Data in Brief 58 (2025) 111280

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Dataset of keywords used by European political parties on Facebook

Francisco Caravaca [a,1,*], Ángel Cuevas [a,b,2], Rubén Cuevas [a,b,3]

[a] Department of Telematic Engineering, Universidad Carlos III de Madrid, Avenida de la Universidad 30, 28911 Leganés, Spain
[b] UC3M-Santander Big Data Institute, Calle Madrid 135, 28903 Getafe, Spain

ARTICLE INFO

Article history:
Received 9 September 2024
Revised 18 December 2024
Accepted 3 January 2025
Available online 10 January 2025

Dataset link: Keywords used by European Political Parties on Facebook (Original data)

Keywords:
Politics
Europe
Terms
Social media

ABSTRACT

This dataset contains the frequency of thousands of terms (or keywords) used by political parties in the posts they have published in their Facebook pages. The data set is composed by 20,317 keywords from posts published by 279 European political parties from 28 countries and spans 5 years, from January 2019 to December 2023. Due to the large diversity of languages in the analysed countries, we have translated every post into English to compile this dataset. We also provide an open-access web portal: EU Political Barometer, in which a wide variety of analyses can be carried out without the need of working directly with the dataset. This allows scientists without a data analysis background to access the information embedded within the dataset. The information included in the dataset may be of value for social scientists that wants to understand the evolution of the topics employed by political parties in Europe based on a widely used political communication tool such as Facebook.

* Corresponding author.
E-mail addresses: fcaravac@it.uc3m.es (F. Caravaca), acrumin@it.uc3m.es (Á. Cuevas), rcuevas@it.uc3m.es (R. Cuevas).
Social media: @fcaravacacrespo (F. Caravaca)
[1] ORCID: https://orcid.org/0000-0002-4696-5276.
[2] ORCID: https://orcid.org/0000-0002-5738-0820.
[3] ORCID: https://orcid.org/0000-0002-1440-8360.

https://doi.org/10.1016/j.dib.2025.111280

Specifications Table

Subject	Social Sciences / Political Science
Specific subject area	Political Parties on Social Media
Type of data	Tables.
Data collection	We have collected the Facebook posts from 279 political parties from the European political scene, using an automatic procedure that browses each of the political parties' Facebook pages. Then, we have translated the texts into English, to then be able to compute the presented keywords dataset. Those political parties Facebook pages were manually identified using the results of recent elections.
Data source location	Facebook.
Data accessibility	Repository name: Repositorio de Datos del Consorcio Madroño Data identification number: doi: 10.21950/ENBQFP Direct URL to data: 10.21950/ENBQFP GitHub Repository with the supporting code is available at: https://github.com/fcaravaca/FB-Keywords
Related research article	

1. Value of the Data

- These data are useful to understand the variations of the political discourse over time on Facebook of European political parties.
- This dataset covers a rather large geographical and temporal coverage. In particular, this dataset covers 28 different countries, which represents an added value that will allow scientists running cross-country analyses using a common and coherent source of information. On the other hand, our dataset expands over a considerable period of 5 years.
- The data can also be used to compare with other sources of information such as political manifestos.
- The data can also be easily visualized on our portal EU Political Barometer; therefore, any interested individual can consult the dataset in a simple way.

2. Background

Computational methods have played a pivotal role in enhancing the understanding of various topics within political science [1]. Notably, election prediction has been a focal point, employing diverse techniques such as sentiment analysis [2], as well as tracking the frequency of mentions of politicians [3].

One of the pioneering studies leveraging textual data was conducted by Laver et al. [4], wherein they derived policy positions from political texts using wordscores techniques.

The investigation around keywords in the political landscape has garnered significant attention on social media platforms, particularly on Twitter. In [5], researchers matched keywords to topics demonstrating that their method effectively mirrors consumer confidence and presidential job approval polls. Meanwhile, in [6], multiple sets of keywords were employed to develop distantly supervised classifiers for topic identification. This method was subsequently applied by Sánchez-Junquera et al. [7] to analyze the most relevant topics discussed during the Spanish November 2019 elections. The examination of keywords proves invaluable in distilling vast volumes of data, as demonstrated by Campos et al. [8]. Therefore, proving a dataset of keywords from a wide range of parties in a very long-time span, is a useful addition that could facilitate researchers to tackle diverse research questions.

3. Data Description

The dataset includes four CSV files, the main one referred to as *keywords.csv* file, and three supplementary files named *post_activity.csv, facebook_parties.csv* and *missing_parties.csv* (final name in the dataset will depend on the dataset version). The first supplementary file includes the posting activity of each party, whereas the second includes the information for each party in the dataset including the FB page URL. All the files are connected through a party ID that allows identifying a specific party.

First, we detail the structure of the data files of the dataset, then we provide a country level dataset description, highlighting some metrics. Finally, we show our portal EU Political Barometer, which is an open-access web portal that can be used to explore the dataset.

3.1. Main data file

The *keywords.csv* file contains a total of 15,927,046 records. Each record describes a keyword or a combination of keywords (as bigrams) used by a political party in a specific week. The presence of a bigram does not exclude the individual appearance of its constituent terms. Having bigrams is pivotal as individual words may not always provide sufficient context. For instance, the term *climate change* imparts more specific information than *climate* or *change* in isolation.

