The Triangle of polarization, political trust and political communication: Understanding its dynamics in contemporary democracies.

(TRI-POL) (2019-2022)

ARGENTINA, CHILE, ITALY, PORTUGAL AND SPAIN

Big Data Codebook

Funding

This research was funded by two competitive grants. First, the Spanish Ministry of Economy and Competitivity. Ministerio de Economía y Competitividad, Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia PID2019-106867RB-I00 /AEI/10.13039/501100011033 (2020-2024), Principal Investigador: Mariano Torcal). Second, by the Fundación BBVA, Ayudas a Equipos de Investigación Científica en Economía y Sociedad Digital 2019 (2020-2022). The views expressed herein are those of the authors and are not necessarily those of these two funding agencies. The PI of the project is also grateful for the

funding provided by the Advance Research Fellowship Programme ICREA, funded by the Catalonian Government.

Index

Index of Tables **3**

Technical Information **5**

1. Citation, Research Team and Contact **5**

Citation 5

Research Team 5

Contact 5

2. Data Description **6**

Overview 6

Files 6

3. General Design of the Study **8**

Field 8

Universe 8

Fieldwork 8

Sampling Method 8

Period of analysis 8

4. Sample and Procedure **10**

Sample Procedure 10

Structure of the Sample 12

5. Variables and Labels in the aggregate data file **17**

Index of Tables

Table 1 Files produced by the TRI-POL big data study 6
Table 2 Subperiods of analysis of tweets 9
Table 3 Structure of the Sample 12
Table 4 Structure of the Sample of Tweets 13
Table 5 List of Global Variables 17
Table 6 List of Variables for news outlets' data 19
Table 7 List of Variables for Twitter data 20
Table 8 List of political entities detected in each country 21

TRI-POL 2021-2022 Big Data Codebook

Technical Information

1. Citation, Research Team and Contact

Citation

This dataset is provided free of charge for all those who wish to use it. Designing this study, retrieving the data, cleaning it, and preparing it for public use meant a lot of work. We are therefore grateful for your acknowledgment of our efforts by citing this codebook when you use it. The suggested citation is the following:

Arcila Calderón, Carlos, Mariano Torcal Loriente, Inmaculada García and David Blanco-Herrero (2022). "The Triangle of Polarization, Political Confidence and Political Competition: Understanding its Dynamics in Five Contemporary Democracies", *Big Data Codebook.*

I

Research Team

Carlos Arcila Calderón (Universidad de Salamanca)
Mariano Tocal Loriente (Universitat Pompeu Fabra)
Inmaculada García
David Blanco-Herrero (Universidad de Salamanca)

Contact

Carlos Arcila Calderón
Mail: carcila@usal.es

Additionally, you can contact:
David Blanco-Herrero
Mail: david.blanco.herrero@usal.es

2. Data Description

Overview

The TRI-POL big data study has collected large volumes of data from digital and social media from five countries: Argentina, Chile, Italy, Portugal and Spain over a period of over one yar and a half between July 2020 and January 2022 (the detailed date information in Table 1). This data will be also studied in combination with other data collected in the framework of the project.

The following pages focus on the technical information concerning the big data approach.

Files

10 datasets: one for the messages collected from Twitter accounts and one for texts collected from the news outlet in each of the five countries (for each of them, a .csv and a .sav file). Each of these datasets includes the detected entities, as well as their frequency of appearance in each text. In the case of the Twitter accounts, metadata such as number of replies, likes and retweets are also present.

1 combined dataset with all the content: for each case, the variables included here are polarization to the right and to the left, negative, positive and total sentiment, date, period, country, and type of content (a .csv and a .sav file). This dataset does not include the political entities analysed in each country, given that they are not comparable; the entities of each country are present in the respective national datasets.

1 qualitative report, with general information about the type of content and the patterns of publication and content in each studied Twitter account and news outlet (a PDF file)

1 data protocol with methodological information to help understanding all the Big Data elements of the TriPol project (a PDF file).

Table 1 resumes these files.

