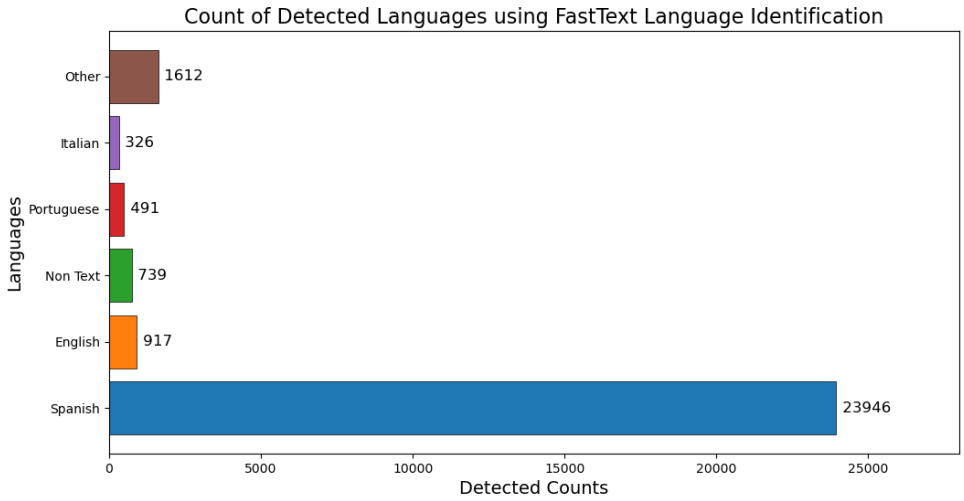

Figure 4.9: Graph showing the count of detected languages for all YouTube collected datasets (Madrid, Barcelona, Valencia).

As we expected, Figure 4.9 shows a much higher share of Spanish comments than any other language. English, Portuguese and Italian are the other most common, but with much lower count than the Spanish. However, we captured examples of the different languages detected with FastText to re-evaluate the counts and make more accurate estimations.

This process revealed that a significant proportion of the comments, which are mostly spoken in the Mediterranean area, were written in Spanish, yet they contained words that are identical in form and meaning to other languages. For instance, we came across a comment in Spanish text misclassified as Portuguese, where the comment read as follows: *"Osea, dar contratos publicos a su marido."* In this case, the words *"contratos"*, *"publicos"* and *"marido"* are nouns that share the same written form and meaning in both Spanish and Portuguese. It is also worth noting that the Non-text category mainly holds comments solely written with emojis.

Furthermore, we highlighted observations in this domain, such as instances of Spanish text laughter representation being misidentified. One such comment contained the expression *"Jajajaja"* which was erroneously detected as Indonesian due to how the word "yes" is written in this language: *"Ja"*. These findings, along with others of interest, have been documented in the appendix for further inspection.

The examination was done manually, but not all comments were checked. Despite this we found that the languages, other than Spanish, constituted less than 1% of the total comments in the Madrid dataset. Within these we found Arabic, Greek, Basque,

Catalan, English and others. The Valencia and Barcelona datasets showed a higher proportion of Catalan comments, but the Spanish language remained within a share of 98% - 99% of the total number of comments.

4.4 Building a Political Orientation Model

4.4.1 Collecting Training Data

The aim of the model we want to build is to predict the political orientation of users based on their comments to certain YouTube videos. The best training data we could hope for is labelled YouTube comments with all of the information of the video, such as video title and video author. Currently, this type of dataset does not exist. Therefore, by relying on the transeferability of tweets to YouTube comments [12], we can use existing labelled Twitter datasets for this research.

We decided to build a custom dataset from three Twitter datasets:

1. Tweets from Spanish politicians of the following parties: PSOE, PP, VOX, Unidas Podemos and Ciudadanos [44]. This dataset contains 245,790 tweets. The support for each party is the following:

   - PSOE: 61,404 (24.98%)
   - Vox: 51,505 (20.95%)
   - Ciudadanos: 38,592 (15.70%)
   - Partido Popular (PP): 45,481 (18.50%)
   - Unidas Podemos: 48,808 (19.86%)

   If we split the parties into left or right orientation, according to the results of the study for *El Español* where supporters place each other in the political stance spectrum [26], by the middle point of 5, then the share is 44.84% left and 55.16% right. These tweets come from a group of 142 politicians, with an average of 1731 tweets each.

2. Tweets from Spanish politicians and journalists used by the team Los Calis in the PoliticES challenge in which they achieved a micro-f1 score of 96% [17]. This dataset has 361,646 comments with an orientation split of 52.21% right and 47.79% left. Similar to the previous dataset, it is a small group of users writing, this time its 430 with an average of 841 comments each.

3. Given the nature of the YouTube comments we collected, there might be users which do not show a stance in the political orientation spectrum, or simply do not talk about politics. With the previous two datasets, we can only classify comments as left or right wing. Therefore, to allow for non-political speach in classification we introduce the *None* label, to target comments not concerning politics. The tweets used for this class come from a dataset on Spanish sentiment classification [45]. The full dataset has approximately 600 million tweets. We retrieved a sample of 3 million tweets. From this sample of tweets we observed some comments writing about Spanish politics. It is not possible to determine the exact amount, but certain searches retrieved thousands of comments. Therefore, to remove these from the *None* category, we made a list of the most common words used in political campaigns [46] (according to official transcriptions collected by El Periódico), and the most popular political actors - including the party leaders and the name of their party. Then to level the three classes of left, right and none, we extracted a random sample of tweets from the cleaned dataset. We did this in a way that it produced a similar number of users with a comparable average to the ones seen the previous datasets, and therefore producing a balanced dataset in terms of labels and users. We collected 288,000 comments, from 288 users, each commenting 1000 times.

The resulting dataset has a total of 895,436 tweets. The representation of the left is 283,197 (31.62%) comments, 324,239 for right (36.21%), and 288,000 for none (32.16%).

It is important to understand that sentiment towards different topics can provide information as to which ideology a user follows, this was tested by Bhatia et al. [47]. The tweets we introduced with the none class may contain sentiment towards these topics. Our cleaning process was focused on Spanish politically related issues. However, topics outside of this area may also be polarising. With this, we wan't to express the difficulty of the task of obtaining *politcally-null* comments.

4.4.2 Training Data Analysis

For this analysis, we did not include stop words, hashtags, mentions, and the attention token of Los Calis.

4.4.2.1 Left & Right ideology

The tweets of journalists and politicians formed a word cloud that shows certain words more salient than others such as *España* (Spain), *gobierno* (government) and *hoy* (today). Again, for this word cloud the Spanish stop words from the NLTK library were removed. There is no clear difference between the word clouds from the left and right wing tweets, it seems the difference could reside in smaller details or less salient words.