Additionally, each record displays the number of posts that include the keyword for a particular party in a given week. In total, we identified 20,317 unique keywords. Table 1 provides a simplified example of the file's structure, showcasing three keywords for a specific week.

Table 1
The *keywords.csv* file structure.

party_id	word	week	freq
902	change	2022–02–06	4
902	food	2022–02–06	4
902	challenge	2022–02–06	1

3.2. Supplementary Files

We include the following supplementary files: *post_activity.csv* and *facebook_parties.csv*. The first one includes the activity of political parties on Facebook i.e., the number of posts published on each of the parties' pages per week. An example is depicted in Table 2. The other file (*facebook_parties.csv*), includes detailed information for each of the parties in our dataset (see Table 3).

Also, it is worth noting we have added an extra file to the repository that includes those parties that received more than 5 % of vote share in the 2019 EU Parliament election or the last parliament election (this will be further explained in the Party Selection subsection) recorded in ParlGov: *missing_parties.csv* and are not included in our dataset because we could not find any reliable Facebook page for them.

Table 2
The *post_activity.csv* file structure.

party_id	Week	postsWithText	totalPosts
902	2022-02-06	32	38
1597	2022-02-06	11	11

Table 3
The *facebook_parties.csv* file structure.

party_id	left_right	Country_name	family_name	party_name	party_name _short	facebook_page
2	7.4	Czech Republic	Conservative	Tradice Odpovědnost	TOP09	top09cz
21	2.5	Lithuania	Green/Ecologist	Lietuvos žalioji	LZP	Lietuvoszaliujupartija

3.3. Country level dataset description

This dataset provides useful content to analyze the use of Facebook made by Political Parties from the European Union landscape and the United Kingdom. As explained before, our dataset aims to include a comprehensive representation of the relevant parties within each country, which by default implies an important heterogeneity across countries. There are countries with more parties than others, there are countries with more active parties than others, etc. Also, within the same countries there is a high heterogeneity across the activity of the parties.

In Table 4 we show different metrics for the keyword usage and party activity for each of the different countries.

Table 4
Summary of the dataset by country, for both keyword usage (*entries* each entry is a record for a particular word used by a party in a given week; $entries_p$ is the median records amount by party in a country; U_k represents the unique keywords used in a country; top_{100} and top_{1000} indicate the percentage of dataset frequency that represents the most frequently used keywords across the entire dataset.), and post activity ($posts_t$ are the posts with text, and $posts_p$ the median amount of posts made by each party).

Country	Keyword Data					Post Activity		
	entries	$entries_p$	U_k	top_{100}	top_{1000}	$posts_t$	posts	$posts_p$
Austria	361,747	61,754.0	13,960	21.73	57.65	19,773	23,479	4369.0
Belgium	647,097	47,367.0	15,282	21.36	57.16	32,088	33,606	2493.0
Bulgaria	663,516	36,656.0	15,325	19.49	57.00	34,848	45,538	3353.0
Croatia	698,842	42,821.5	15,701	19.63	56.42	27,249	29,114	2015.5
Cyprus	462,680	59,277.0	14,073	17.39	54.63	28,821	32,965	3575.5
Czech R.	998,853	114,409.0	16,032	20.63	58.17	46,525	48,971	5371.0
Denmark	510,809	30,991.0	14,164	24.00	58.85	21,444	23,874	1385.0
Estonia	466,559	80,364.5	14,840	20.57	59.28	20,596	21,903	3604.5
Finland	552,790	70,243.0	14,695	21.43	59.44	26,197	27,934	3002.0
France	357,242	34,719.0	13,895	19.24	54.64	30,346	34,770	3291.5
Germany	730,396	44,845.0	15,543	20.64	57.26	30,222	36,617	2411.0
Greece	364,096	36,233.0	14,014	17.90	54.22	24,061	39,607	2398.0
Hungary	749,007	112,030.5	15,548	21.02	58.43	49,490	52,009	6819.0
Ireland	317,674	37,428.0	12,658	21.11	55.80	20,260	22,691	2152.0
Italy	850,930	71,214.0	16,242	17.36	56.04	81,461	86,755	4962.0
Latvia	655,849	45,209.0	15,619	18.95	58.42	25,902	29,936	1897.0
Lithuania	426,321	44,787.5	14,514	19.56	57.53	12,418	14,702	1571.0
Luxembourg	216,399	32,869.0	13,035	21.36	58.67	10,661	11,847	1649.0
Malta	108,946	24,679.0	10,323	20.16	54.15	9365	11,692	3157.5
Netherlands	295,480	19,098.0	12,955	22.63	57.88	19,010	20,474	1463.0
Poland	671,065	69,633.0	15,465	20.56	57.02	38,345	42,221	4856.0
Portugal	819,485	61,773.0	16,131	18.93	56.88	50,791	54,386	2985.0
Romania	748,074	70,218.0	15,903	20.21	56.71	32,449	34,793	3424.0
Slovakia	771,747	61,248.0	15,836	20.39	56.91	33,004	39,712	2937.0
Slovenia	488,829	43,317.0	15,190	20.18	58.22	23,051	25,239	2472.0
Spain	894,641	55,014.0	15,553	19.32	53.29	74,478	87,799	5262.5
Sweden	584,892	56,889.5	14,486	21.73	59.58	25,362	28,830	2197.5
United Kingdom	513,080	43,315.0	13,327	21.15	55.77	37,104	41,033	3646.0