Table 1 Files produced by the TRI-POL big data study

Name	Format	Brief description
argentina_for_analysis	.sav / .csv	Analysed texts collected from Argentinean news outlets
argentina_twitter_for_analysis	.sav / .csv	Analysed texts collected from Argentinean Twitter accounts
chile_for_analysis	.sav / .csv	Analysed texts collected from Chilean news outlets
chile_twitter_for_analysis	.sav / .csv	Analysed texts collected from Chilean Twitter accounts
italy_for_analysis	.sav / .csv	Analysed texts collected from Italian news outlets

italy_twitter_for_analysis	.sav / .csv	Analysed texts collected from Italian Twitter accounts
portugal_for_analysis	.sav / .csv	Analysed texts collected from Portuguese news outlets
portugal_twitter_for_analysis	.sav / .csv	Analysed texts collected from Portuguese Twitter accounts
spain_for_analysis	.sav / .csv	Analysed texts collected from Spanish news outlets
spain_twitter_for_analysis	.sav / .csv	Analysed texts collected from Spanish Twitter accounts
Joint variables	.sav / .csv	Combined dataset with the general analysis for all the texts from all countries and type of platform
Big_Data_Qualitative_Report	.pdf	Report with a qualitative analysis of the studied countries and type of platform
Big_Data_Protocol	.pdf	Report with the codebook and methodological information of the big data study of the TRI-POL project

Source: own elaboration.

3. General Design of the Study

Field

International (Argentina, Chile, Italy, Portugal and Spain).

Universe

1. Opinion articles and editorials from Spanish, Argentine, Chilean, Portuguese and Italian media,

2. Posts from Twitter accounts of parties, politicians and entities of these countries.

Table 2 shows the specific news outlets and Twitter account present in the study.

Fieldwork

Two strategies: 1) Access to the archives of the selected news outlets and download of the relevant materials,

2) Use of the REST function of the Application Programming Interface (API) provided by Twitter for developers, downloading the tweets published by the selected accounts during the stablished period.

Sampling Method

All the relevant Twitter accounts and news outlet were selected following the criteria of national experts participating in this project from each country. For the Twitter accounts, all the content could be downloaded, but for the news outlets the diagonally composed week was used as a sampling method, reaching 22% of the universe. Not all the news outlet could be studied, given that some of them did not have opinion sections or they could not be accessed.

Period of analysis

The total period of analysis goes between July 1, 2020 and January 6, 2022 both for the news outlets and for the Twitter accounts. However, the content produced by the Twitter accounts was subdivided into three periods, given that this content was being compared with experimental surveys and passive meter data collected in other fields of the project. The content collected from news outlet was also divided into these three periods, even though it does not belong to the experimental study, in order to establish

potential comparison with the Twitter information. Table 2 details the distribution of the three phases in which the downloaded data can be classified.

Table 2 Subperiods of analysis of tweets

Period	Begin	End	Days	Downloaded tweets	Downloaded articles
Baseline period	01/07/2020	22/09/2021	448	185426	58357
First experimental period	23/09/2021	18/11/2021	56	21667	10398
Second experimental period	19/11/2021	06/01/2022	58	11985	6146
Complete period	01/07/2020	06/01/2022	554	219078	74901

Source: own elaboration.

4. Sample and Procedure

Sample Procedure

The first phase of this part of the study began in August 2021 and it had the goal of collecting all Editorials, Opinion articles and Columns with political content from a selection of media from each of the studied countries. Once all the archives of the media were identified, they were accessed (using a subscription if necessary), and when the archives were not an option, the corresponding sections were searched in order to copy the political opinion texts in TXT files. As aforementioned, not all the originally planned outlets could be included in the sample. The sampling method was the diagonally composed week, reaching 22% of the universe (all the opinions texts about politics produced by the studied media between July 1st, 2020 and January 6th, 2022). Given that the unit of analysis was the paragraph, all the texts were automatically divided into different paragraphs, so that they could all be added later to the dataset as independent units.

Similarly, all the tweets (excluding retweets) from the Twitter accounts of the main political actors of the countries under study from the same dates were collected. For the download, the Application Programming Interface (API) provided by Twitter for developers was used, employing the REST function, which allows downloading the tweets published by the selected accounts, filtering by hashtags or keywords, during a particular period of time or since the creation of the account. The Twitter messages were downloaded in .csv and stored in Excel format.

Methodology

Once the collection of content was concluded, the **first task** was to detect the different political entities (see Table 6) present throughout the texts as well as their frequency of appearance. The procedure followed during this phase was:

(1) A local expert from each country was asked to identify those surnames or terms that have a double meaning and that may a problem for the correct detection within our case study. For example, the term "Ciudadanos" (citizens in Spanish) can lead to confusion within the political sphere given its double interpretation: as the name of a political party or as a generally used noun.