Figure 4.10

We attempted to identify the significant words, that are less common, but are distinctive for each ideology. We achieved this by ranking words through Pointwise Mutual Information (PMI). These wordclouds are available in the appendix. Among the words the left uses more distinctively than the right, are *"haiku"* and *"bdía"*. The former is a type of Japaenese poem, which term was only used by a left wing politician, who wrote a haiku daily, and started the tweet in this same way: *"Cada día, un haiku..."*. Similarly, the latter is an abreviation of the greeting *"buenos días"*, which was, again, only used by one particular left wing politician.

The distinctive words the right used were more politically charged, and focus on derogating characteristics of the contrary ideology. This shows a similar conclusion to that reached by Darwish et al. in their study analysing the tweets in the 2016 US presidential elections [48]. Their study revealed that messages from Trump, predominantly backed by right-leaning sources, were more adept at framing and criticizing Clinton, who had stronger support from left-leaning outlets.

The following analysis is made exclusively from what the tweets read:

- *"**multiculturalismo**"* and *"**globalismo**"* - When the right mentions multiculturalism or globalism, it refers to the immigration Spain receives. Specifically they criticise the lack of integration from immigrants, and deem the multiculturalism project a failure, sometimes even redefining it as *"invasión cultural"* (cultural

invasion). More on how the right comments on this topic can be seen in an article on Abascal, the president of the right wing party Vox [49].

- *"**narcocomunismo**"* - The term narcocomunism is used by the right to describe the political situation of some countries in Latin America, the most common one being Colombia, lead by president Petro [50].
- *"**sánchezstein**"* - This refers to the Spanish coalition government, which is lead by the socialist Pedro Sánchez, but its also constituted by the left party Podemos. We can assume that the addition of the termination "stein" is a play of words to derogate a government formed by different parties - a "Frankenstein" government [51].
- *"**separatista**"* - The *"separatist"* term is used in the context of Catalonia's separation from Spain. Politicians use it to warn people about *"separatist"* parties getting into government.
- *"**zaldívar**"* - Zaldívar is a city in the Basque Country, in the north of Spain. It is home to an environmental disaster that occurred in the year 2020, where tonnes of industrial waste broke out of a landfill killing two operators and invading two lanes of a major highway in the city [52].

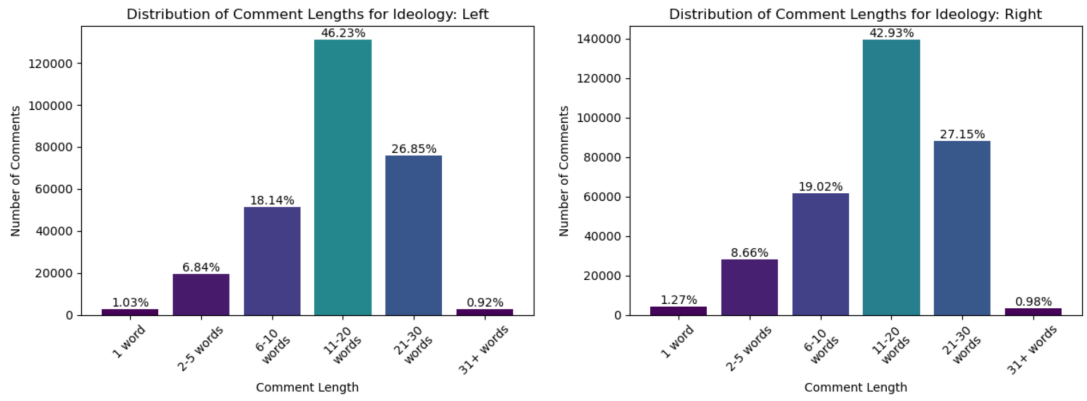

Figure 4.11

4.4.2.2 None ideology

The word cloud formed by the tweets classed as not supporting a specific ideology contains mostly temporal adverbs (*siempre*, *ahora*), common nouns (*hoy*, *día*) and

adjectives (*mejor*, *solo*, *bien*). It is clear that the words for the none ideology differ greatly from those of the left and right orientations. This could translate to an easy way of distinguishing between politically opinionated comments and non-political comments. It is also not possible to extract a salient theme from the wordcloud, which is within our expectations as these are tweets collected randomly.

Figure 4.12

The tweet length distribution for the none ideology clearly differs from that of the right and left orientation. The apolitical comments show shorter length compared to the political ones. This could be another factor our model finds patterns with to classify comments.

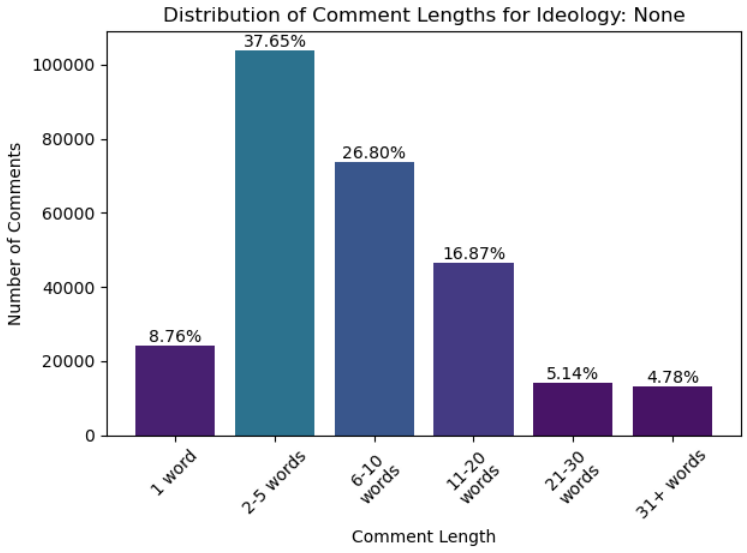

Figure 4.13

To put these findings into perspective here are some examples of the tweets corre-

sponding to each class.

- **Right**: *"Agradezco las aportaciones y las propuestas que ha realizado el alcalde de #Sevilla, user, en la reunión que hemos mantenido esta mañana. La buena sintonía entre administraciones es imprescindible para lograr el objetivo común y hacer frente al #Covid_19."*
- **None**: *"user Vi tus fotos hermoso lugar !!"*
- **Left**: *"Esta mañana el user fija posición y debate muchos temas importantes que afectan a nuestro país en la Comisión de #AsuntosExteriores. Muchos retos comunes que exigen la cooperación entre países"*

These are randomly chosen examples, but it comes to show how the opinionated political comments can provide more information than those classified as none.