Regarding the keyword data usage, we first include the total number of entries, these are the number of records for each country in the *keywords.csv* file, i.e., a record is the usage of a word (at least once) in a given week by a party. As it is expected, we found that there is more data in some countries than others, being for instance the Czech Republic, the country with the greatest number of records (998k), and Malta (as was expected as being the country with the least parties) the one with the least number of rows in the dataset (109k). We also include the median number of entries by party ($entries_p$), which shows a more representative metric to see the differences between countries.

Additionally, there is diversity in terms of keyword usage across different countries, as we account for the number of unique words found in each country, denoted as U_k. These unique keywords also encompass bigrams. We identified a median value of 15,015 unique keywords per country, and most countries closely align with this value.

We have also calculated the list of the most frequently used keywords in the dataset at the European level, along with their frequencies. For example, the four most common keywords are: *will, government, people*, and *can*, which are extensively utilized across the entire European Union. Using this list, we determined the percentage that these top keywords (the 100 or 1000 most common) represent in each country, denoted as top_{100} and top_{1000}.

As mentioned earlier, we have included a post activity analysis in Table 4. This analysis provides information on the number of posts containing text in each country, denoted as $posts_t$, the total number of posts, and the median number of posts per party, denoted as $posts_p$. This is relevant as the differences in posting activity influence the data available for each country.

Another aspect which is necessary to explore is whether the activity of political parties have varied over time, in particular the most relevant issue could be that there is a generalized decrease on activity of several countries, e.g., that could happen if the parties are less incentive to participate on Facebook, due to being a less popular social network on a particular country. Therefore, we want to check whether there is a relevant decrease in participation by political parties in the presented dataset. To analyze this, we present in Fig. 1a boxplot that depicts the weekly activity over the different countries separated by years.

The parties' activity has remained very regular in many countries during the analyzed period (e.g., Estonia, Netherlands, Spain, Estonia, Lithuania, Luxembourg, United Kingdom). There are also countries where the increase in published posts can be easily seen, such as Bulgaria and the Czech Republic, the latter going from 100 weekly posts to 250 posts from 2019 to 2023 (an increase of +150 % in post activity).

On the other hand, there are countries which have their minimum number of posts in the last year of the dataset (2023), such as Sweden or Denmark. However, these decreases in those countries are insufficient to state that parties are less willing to participate in this social media. Furthermore, we do not find a decrease similar to the post-activity increase in the Czech Republic.

In summary, our dataset provides a detailed overview of the linguistic and posting activity of political parties on Facebook across the different European countries.

3.4. EU political barometer

In addition to collecting the dataset, we have developed an interactive web tool: the *EU Political Barometer*. This website is focused on displaying information related to political parties on Facebook, and it is publicly available at https://eupoliticalbarometer.uc3m.es. The data shown on the website, as the dataset presented here, contains the data since January 1st, 2019. The current version of the dataset includes data until December 31st, 2023, whereas the information on the website is regularly updated.

The site has been designed to facilitate access to the data to other researchers or any interested stakeholder. The tool provides a wide range of visualizations and features to allow comparisons between parties or countries. Finally, we have tried to develop a very friendly user