(2) Once these terms had been identified, a list was generated with the most frequent neighbors in order to proceed to their correct detection. To do this, all the words neighboring these terms were extracted from the text through the tokenization process that allows the sample texts to be broken down into all its linguistic units; then, all the stop-words (words that accompany the terms but do not provide relevant information, such as articles) were removed; then the term that interests us in each text was identified and the neighbors to its right and left were selected, and the most frequent ones were selected returned to the experts so that they could identify which ones can help to correctly and unequivocally detect the problematic term. For example, in the case of "Ciudadanos", the neighbor "party" or "Arrimadas" (the national leader of said party) would help to the correct

detection.

(3) Finally, once the neighbors that help the correct detection had been identified, the entities were detected using tokenization and dividing the texts into bigrams that contemplate the options of appearance right and left in the texts ("arrimadas – Ciudadanos" and "Ciudadanos - arrimadas") so that the co-occurrence of the entities with the adequate neighbors would help counting the number of appearances of the entities. In this way, it is ensured that the frequency and detection is, although conservative, safe.

The **second task** was the creation of dictionaries to measure the presence of positive and negative feelings, and of polarization towards left and right in each of the units of analysis. For the measure of the sentiment, the already validated and well-used SentiStrenght dictionaries (in Spanish, Italian and Portuguese, depending on the country) were employed. These dictionaries are based on terms with negative (from -1 to -5, being -5 the terms expressing a most negative sentiment) and with positive values (from 1 to 5, being 5 the terms expressing a most positive sentiment).

This model was used as a basis for the measure of left and right polarization, defined as the expression of positive valence and the direction it takes (right, left, neutral). For measuring polarization five ad-hoc dictionaries (one for each country in the sample) with 100 words each were developed using the words with highest frequencies of appearance in each the text of each country, removing stop-words and entities (Pablo, Casado, Ciudadanos, etc.) and manually detecting which ones can generate polarization, eliminating the rest. With this, a list of 100 potential polarizing words were generated so that local experts could check whether these words were likely to generate polarization, to assign a value between 0 and 5, where 0 shows no polarization and 5 shows maximum polarization, to each direction (left, right or both), and to add as many words as they deemed appropriate. Then, using these dictionaries in a similar way to the SentiStrength ones, each text could be classified between 1 and 5, where 1 shows no polarization and 5 shows maximum polarization either to the right or left, depending on the dictionary (this change from 0 to 5 into 1 to 5 was necessary because the SentiStrength structure used as a basis required it). The dictionary of polarization to the left is made up of words used by supporters of leftist ideologies to refer to voters/supporters/leaders of the opposite political spectrum, and the opposite is valid for the polarization to the right.

Finally, and after running the dictionaries –executing Python code, the final datasets containing analyzes of politics content in the media and tweets were obtained, using the aforementioned dictionaries on (a) extraction of entities, (b) positive and negative sentiment analysis; and (c) left and right polarization. Before starting the data analysis, a final cleaning was carried out, unifying formats (for example, in dates) or removing paragraphs (unit of analysis) that do not contain textual information such as blank spaces, asterisks, etc., and therefore, obtain null or neutral evaluations in all the variables.

Structure of the Sample

As it has been explained, the sample has two types of content: articles from news outlets and tweets. Table 3 shows the structure of the sample of articles, whereas Table 4 shows the one of the sample of tweets. The figures referred in these tables are those of the final sample, the one used for the study; however, it should be noted that the original database included a total of 318,314 units of analysis, that were reduced to the final 293,979 after the cleaning of the dataset, mostly by removing text without metadata for the analysis, or texts produced by the automatic division of the news media articles into paragraphs (the unit of analysis in this category) that had no textual information (such as symbols between paragraphs, for example).

Table 3 Structure of the Sample

Source	Subperiod 1	Subperiod 2	Subperiod 3	Total
Argentina	**11582**	**2121**	**1089**	**14792**
Clarín	2665	432	0	3097
La Nación	1766	368	0	2134
Página 12	2111	429	499	3039
Perfil	1297	258	0	1555
Telam	2323	365	136	2824
TN	1420	269	454	2143
Chile	**8485**	**1503**	**1675**	**11663**
Cooperativa	1792	416	301	2509
El mercurio	2557	444	413	3414
El mostrador	2629	364	623	3616
La tercera	1507	279	338	2124
Italy	**5604**	**995**	**511**	**7110**
Corriere della Sera	1677	267	290	2234
Il Sole 24 Ore	429	161	166	756
La Repubblica	2366	304	0	2670
La Stampa	735	116	0	851
Libero	397	147	55	599
Portugal	**12798**	**3408**	**1481**	**17687**
Correio da Manha	998	179	0	1177
Diario de Noticias	2128	478	428	3034
Expresso	4158	601	0	4759
Jornal de Noticias	664	171	294	1129
Observador	2871	1069	0	3940
Publico	1793	217	0	2010
Sapo	186	693	759	1638

Spain	**19888**	**2371**	**1390**	**23649**
El Español	4901	619	0	5520
El Confidencial	4176	401	421	4998
El Periódico	1392	119	130	1641
Público.es	1124	207	223	1554
El Mundo	2560	230	0	2790
La Razón	253	160	257	670
ABC	704	181	0	885
El País	2785	269	359	3413
La Vanguardia	1993	185	0	2178
TOTAL	**58357**	**10398**	**6146**	**74901**

Source: own elaboration.