4.4.3 Choosing a Model

The study we want to emulate with this experiment is that of Los Calis for the PoliticES competition [17]. This model consists of three parts:

1. **Comment representation** - They use a concatenation of two Spanish pre-trained models equivalent in structure to BERT and RoBERTa. These are known as BETO [22] and MarIA [23]. We use an uncased version of BETO as the comments we will process from YouTube do not follow conventional guidelines on casing. This was noted through observations such as comments including cities without capitalisation. The collected comments are transformed into tokens that feed into these two models. Each model generates a classification token [CLS], which is a 768-dimensional vector representing the meaning of the comment. Both vectors are concatenated and fed into the classification layer.
2. **Classification Layer** - Despite the larger size of our dataset, we started our experiments with the same complexity compared to that of Los Calis - three layers of 768 units, with a *tanh* activation function. The last layer feeds into a sigmoid activation function which returns the probability score for each ideology class.
3. **Voting System** - After each comment has been classified by ideology, to determine the political stance of a user, each of their comments are aggregated and the modal class is chosen to classify that user.

These three parts make the system that predicts the political orientation of a user based on their comments. However, before feeding the text to the BETO and MarIA models to build the representations we must first clean the comments from possible noise and standard characters used in Twitter, which will allow us to normalise text and therefore generalize better. The cleaning procedure is carried out by using the following pre-processing methods:

- **HTML entities removal** - Due to aggregation of training data from different sources, we found many rows with quote imbalances that were resolved automatically by substituting them with HTML entities. Some of these entities include: `&quot;`, `&#34;`
- **Normalising mentions** - Mentions are common in tweets, these take the form of `@user123`. To standardise them we rename the users to *user* removing the @. So, `@usuario123` turns into `user`.
- **Removal of URLs** - This is a common pre-processing technique used for tweets, including in election prediction papers that use tweets [32, 19]. Many tweets reference or react to external content, and links in the form of text do not provide relevant information so they muust be removed.
- **Removal of excess white-space** - This is another popular pre-processing technique. When handling text written in social media there is a chance that some comments contain excessive whitespace that does not contain useful information.
- **Processing camel case hashtags** - In previous papers, such as that of Singh et al. [35] hashtags were removed. However, we saw that many of the hashtags in our training dataset were used in place of people, cities or topics. These could portray relevant information for the prediction of someones' ideology. Therefore, we propose the following method of pre-processing hashtags, based on their common *Camel Case* structure. Through a regular expression we dismount hashtags into words separated by change in case. For instance, `#municipalityElections` becomes `municipality Elections`.
- **Reduction of repeated characters** - Inspired by Singh et al. paper [35], since we will deal with user-produced data, there could be cases in which repetition of characters is used for exaggeration or simply by mistake. This will result in differently written words that mean the same. We can normalize them by

reducing the repeated characters through a regular expression. We also take into account the natural occurring repeated characters in the Spanish language (c,l,r,n). When these are repeated we condense them into two. For example, in the Madrid dataset this comment: *"¿Si **toooodo** el mundo quiere primarias [...] **Belarra** a aseverarlo [...]"*, would be processed into: *"¿Si **todo** el mundo quiere primarias [...] **Belarra** a aseverarlo [...]"*

- **Lower case conversion** - Finally, convert all the text into lower case.

It is also worth noting that we considered the removal of stop words. We decided not, as according to a paper on the behaviour of BERT [53], stop words received as much attention as non-stop words, but their deletion has not effect in Mean Reciprocal Rank performances. Some stop words can also provide context, for example through negation words, which are considered stop words. In an ideal scenario, we would test the effect of their removal, but given the high cost of encoding the training set, we only processed it once.

Chapter 5

Results and Analysis

5.1 Training and Fine Tuning

Before training, we splitted the data into training (64%), development (16%) and testing (20%). We leverage our large collection of data to make generous splits. The first training experiment on the model described was carried out with the hyperparameters and settings described in Figure 5.1.

Parameters	**Value**
Learning Rate	5e-4
Training epochs	50
Activation function	Tanh
Layers	3
Dense layer units	768
Dropout	0.15
Optimizer	Adam

Table 5.1: Parameters and their respective values used in the first experiment of political ideology classifier

The results of this initial trial achieved a **75% macro f1 score** for the prediction of tweets, not of users. We would expect the prediction for users to be higher as true positive predictions accumulate and the they are voted as the most likely ideology for that user.

We followed this preliminary test with a grid search. We established the ranges based on the Los Calis study and our own observations from the first experiment. The ranges and optimal values can be seen in table 5.2.

Parameters	Value	Optimal
Learning Rate	[1e-4, 5e-4, 1e-3]	5e-4
Activation function	[Tanh, ReLu]	ReLu
Layers	[2, 3]	2
Dense layer units	[128, 512, 768]	512
Dropout	[0.15, 0.20, 0.25]	0.20

Table 5.2: Range of values, and their optimal outcome, for parameters in the political ideology classifier

The most significant change was the activation function. Tanh seemed to limit the macro g1 score to approximately 76%, whereas ReLu appeared to have a higher bound getting closer to 80%. We were surprised to see the little effect of dense layer units and number of layers within the proposed range. Simpler models performed within the same range of f1 scores as more complex ones. When debating which value for dense layer units and layers to choose, we favoured simpler structures when the results were within $\pm$ 0.05%.

We followed the training for the classifier with optimal parameters. We checked for overfitting by plotting the loss and macro f1 score over epochs. We chose to keep the model in epoch 48, as the development loss was its lowest and the macro f1 at its highest with **80.37%**. At this epoch there was no signs of overfitting, see Figure 5.1.

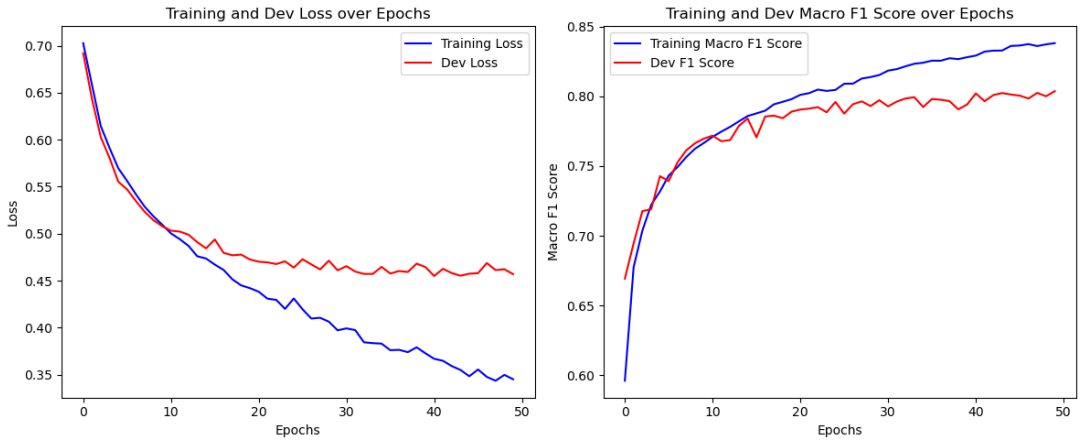

Figure 5.1: Graphs showing the training and development loss and Macro F1 scores over epochs

The chosen model of 80.37% macro f1 score in the development set, showed similar results when applied to the test set. Table 5.3 describes the classification report of the test set.