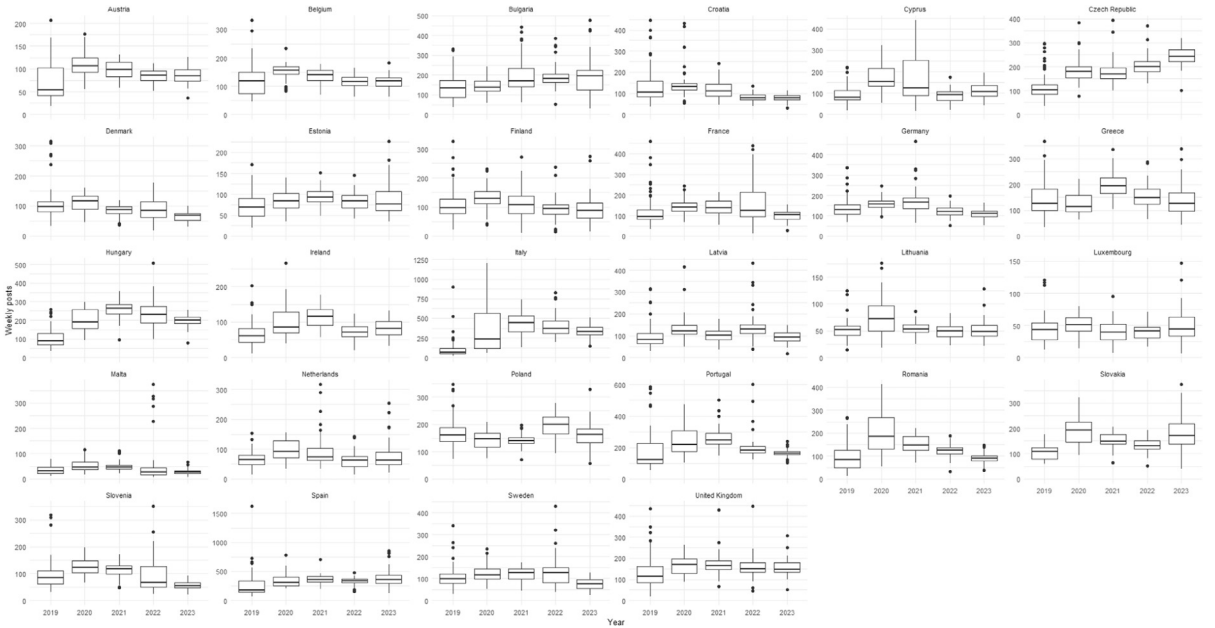

Fig. 1. Weekly posting activity of each country across years. Boxplot that provides the distribution of the post activity of each country each year during the five-year period of the dataset.

interface, and the website does not require the user to have previous knowledge on data management to use it.

The portal is divided into different dashboards each of them addressing a different feature. The dashboard including the information of this dataset is referred to as *Trends*. Each of the dashboards includes different charts and maps, to observe the evolution of certain aspects, as well as the differences between parties or countries on those elements. Next, we explain the main functionalities of the *Trends* Dashboard.

3.5. Trends dashboard

This dashboard allows anyone to analyze and visualize in an easy manner the dataset introduced in this paper. Particularly, the dashboard is accessible through the main page of the site, or directly at https://eupoliticalbarometer.uc3m.es/dashboard/trends.

The main functionality of this dashboard is inspired by Google Trends, even though the information displayed is different in nature from Google Trends. The user just needs to select or search for a keyword, and the dashboard displays the information about the evolution of the usage of this term across EU and UK political parties. For instance, someone could select the keyword *coronavirus*, and a general chart about the aggregated use of this term will be displayed (see Fig. 2). It is also possible to observe this data in a cumulative way. A *Relative usage* chart is also displayed, which shows the percentage of posts that use a certain keyword over the total number of posts in a specific time window.

Furthermore, for the selected keyword, if the user scrolls down in the web portal, the dashboard displays the differences between different countries or ideologies, over three groups: left, center, and right parties. For instance, Fig. 3 depicts the usage of the term *coronavirus* over the different countries and political orientations. This dashboard also allows the possibility of selecting a specific time window and modifying the ideological groups in the 10-point ideological scale. It is also possible to filter out countries to only focus on the ones of interest to the user.

The tool also allows the possibility of selecting several terms at the same time, with all the previous features. This will facilitate seeing the differences in the aggregated use of those keywords over time as well as their use in different countries. Additionally, it is possible to select specific parties or countries, to see how they make use of any kind of keyword, as is displayed in Fig. 4, in which we selected three parties Vox (Spain), Rassemblement National (France) and Alternative für Deutschland (Germany) using the keywords *climate change* and *immigration*.

4. Experimental Design, Materials and Methods

In this section, we present the methods used to create the presented dataset, that includes the keywords extracted from more than one million posts (885,000 containing text) publicly posted by political parties within 28 different countries spanning 5 years, between January 1, 2019, and December 31, 2023. We first describe the political party selection. Second, we describe the methodology used to retrieve the raw text data from the FB pages of political parties, i.e., FB posts. Third, we explain the methodology used to translate all the posts into English. Finally, we explain how the dataset has been generated. In Fig. 5 we include the flowchart of the dataset creation, including party selection, data collection and dataset generation.

4.1. Party selection

As a first step, it is important to have a comprehensive list of the political parties that covers the majority of the vote share in elections of any of the analyzed countries. The goal is to select the main Facebook page of each political party. This process required manual verification, as parties may have different Facebook pages (e.g., this could include pages for a region inside the

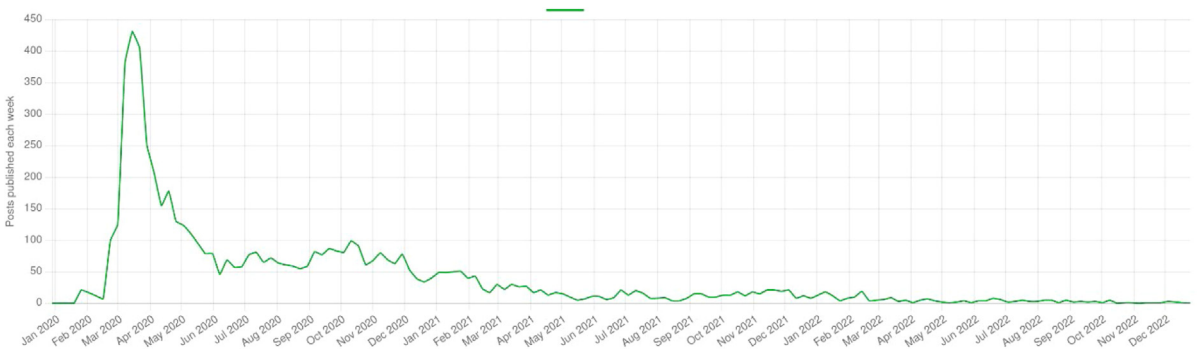

Fig. 2. Web portal screenshot about the use of *coronavirus* keyword. The total number of posts containing the term *coronavirus* from January 2020 to December 2022, broken down by each week. For instance, we observe a peak around March 2020 when the pandemic irrupted.