Classes	**Precision**	**Recall**	**F1-Score**	**Support**
Left	0.7340	0.7421	0.7380	56554
Right	0.7839	0.7351	0.7587	65109
None	0.8847	0.9376	0.9103	57425
Metrics				
Accuracy		**0.8022**		179088
Macro avg	0.8009	0.8049	**0.8024**	179088
Weighted avg	0.8005	0.8022	0.8008	179088

Table 5.3: Full classification report of best model on test set (per comment).

By utilizing pre-trained deep learning models and neural networks, we have enhanced the results from Prati and Hung's 2019 study. While their research achieved a 77% accuracy using Random Forest [15] for binary classification [14], our approach surpassed this by 3 percentage points, achieving an 80% accuracy even with the addition of a third category.

The reported results in table 5.3 are the most relevant to the purpose of election prediction using YouTube data, as the comment frequency per user in this platform is quite low, per our analysis in Figure 4.6. Therefore assuming full transferability between our model and YouTube we could expect an accuracy of 80%. However, the paper written by Los Calis team, showed their results on user ideology prediction, rather than prediction per comment. Therefore, we applied the voting system to categorize users on their ideology. Table 5.4 shows the results.

Classes	**Precision**	**Recall**	**F1-Score**	**Support**
Left	0.9611	0.9379	0.9494	290
Right	0.9487	0.9418	0.9453	275
None	0.9697	1.0000	0.9846	288
Metrics				
Accuracy		**0.9601**		853
Macro avg	0.9598	0.9599	**0.9598**	853
Weighted avg	0.9600	0.9601	0.9599	853

Table 5.4: Full classification report of best model on test set (per user).

In the task of user ideology classification, we've refined existing datasets and combined them to feed it into a revised version of the model built by the team Los Calis, to match their results, achieving a **macro f1 score of 96%**. Notably, this was done while incorporating a third category for non-political speech.

5.2 Testing on YouTube data sample

Throughout this study we have hypothesized that we can build a model trained on tweets, and use it on YouTube comments. This speculation arises from Shevtsov et al. study [12], in which they found that Twitter communities correlate with YouTube comment communities. From this finding, we draw the hypothesis that our political orientation model can be applied to the comments collected for the municipality elections of Madrid, Barcelona and Valencia.

To test this, we manually labelled a random sample of 100 comments from the Madrid dataset. In the process of labelling comments we only considered the author, video title and comment. This information was then condensed into a single string to feed the model all the data related to one comment. The format this string takes is the following: *Video de [Author]: "[Video title] [Comment]"*. This structure aims to emulate how a tweet would be structured when reacting to a video. The results of this test are shown in Table 5.5.

Classes	**Precision**	**Recall**	**F1-Score**	**Support**
Left	0.2963	0.2963	0.2963	27
Right	0.4286	0.3191	0.3659	47
None	0.1842	0.2692	0.2188	26
Metrics				
Accuracy		**0.3000**		100
Macro avg	0.3030	0.2949	**0.2936**	100
Weighted avg	0.3293	0.3000	0.3088	100

Table 5.5: Full classification report of model tested on a labelled random sample of 100 YouTube comments from the Madrid dataset.

The results show an extremely low score for both accuracy and macro f1 score, suggesting that the classification is completely random, as it is close to 33%. From this outcome we can draw with confidence that **our initial hypothesis cannot be accepted**, and therefore, that we cannot learn to classify YouTube comments having trained a model on tweets. Even when attempting this test with different structures, such as, only including the comments, or the video title and comments, the results were similar in that they classified at random level.

Since this experiment has proven to classify YouTube comments randomly, we are not able to adequately forecast the outcome of the municipality elections through political ideology prediction.

Chapter 6

Conclusion and Future Work

6.1 Conclusions

The first objective of this study was to devise a methodology for the collection of comments through the YouTube API. We used previous research on election predictions, that used the Twitter API, to guide us through the process of collecting data. We found a unique way of collecting comments for elections. This involved the elaboration of a query structure that takes as input the name of a candidate or a party. Once all parties and candidates had their own query, we searched each with the YouTube API, and the stated filters (including the collection window), to find videos relevant to the elections. Once the video ids were aggregated, we extracted all the comments from them. Following the collection of comments, we devised a cleaning process through three different attributes: the publishing date, video title and video author. We consider that this objective was met as we obtained around 25,000 clean comments for the the smallest elections held in the cities of Madrid, Barcelona and Valencia, in a platform that produces much less comments than Twitter, as seen in Shevtsov study comparing these platforms [12].

We also proposed to test the validity of YouTube comments for the task of election prediction, and confirm Mejova et al. study suggesting that YouTube is a rich source of politically opinionated data [11]. We achieved this by thoroughly analysing the collected comments. As we saw in section 4.3, the topics discussed in both the videos and the comments are heavily related to politics, as they included most candidate names and parties. Thus confirming the presence of politically polarised comments in YouTube. Additionally, we saw that the YouTube community is formed by many users commenting several times, which supports the idea that we can generalize over

this data as many users participate. We also investigated the nature of the comments by analysing the type and political ideology of the video creators. These were mostly, political parties and news outlets, with a slightly majority towards the right-leaning ideology. During the comment analysis we also found a small proportion (1%) of them being written in different languages, but the great majority in Spanish.

We were able to construct a large dataset by grouping data from three different sources. This dataset of 895,436 tweets, balanced within 5%, included users of right and left stance, as well as the none-political speech class. This newly introduced category, allows for comments that are not related to politics, to be disregarded from the final calculation of votes. After training and fine-tuning, we built a model that improved the results of the winning team Los Calis in the PolitcES competition of 2022 [17]. We achieved the same macro f1 score of 96% with the addition of the none ideology class.

Finally, our last objective was to test the hypothesis that a model trained with tweets can classify YouTube comments. We drew this idea from the outcome of Shevtov et al. study [12]. Our results obtained a macro f1 score of 29% for this task, thus proving that we cannot perform transfer learning from tweets to YouTube comments in the context of political ideology prediction. Following this outcome, we did not produce a predicted share of votes for each ideology, for the collected cities, as the classification is random.

6.2 Future Work

In this study we explored a possible route to start data collection through the YouTube API, focusing on the retrieval of videos by keyword searches in their title, and the posterior collection of their comments. It is difficult to imagine a plausible alternative for comment collection, as the YouTube API does not allow for comment searches by their content. However, we do consider that this procedure can be enhanced by crafting more queries that search for videos on political-related topics, and not necessarily as direct as the search for candidates and parties. Possible topics include healthcare, free markets, and other polarising topics. These could be adequately chosen with enough domain knowledge (of the specific elections).

We highlighted the difficulties of transfer learning from a model trained with tweets and applied on YouTube comments. One of the distinctions we expected between these platforms, was that tweets are reactionary to other tweets, through replies, allowing us to access that context. However, comments could be reactions to the video, in which we only have the title and description as sources of context. A deeper study on their

differences, in relation to their syntax and semantics, could help us understand the challenges of their transferability, and allow for a more appropriate methodology to be followed.

Another, more direct approach, that would catapult election predictions through YouTube comments, is the labelling of user data. If a large dataset of YouTube comments is ideologically classified by sociologists and political scientist, this would encourage academics to apply machine learning algorithms on this supervised data. We are aware of the difficulty this supposes and the large cost of the labelling process, therefore a thorough investigation on the transferability between Twitter and YouTube is preferred.

The ideology prediction model we described in this study, was trained with less than 1000 users, despite being formed by more than 800,000 tweets. This type of tweet frequency per user is not appropriate for election prediction, as we do not have a small group of people casting the majority votes, but rather many users with varying influences (identified through retweets or likes, as seen in other studies [19, 20, 21]). To better address the election paradigm, we must build a model trained with many users and if possible as varied in ideology and writing style as possible. This will not only make the model more robust to a variety of comments, but also generalize over a more realistic distribution of people. The reason this is difficult is due to the lack of labelled data. Most of the datasets we described here, and that we encountered online, are tweets made by politicians and journalists, which are a small group of users that are easily labelled, as they are usually linked to a particular party or ideology. This data covers the ideological spectrum, but does not address the writing differences among voters. Therefore, we encourage academics to investigate the writing style differences between politicians and other users of the same political orientation.

Finally, we would like to comment on the challenges of election forecasting through ideology prediction. This area is heavily reliant on the current circumstances of the country of study. The political spectrum is not universal to all countries and, more importantly, not general to all temporal landscapes. Contemporary left-leaning viewpoints might evolve or shift in the next half-decade. Similarly, the pressing needs of a country can sway the policies of a parties towards ideas that are divergent from their historical ideologies. All of these factors play an important role in the classification of a users' political orientation. We believe the understanding of both historical context and current dynamics is crucial to determine which side of the political spectrum (or its detachment from it) a user is in.

Bibliography

[1] European Union. Spain – eu member country profile.

[2] ECLAC United Nations. Spain - political and electoral system.

[3] European Parliament. Understanding the d'hondt method.

[4] Andalucia.com. Municipal elections in spain, May 2023.

[5] CIS. Estimación de la intención de voto.

[6] Jeff Manza and Clem Brooks. How sociology lost public opinion: A genealogy of a missing concept in the study of the political. *Sociological Theory*, 30(2):89–113, 2012.

[7] Andrew J Perrin and Katherine McFarland. Social theory and public opinion. *Annual Review of Sociology*, 37:87–107, 2011.

[8] Jennifer Bachner and Kathy Wagner Hill. Advances in public opinion and policy attitudes research. *Policy Studies Journal*, 42:S51–S70, 2014.

[9] Asif Khan, Huaping Zhang, Nada Boudjellal, Arshad Ahmad, Jianyun Shang, Lin Dai, and Bashir Hayat. Election prediction on twitter: A systematic mapping study. *Complexity*, 2021:5565434, 2021.

[10] Chris Stokel-Walker. Twitter's $42,000-per-month api prices out nearly everyone, Mar 2023.

[11] Yelena Mejova and Padmini Srinivasan. Political speech in social media streams. *Proceedings of the 4th Annual ACM Web Science Conference*, 2012.

[12] Alexander Shevtsov, Maria Oikonomidou, Despoina Antonakaki, Polyvios Pratikakis, and Sotiris Ioannidis. What tweets and youtube comments have in

common? sentiment and graph analysis on data related to us elections 2020. *PLOS ONE*, 18(1), 2023.

[13] Panagiotis T. Metaxas, Eni Mustafaraj, and Dani Gayo-Avello. How (not) to predict elections. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 165–171, 2011.

[14] Ronaldo Cristiano Prati and Elias Said-Hung. Predicting the ideological orientation during the spanish 24m elections in twitter using machine learning. *AI & SOCIETY*, 34(3):589–598, 2019.

[15] L Breiman. Random forests mach learn 45 (1): 5–32, 2001.

[16] José Antonio García-Díaz, Salud María Jiménez-Zafra, María-Teresa Martín Valdivia, Francisco García-Sánchez, L. Alfonso Ureña-López, and Rafael Valencia-García. Iberlef 2022 task - politices. spanish author profiling for political ideology.

[17] Sergio Santamaria Carrasco and Roberto Cuervo Rosillo. Loscalis at politices 2022: Political author profiling using beto and maria. *Iberian Languages Evaluation Forum 2022*, page 1–10, Sep 2022.

[18] World Population Review. Population of cities in spain 2023.

[19] Didier Grimaldi. Can we analyse political discourse using twitter? evidence from spanish 2019 presidential election. *Social Network Analysis and Mining*, 9(1):1–9, 2019.

[20] Marián Alonso González. Predicción política y twitter: Elecciones generales de españa 2015. *ZER - Revista de Estudios de Comunicación*, 22(43):13–30, 2017.

[21] Luis Deltell, Florencia Claes, and José Miguel Osteso. Predicción de tendencia política por twitter: Elecciones andaluzas 2012. *Ámbitos. Revista internacional de comunicación*, (22), 2013.

[22] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

[23] Asier Gutiérrez Fandiño, Jordi Armengol Estapé, Marc Pàmies, Joan Llop Palao, Joaquin Silveira Ocampo, Casimiro Pio Carrino, Carme Armentano Oller, Carlos Rodriguez Penagos, Aitor Gonzalez Agirre, and Marta Villegas. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68, 2022.

[24] Marie Chaput and Julio Pérez-Serrano. *La transición española*. Biblioteca Nueva, 2015.

[25] La Constitución española de 1978. La constitución española de 1978.

[26] Alberto D. Prieto. El 75podemos de extrema izquierda, Jan 2019.

[27] The Guardian. Is catalonia still dreaming of independence from spain?, Sep 2022.

[28] RTVE.es. El independentismo catalán llega al 23j más dividido que nunca, Jul 2023.

[29] Montserrat Fernández Crespo. Predicción electoral mediante análisis de redes sociales. *Ene*, 12:27, 2019.