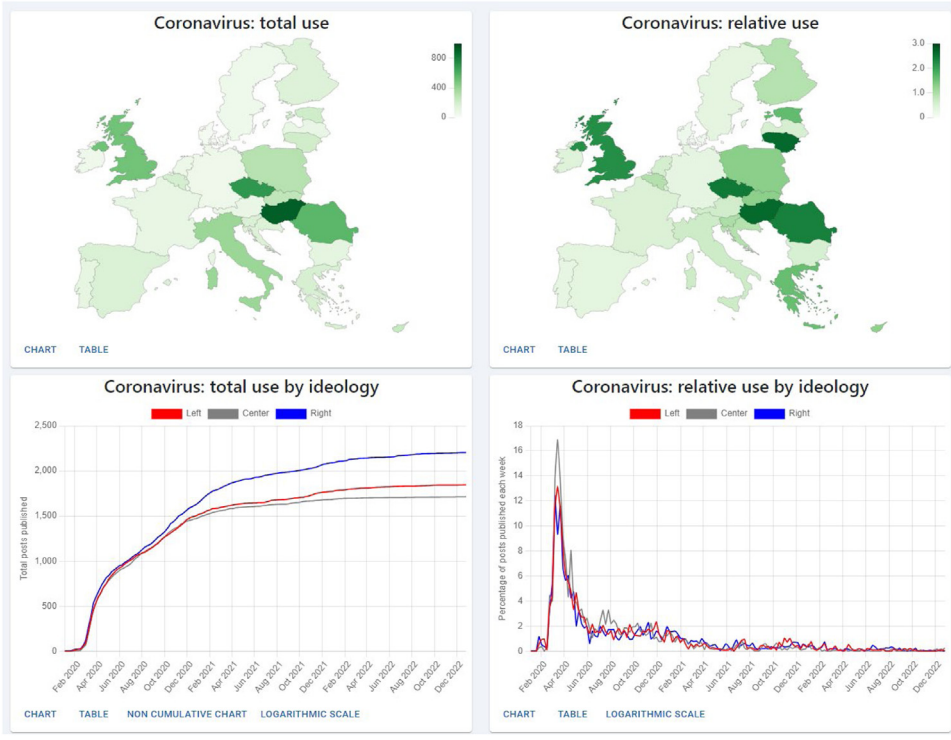

Fig. 3. Web portal screenshot with more detailed information about the *coronavirus* keyword. Detailed information about the usage of the term *coronavirus* is provided. This includes two maps: the first displays the total frequency of the keyword, while the second illustrates its frequency relative to the overall number of posts. Additionally, the information encompasses usage categorized by ideology. In the bottom left corner, there is a chart depicting the cumulative total of posts categorized by ideology. Lastly, another chart displays the percentage of posts in each week categorized by ideology.

country). It is also important to specify that some parties do not have a Facebook page, or they may have an inactive one. We skipped these political parties.

To compile the list of Political Parties, we utilized ParlGov [9], a comprehensive database encompassing information from the majority of OECD countries regarding political parties and elections, which encompasses all the European Union member states. This resource provides details such as party names and abbreviations for political parties, as well as election-related information, including results from both national parliamentary elections and European Union elections, specifying the vote share and seats held by each party. This dataset plays a crucial role in enabling us to identify the relevant parties within each country.

Although there are other alternatives such as Global Party Survey 2019 [10] that provide information on political parties in other countries, we selected the ParlGov dataset, for the following reasons: (i) this database includes information from a wide variety of political parties; (ii) it provides specific information from each political party, such as their political family and their ideology; (iii) it also includes information from elections. Therefore, ParlGov aligns with our objectives in comparison with the other options.

As our objective is to identify relevant parties that collectively hold the majority of the vote share in each country, we initially prioritize parties that obtained at least 5 % of the votes in either the 2019 European Elections or the most recent parliamentary elections documented in ParlGov for each respective country. By applying this threshold, we compile a list of parties,

Fig. 4. Example of a simple comparison of the use of keywords by different parties. Total number of posts using the keywords *climate change* and *immigration* by the parties Vox (Spain), Rassemblement National - RN (France) and Alternative für Deutschland - AfD (Germany).