[30] Juan M. Soler, Fernando Cuartero, and Manuel Roblizo. Twitter as a tool for predicting elections results. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1194–1200, 2012.

[31] José Rúas-Araújo, Iván Puentes-Rivera, and María Isabel Míguez-González. Capacidad predictiva de twitter, impacto electoral y actividad en las elecciones al parlamento de galicia: un análisis con la herramienta liwc. *Observatorio (OBS*)*, 10(2), 2016.

[32] Didier Grimaldi, Javier Diaz Cely, and Hugo Arboleda. Inferring the votes in a new political landscape: The case of the 2019 spanish presidential elections. *Journal of Big Data*, 7(1), 2020.

[33] Fabio Franch. (wisdom of the crowds)2: 2010 uk election prediction with social media. *Journal of Information Technology & Politics*, 10(1):57–71, 2013.

[34] Daniel Gayo-Avello. A meta-analysis of state-of-the-art electoral prediction from twitter data. *Social Science Computer Review*, 31(6):649–679, 2013.

[35] Prabhsimran Singh, Ravinder Singh Sawhney, and Karanjeet Singh Kahlon. Predicting the outcome of spanish general elections 2016 using twitter as a tool. *Communications in Computer and Information Science*, page 73–83, Jul 2017.

[36] congreso.es. Composición - congreso de los diputados.

[37] Google. Get started creating youtube shorts.

[38] Arooj Ahmed. Data shows, youtube shorts gives tough competition to tiktok soon after its global launch, May 2021.

[39] La Vanguardia. Resultado elecciones municipales en madrid, (pp) gana: última hora con el 100.0

[40] EL PAÍS. Resultados electorales en madrid: Elecciones municipales 2019, May 2019.

[41] Frederic Guerrero-Solé. The ideology of media. measuring the political leaning of spanish news media through twitter users' interactions. *Communication amp;amp; Society*, 35(1):29–43, 2022.

[42] YouTube API. Search: Listnbsp; —nbsp; youtube data apinbsp; —nbsp; google for developers.

[43] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[44] Ricardo Moya. Tweets política españa, Mar 2023.

[45] Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. RoBERTuito: a pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France, June 2022. European Language Resources Association.

[46] El Periódico. Las palabras más usadas en los programas de los partidos, Nov 2019.

[47] Sumit Bhatia et al. Topic-specific sentiment analysis can help identify political ideology. *arXiv preprint arXiv:1810.12897*, 2018.

[48] Kareem Darwish, Walid Magdy, and Tahar Zanouda. Trump vs. hillary: What went viral during the 2016 us presidential election. In Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri, editors, *Social Informatics*, pages 143–161, Cham, 2017. Springer International Publishing.

[49] Europa Press. Vox llama a frenar el multiculturalismo y cambiar las políticas migratorias para evitar que españa acabe como francia, Jul 2023.

[50] Lgi. Gustavo petro: El candidato del narcotráfico que amenaza a colombia, May 2022.

[51] Clara Pinar. Sánchez: “el gobierno de coalición tendrá varias voces pero caminará en una única dirección”, Jan 2020.

[52] Alberto León. Vertedero de zaldibar: Un año del derrumbe, Feb 2021.

[53] Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*, 2019.

[54] elEconomista.es. Así funciona desokupa: Esto es lo que cobra el polémico negocio español para recuperar viviendas ocupadas, May 2023.

[55] Ediciones EL PAÍS. Resultados electorales en barcelona: Elecciones municipales, May 2023.

[56] Público. Resultados elecciones municipales valencia 2023, May 2023.

Appendix A

List of Parties and their Candidates

Madrid		Barcelona		Valencia	
Party	**Candidate**	**Party**	**Candidate**	**Party**	**Candidate**
Más Madrid	Rita Maestre	Esquerra Republicana de Cataluña (ERC)	Ernest Maragall	PP	María Jose Catalá
Partido Popular (PP)	José Luis Martínez-Almeida	Junts per Cat	Xavier Trias	Compromis	Joan Ribó
Ciudadanos	Begoña Villacís	PSOE	Jaume Collboni	PSOE	Sandra Gomez
Partido Socialista Obrero Español (PSOE)	María Reyes Maroto	Barcelona en Comú-En	Ada Colau	Vox	Juan Manuel Bádenas
Vox	Javier Ortega Smith	PP	Daniel Sirera	Ciudadanos	Fernando Giner
Unidas Podemos	Roberto Sotomayor	Valents	Eva Parera	Unides Podem	Pilar Lima
Recupera Madrid	Luis Cueto	Candidatura de Union Popular (CUP)	Basha Changue		
Tercera Edad en Acción	Guillermo Hernando	Vox	Gonzalo de Oro Pulido		
Por un Mundo Más Justo	Raquel Torrejón	Ciudadanos	Anna Grau		
Unión por Leganés	José García				
Progreso de Ciudades	Edgar Silva				
Falange Española de la JONS	Jesús Muñoz				
Partido Humanista	Arturo Viloria				
Partido Feminista de España	Lidia Falcón				
Partido Animalista Contra el Maltrato Animal (PACMA)	Asunción Estévez				
Partido Cannábico Luz Verde	Alberto Boira				
Escaños en Blanco	Aurora Rojas				
Partido Comunista de los Trabajadores	Javier Martín				
Partido Comunista de los Pueblos	Javier Martorell				
Madrid Capital	José Ángel Baeza				
Partido Castellano-Tierra Comuner	Julián Martínez				

Table A.1: Table showing the full list of parties and their candidates for each city used in this study.

Appendix B

Data Analysis on Barcelona

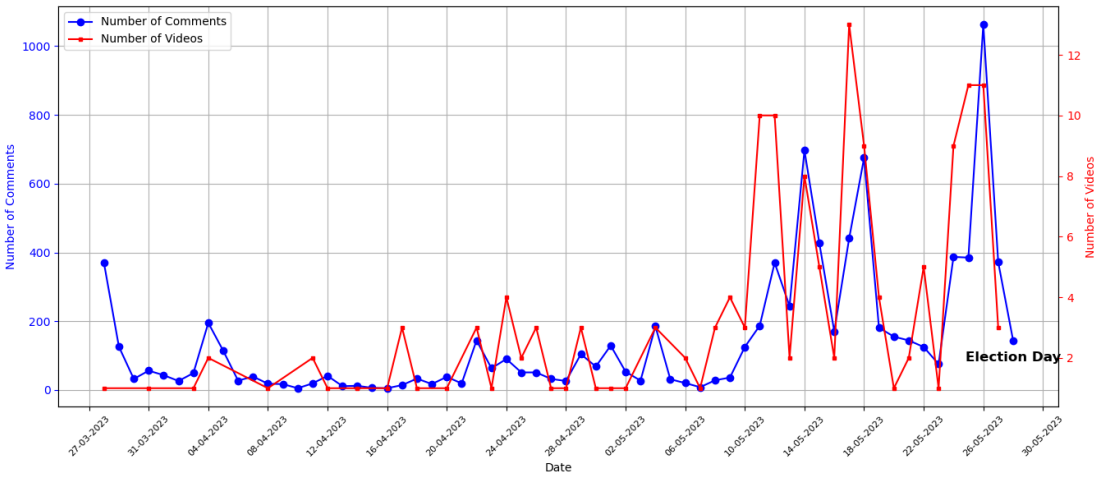

Figure B.1: Timeline of collected comments and videos for the Barcelona dataset

This timeline shows the same patterns as those seen in the main study for Madrid.

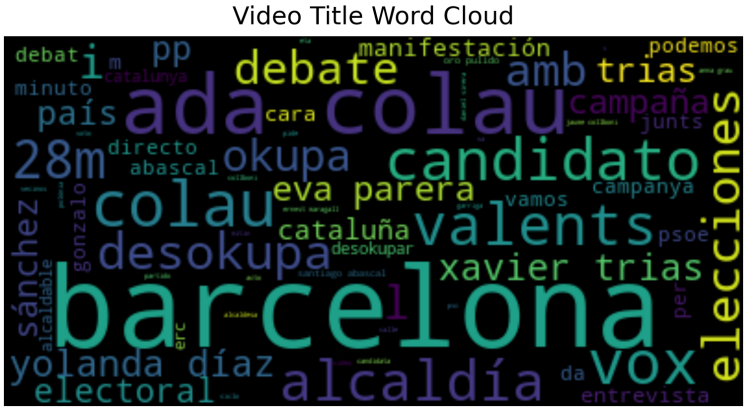

Figure B.2: Word cloud of the video titles for the Barcelona dataset

As seen in the study, the name of the city is highly prevalent in the video titles. This is caused by our own methodology for search and cleaning. For the same reason the name of candidates such as *Ada Colau* or *Xavier Trias* stand out in this word cloud. A term we did not see with such high frequency in the Madrid dataset is *"desokupa"*. This is the name of a business that reclaims occupied households. Occupation of households is a recurrent and polarising event that occurs in many parts of Spain but mostly in Barcelona [54]. Again, we see the relevance of these videos to the political situation of each city.

Figure B.3: Word cloud for the comments in the Barcelona dataset

Previously, we found how the frequency of certain parties or candidates did not match their performance in the elections. This word cloud indicates the same finding. The party *Vox* is the biggest party name in the comment word cloud, however, the results of the Barcelona municipality elections left this party as the sixth power [55].

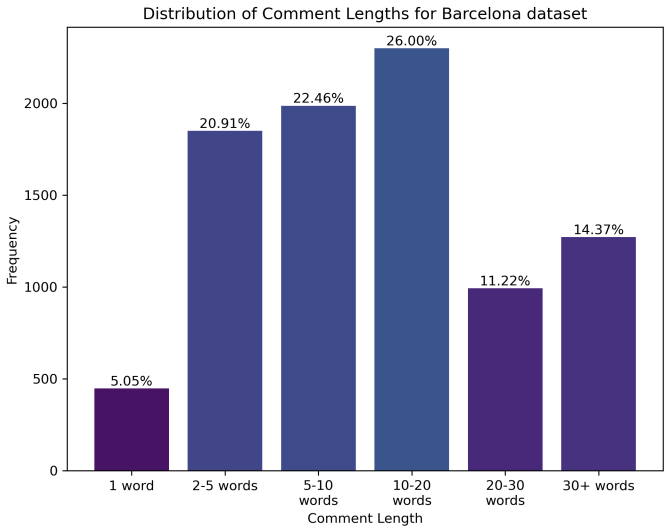

Figure B.4: Comment length bar graph for the Barcelona dataset

The comment length distribution for Barcelona is the same for Madrid and Valencia.

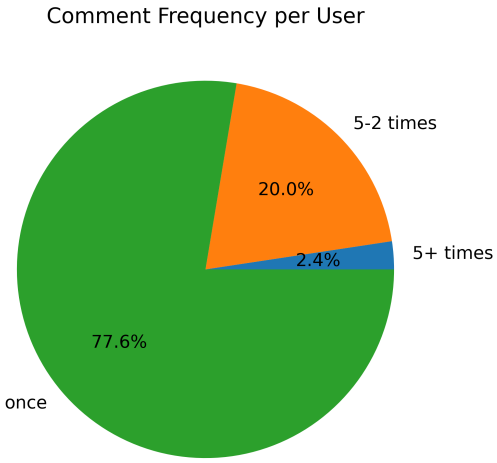

Figure B.5: Pie chart showing the comment frequency per user in the Barcelona dataset

The comment frequency per user is similar to those seen in other cities, the share of users that comment more than five times lies in the middle when compared to Madrid and Valencia

Appendix C

Data Analysis on Valencia

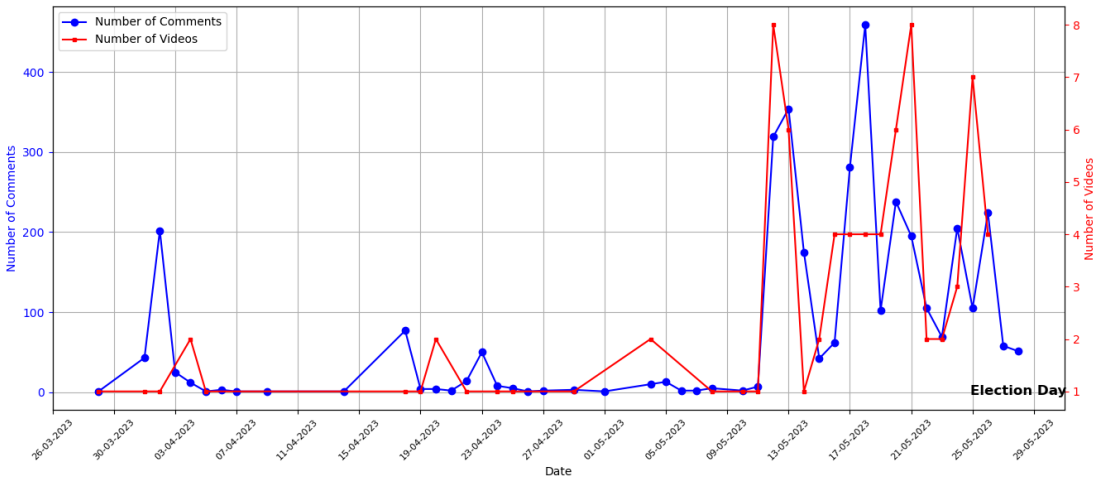

Figure C.1: Timeline of collected comments and videos for the Valencia dataset

This timeline shows the same patterns as those seen in the main study for Madrid.