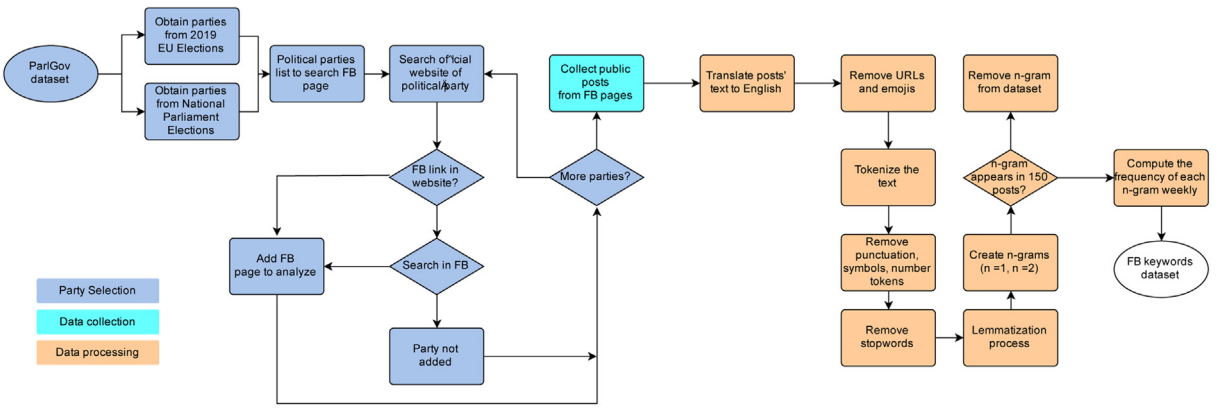

Fig. 5. Flowchart of the generation process of the dataset, including party selection, data collection and processing.

from which we try to retrieve their Facebook pages. Specifically, for the 2019 EU Elections, we collect information from 156 out of 165 entries (independent candidates were not considered), and for the parliamentary elections, we obtained data from 149 out of 152 parties. It is worth noting that these lists largely overlap, with a significant portion of parties being present in both categories.

Our approach to identifying all Facebook party pages followed this procedure: after obtaining information from ParlGov about a particular party, we initially attempted to locate their official website, as a significant portion of parties provide a link to their primary Facebook page there. This method proved successful for most parties. However, in certain instances, we encountered difficulty in finding the parties' websites. In such cases, we conducted a direct search using Facebook's search bar, ensuring that the page belonged to the national organization.

While we believe that the selected parties already provide extensive coverage of political entities, we endeavored to collect information from additional parties that participated in the elections. However, it is important to note that this process comes with certain challenges and limitations: (i) smaller parties may be less likely to have a presence on Facebook; (ii) finding their Facebook pages may not be straightforward for various reasons, such as a lack of a dedicated website or discrepancies in their Facebook page names compared to the party's name on ParlGov; (iii) smaller parties may create a Facebook page but not actively use it; (iv) some parties may utilize Facebook Groups instead of pages, and it may be difficult to verify if these were created by the political organization. While we acknowledge the possibility of smaller parties being absent from the dataset due to these challenges, our aim is to construct a dataset that encompasses as many political parties as feasible. Therefore, we have made diligent efforts to include as many parties as possible, but we can only guarantee the presence of those with a vote share above 5 %.

The final 279 parties forming our dataset accounted for the 90.50 % of vote share on average in the 2019 EU Elections and 91.44 % for the last parliament elections across the 28 countries under consideration. In terms of seat share in each of the elections, we obtained a bit better representation, with more than 95 % coverage.

In summary, although we may be missing some parties, our dataset is very comprehensive and offers a very large coverage in terms of vote share and seats. File *missing_parties.csv* includes the list of parties not included in our dataset that had more than 5% vote share in either election. Table 5 breakdowns the coverage information of our dataset per country showing: (i) the number of parties included in our dataset; (ii) the coverage in vote share and seats according to the 2019 EU elections; (iii) the coverage in vote share and seats according to the last national election.

4.2. Data collection

We gather information pertaining to the posts on the main Facebook pages of the 279 political parties included in our dataset. These posts are accessible to the public without the need for a Facebook account. For each post, we gather the content (including text and URLs) and record the level of engagement it has received, which encompasses the number of likes, shares, comments, and video views (if applicable). It is important to emphasize that this information does not encompass any personal data and thus is not governed by data protection regulations such as the GDPR.

Due to the large number of political parties in our dataset, we built an automatic procedure to browse through each of the political parties' pages and scrape the publicly available information related to the posts. We browse through the posts published back to January 1st, 2019, when we stopped collecting data. We have collected more than 1,000,000 posts, from which 88.3 % include text. This collected data was already used for the results of one of our previous works [11]. However, in that previous article we did not analyze the content of the post (i.e., the text of the posts) but the post activity and engagement. Therefore, the dataset presented in this article is intrinsically different.

Table 5
Parties Representativeness of the dataset in the 2019 EU Elections and the last Parliament Elections available in ParlGov. *EU* stands for the EU elections and *P* for the parliament elections. The table reports the vote share V and the percentage of the seats represented *S*.