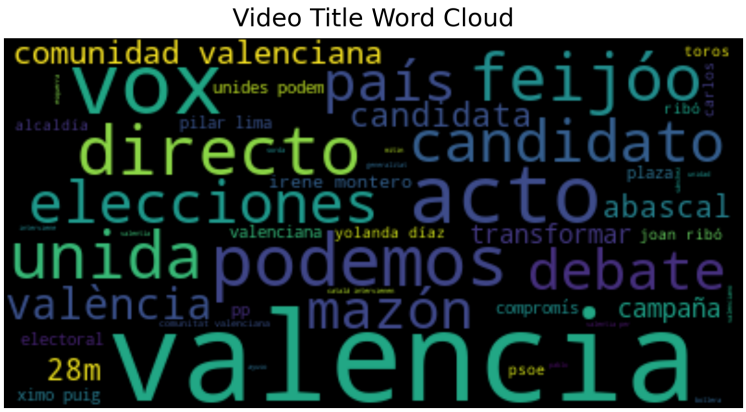

Figure C.2: Word cloud of the video titles in the Valencia dataset

This time, the video title word cloud does not show candidates for the city, the two names that appear are *Feijóo*, the president of *PP*, and *Mazón*, *PP's* candidate for the community of Valencia, not the city. This number of videos around the candidates for the municipality elections for this city has dropped considerably, as understood from the video title word cloud. This is likely to happen as we study cities with lower population. We do have parties mentioned in the titles, but without the name of candidates we cannot be sure these are related to the municipality instead of the community of the same name.

Figure C.3: Word cloud of the comments in the Valencia dataset

Again, the party *Vox* is the most frequent among all. This time in the municipality elections they had the fourth most support [56]. Once again, showing that frequency is not adequate for election prediction.

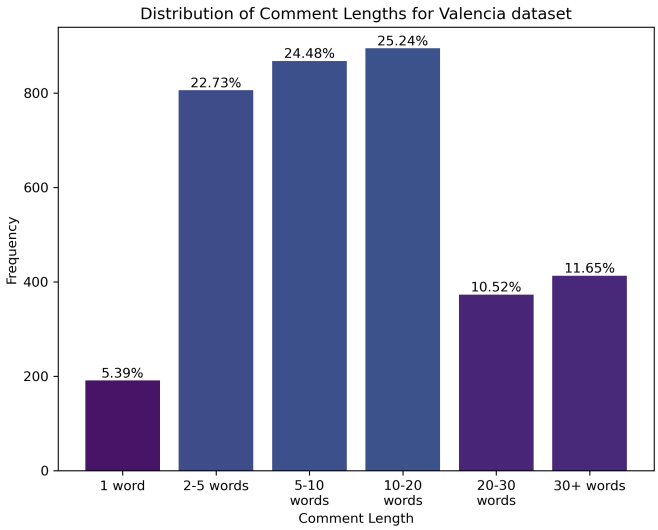

Figure C.4: Comment length bar graph for the Valencia dataset

The comment length distribution for Valencia is the same as the other cities studied.

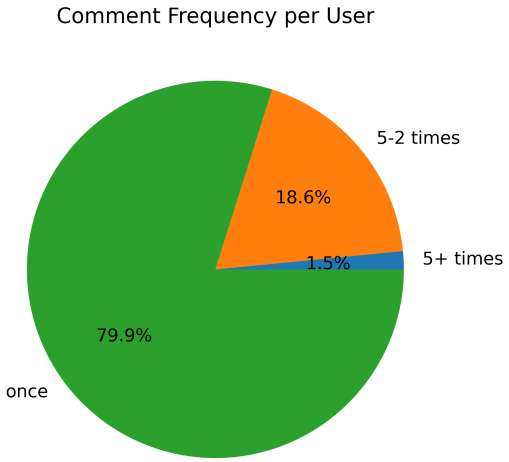

Figure C.5: Pie chart showing the comment frequency per user in the Valencia dataset

The comment frequency per user distribution for Valencia is the same as the other cities studied. The share of users that commented 5 times or more is slightly less than the other cities.

Appendix D

Language Analysis of YouTube data

In the analysis of language detection we saw two cases of confusion by the FastText [43] model, in which some Spanish comments were being classified as Portuguese and the Spanish laughing expression was being detected as Indonesian. We were also able to identify other peculiar patterns.

- **German** - The vomiting emoji was highly related to the German language, as most of the examples seen for this language were comments in which the only characters present was the vomiting emoji. This emoji, when present alone, was not detected as any other language. Perhaps the training the model underwent had many examples of this emoji for the German language and it associated the two.
- **Russian & Dutch** - The common laughing emoji was mostly associated with these two languages. It was interesting to see that when the frequency of this emoji surpassed the mark of three, it went from being identified as Dutch to being classified as Russian.
- **Others** - There are many other examples, but of similar nature to those seen in the main study, where a word that is the same for different languages confuses the model into classifying that comment incorrectly.

Appendix E

Political Ideology Prediction model - Training Data analysis

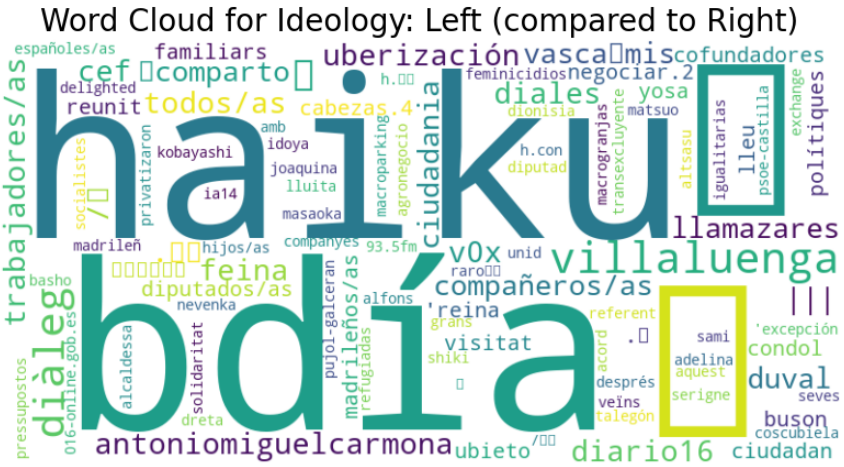

Figure E.1: Word cloud showing the distinctive words the left uses that the right does not

The main words in the above word cloud are *"haiku"* and *"bdía"*, these are explained in the main text, section 4.4.2.1.

Figure E.2: Word cloud showing the distinctive words the right uses that the left does not

The most frequent words in this word cloud are commented in the main text, section 4.4.2.1.