Country	Parties	EU V.(%)	EU S.(%)	P. V.(%)	P. S.(%)
Austria	5	98.20	100.00	96.90	100.00
Belgium	13	97.85	100.00	95.92	100.00
Bulgaria	11	92.18	100.00	86.21	95.00
Croatia	12	71.74	81.82	94.73	96.03
Cyprus	8	87.98	100.00	80.94	92.86
Czech Republic	9	88.59	100.00	88.37	100.00
Denmark	17	100.00	100.00	98.40	97.77
Estonia	6	88.70	100.00	92.80	100.00
Finland	9	96.80	100.00	95.07	99.00
France	10	89.65	100.00	81.38	91.68
Germany	13	95.50	98.96	96.12	99.86
Greece	7	73.21	80.95	87.86	95.00
Hungary	8	97.40	100.00	99.02	98.99
Ireland	9	76.07	72.73	86.51	87.50
Italy	9	94.67	100.00	88.42	94.00
Latvia	13	78.87	75.00	77.63	100.00
Lithuania	10	76.25	81.82	87.02	94.33
Luxembourg	7	96.21	100.00	92.24	100.00
Malta	4	94.22	100.00	99.55	100.00
Netherlands	13	97.54	100.00	91.40	94.67
Poland	9	95.70	100.00	98.91	90.87
Portugal	11	95.14	100.00	92.45	96.52
Romania	8	93.43	100.00	93.19	94.55
Slovakia	11	82.37	84.62	91.39	100.00
Slovenia	10	93.22	100.00	83.76	97.78
Spain	16	91.18	96.30	92.39	96.57
Sweden	10	98.27	95.00	93.85	95.42
United Kingdom	11	93.10	98.63	97.96	98.92
Total	279	90.50	95.21	91.44	96.69

4.3. Data translation

One challenge that may arise when analyzing text data across different countries is the presence of multiple languages. The European Union, for instance, acknowledges 24 official languages [12], contributing to a rich linguistic diversity. Moreover, some political parties on Facebook may employ additional official or co-official languages in certain countries, such as Luxembourgish or Catalan. This extensive variety implies that effectively analyzing the data in each of these languages would necessitate a deep understanding of each, rendering the task impractical for us.

Hence, we opted to translate the text of the posts into English, considering it to be the most accessible choice given its widespread usage. Moreover, as the performance of the translators is related to the training dataset size [13], this usually means that translations into English are generally better than to any other language. Finally, libraries usually have better support for this language.

Because of all of the above, the text of each of the posts is translated by using online translators into English, in particular, we used the following providers: Google Translator, DeepL, Microsoft Translator Text API, and Libre Translate.

While automatic translations are not flawless and may contain errors due to the absence of contextual understanding, studies have shown that the use of automatically translated texts is a viable option in bag-of-words models and LDA topic models [14].

4.4. Dataset generation

To generate the dataset contributed to this work we have used the *quanteda* R package [15] to process the posts' text. The data preprocessing pipeline involves several key steps. First, we implement a basic data cleaning process, which encompasses the removal of URLs and emojis from the text. Second, we proceed to generate tokens from the corpus. Within this tokenization process, tokens corresponding to punctuation marks, symbols, and numerical values are eliminated to streamline the dataset.

Third, we undertake the removal of stopwords using the dedicated function from the *tm* R Package [16]. This function encompasses a curated list of 174 common English words that typically impart limited informational value. Then, the tokens undergo lemmatization, a process that standardizes words to their base form. For this task, we employ the *lexicon* R package [17], leveraging Měchura's extensive English lemmatization list [18] which comprises over 41,000 transformation pairs.

Once we have applied the lemmatization process, we aggregate terms into groups consisting of one or two words, denoted as n-grams with n values of 1 or 2. Specifically, bigrams, which are combinations of two adjacent words, are selected. After that, we only retain the terms that appear consistently across a series of posts, with a minimum threshold of 150 occurrences.

Finally, from each political party, we extract the frequency of appearance for each keyword within weekly intervals. This step enables a nuanced analysis of the prominence and recurrence of specific terms over time.

The resulting dataset offers comprehensive insights into the topics that political parties communicate through their Facebook posts. This enables an understanding of the most frequently disseminated subjects by a specific political party within a defined time frame. The high granularity, on a weekly basis, allows the observation of variations in the intensity of a certain keyword.

Limitations

We acknowledge that there are some parties that are not represented in the dataset (due to not having a Facebook account or we not been able to find it), as it is mentioned in the Party Selection section. We further include a file in the dataset missing_parties.csv that shows which relevant parties are missing from the dataset. We also acknowledge that it is relevant to report that post activity is not heterogeneous across parties, countries nor time, which will affect the amount of data available. It is also important to note that there is the possibility that some posts were deleted by parties before we were able to collect them. Related to the previous issue, there is also the possibility that some parties have closed their accounts before we could collect data from them.

Ethics Statement

Considering this work, we have obtained the approval of the Ethics Committee of our institution. Relevant points to consider:

1. **Terms of Service (ToS):** The Facebook posts used as the basis for our research are publicly available information that anyone can access even without a Facebook account. Therefore, to the best of our knowledge, this is public information open to anyone. Roughly, speaking a user does not have to subscribe to the Terms of Use of Facebook to access that information. That happens when someone opens an account. We use that public information to create this dataset, therefore what we are making public is the information we generate.
2. **Copyright:** The resulting dataset that we are presenting has been designed and created by us, which is based on the publicly available information discussed in the previous point.

3. **Privacy:** The data is not related to individual users; therefore, anonymization is not relevant to this dataset.
4. **Scraping policies:** To the best of our knowledge there is not specific scraping policies.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Grammarly in order to improve the quality of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Credit Author Statement

Francisco Caravaca: Software, Investigation, Data curation, Writing – Original Draft, Visualization. **Ángel Cuevas**: Conceptualization, Methodology, Validation, Writing – Review & Editing, Supervision. **Rubén Cuevas:** Conceptualization, Writing – Review & Editing.

Data Availability

Keywords used by European Political Parties on Facebook (Original data) (Repositorio de Datos del Consorcio Madroño).

Acknowledgments

This research has received funding from: the project ENTRUDIT (Grant TED2021-130118B-I00) funded by the MCIN/AEI/10.13039/501100011033 and the NextGeneration EU/PRTR funds; the Ministerio de Asuntos Económicos y Transformacion Digital and the European Union-NextGenerationEU through the project PRITICACLOUD and the AImpulsa chair (Grant TSI-100922-2023-1).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. Glavaš, F. Nanni, S.P. Ponzetto, Computational analysis of political texts: bridging research efforts across communities, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Association for Computational Linguistics, Florence, Italy, 2019, pp. 18–23, doi:10.18653/v1/P19-4004.
[2] P. Burnap, R. Gibson, L. Sloan, R. Southern, M. Williams, 140 characters to victory?: using Twitter to predict the UK 2015 general election, Elect. Stud. 41 (2016) 230–233, doi:10.1016/j.electstud.2015.11.017.
[3] J. DiGrazia, K. McKelvey, J. Bollen, F. Rojas, More tweets, more votes: social media as a quantitative indicator of political behavior, PLOS ONE 8 (11) (2013) e79449, doi:10.1371/journal.pone.0079449.
[4] M. Laver, K. Benoit, J. Garry, Extracting policy positions from political texts using words as data, Am. Polit. Sci. Rev. 97 (2) (2003) 311–331, doi:10.1017/S0003055403000698.
[5] B. O'Connor, R. Balasubramanyan, B. Routledge, N. Smith, From tweets to polls: linking text sentiment to public opinion time series, in: Proceedings of the International AAAI Conference on Web and Social Media, 4, May 2010, pp. 122–129, doi:10.1609/icwsm.v4i1.14031.

[6] M. Marchetti-Bowick, N. Chambers, Learning for microblogs with distant supervision: political forecasting with Twitter, in: W. Daelemans (Ed.), Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 603–612. https://aclanthology.org/E12-1062. Accessed: 22 August, 2024. [Online]. Available:.
[7] J. Sánchez-Junquera, S.P. Ponzetto, P. Rosso, A Twitter Political Corpus of the 2019 10N Spanish Election, in: P. Sojka, I. Kopeček, K. Pala, A. Horák (Eds.), Text, Speech, and Dialogue, Springer International Publishing, Cham, 2020, pp. 41–49, doi:10.1007/978-3-030-58323-1_4. vol. 12284 in Lecture Notes in Computer Sciencevol. 12284.
[8] R. Campos, A. Jatowt, A. Jorge, et al., Text mining and visualization of political party programs using keyword extraction methods: the case of Portuguese legislative elections, in: I. Inclusivity, A. Sserwanga, H. Goulding, J.T. Moulaison-Sandy, A.L. Du, V.H. Soares, et al. (Eds.), in Information for a Better World: Normality, Virtuality, Physicality, Springer Nature Switzerland, Cham, 2023, pp. 340–349, doi:10.1007/978-3-031-28035-1_24. Lecture Notes in Computer Science.
[9] H. Döring, C. Huber, and P. Manow, "ParlGov 2022 Release." Harvard Dataverse, 2022. doi: 10.7910/DVN/UKILBE.
[10] P. Norris, "Global Party Survey, 2019." Harvard Dataverse, 2020. doi: 10.7910/DVN/WMGTNS.
[11] F. Caravaca, J. González-Cabañas, Á. Cuevas, R. Cuevas, Estimating ideology and polarization in European countries using Facebook data, EPJ Data Sci. 11 (1) (2022) 56, doi:10.1140/epjds/s13688-022-00367-1.
[12] European Union, "Languages, multilingualism, language rules | European Union." Accessed: 10 November 2023. [Online]. Available: https://european-union.europa.eu/principles-countries-history/languages_en.
[13] A. Naveen, et al. Massively multilingual neural machine translation in the wild: findings and challenges, (2019) arXiv preprint arXiv:1907.05019.
[14] E. de Vries, M. Schoonvelde, G. Schumacher, No longer lost in translation: evidence that google translate works for comparative bag-of-words text applications, Polit. Anal. 26 (4) (2018) 417–430, doi:10.1017/pan.2018.26.
[15] K. Benoit, et al., Quanteda: an R package for the quantitative analysis of textual data, J. Open Source Softw. 3 (30) (2018) 774, doi:10.21105/joss.00774.
[16] I. Feinerer, K. Hornik, A. Software, tm: text Mining Package. (2024). Accessed: 22 August 2024. [Online]. Available: https://cran.r-project.org/web/packages/tm/index.html.
[17] T. Rinker et al., lexicon: lexicons for Text Analysis. (2019). Accessed: 22 August 2024. [Online]. Available: https://cran.r-project.org/web/packages/lexicon/index.html.
[18] M. Měchura, Lemmatization Lists. (2023). Accessed: 01 March, 2023. [Online]. Available: https://github.com/michmech/lemmatization-lists.