The news outlet that could not be part of the sample are:

- In Argentina: *Perfil*, *Infobae*, *El Intransigente*, *Diario Panorama* and *El Litoral*. Specifically, in the the case of *Perfil* it has not been possible to follow the method of the composed week in the article selection process since no sequence has been found in them. For this reason, all the political opinion articles found have been archived. The rest of the media that were part of the sample do not have an opinion section and consequently have not been part of the coding: *Infobae*, *El Intransigente*, *Diario Panorama* and *El Litoral*.

- In Chile: *Biobiochile*, *Lun*, *Emol*, *24 Horas*, *Publimetro* and *T13*. *Biobiochile* has a section in video format, so it was not possible to transfer it to TXT for the analysis. *Lun*, *Emol*, *24 Horas* and *Publimetro* have no opinion section, whereas *T13* has no opinion section and it is in video format.

- In Italy: *Gazzetta Del Sud, Ansa, Dagospy*, and *Virgilio*, *Giornale Di Sicilia*. they do not have an opinion section. In the case of *Gazzetta Del Sud* it has a blog, but not an opinion section.

- In Portugal: *Noticias ao Minuto*, *RTP* and *TVI 24*. They do not have an opinion section. Regarding *Sapo*, the articles found are minimal because it does not have an opinion section, but instead has a blog dedicated to social issues, the health field and with few political opinion articles.

- In Spain: All originally selected media could be included, but *Público* and *La Razón* have a lower volume of units due to the lack of an archive.

Table 4 Structure of the Sample of Tweets

Source	**Baseline period**	**First experimental period**	**Second experimental period**	**Total**

Argentina	**22567**	**2298**	**1419**	**26284**
@alferdez	845	196	101	1142
@alfredocornejo	508	113	51	672
@AntonioBonfatti	54	4	0	58
@CasaRosada	1602	192	121	1915
@CFKArgentina	296	53	36	385
@danielscioli	849	81	84	1014
@DiputadosAR	3414	386	205	4005
@elisacarrio	912	206	86	1204
@GerardoMorales	815	81	40	936
@horariorlarreta	2172	220	156	2548
@joseluisgioja	206	19	10	235
@mauriciomacri	156	44	16	216
@NicolasdelCano	1225	141	87	1453
@PatoBullrich	9513	562	426	10501
@SenadoArgentina	0	0	0	0
Chile	**38840**	**5269**	**2856**	**46965**
@Camara_cl	5690	639	463	6792
@danieljadue	1602	248	109	1959
@desbordes	950	66	187	1203
@evelynmatthei'	2120	356	243	2719
@gabrielboric'	2850	549	510	3909
@GobiernodeChile	2127	265	245	2637
@HeraldoMunoz	623	37	29	689
@ignaciobriones_	866	95	23	984
@joseantoniokast	3931	752	364	5047
@LavinJoaquin	5437	0	0	5437
@MaldonadoCurti	408	27	0	435
@PamJiles	3728	173	148	4049
@paulanarvaezo	620	43	37	700
@ProvosteYasna	638	259	49	946
@sebastiansichel	664	1136	26	1826
@Senado_Chile	4536	579	384	5499
@ximerincon	2050	45	39	2134
Italy	**26893**	**31513**	**2229**	**32275**
@berlusconi	604	54	18	676
@EnricoLetta	1040	162	61	1263
@euronewsit	8731	1022	816	10569
@Europarl_IT	1226	183	152	1561
@GiorgiaMeloni	1950	268	186	2404
@GiuseppeConteIT	299	79	59	437
@matteorenzi	932	110	55	1097
@matteosalvinimi	7194	330	228	7752

@Montecitorio	3731	664	426	4821
@Palazzo_Chigi	960	253	198	1411
@robersperanza	226	28	30	284
Portugal	**10852**	**1436**	**937**	**13225**
@AndreCVentura	1349	190	180	1719
@catarina_mart	1058	113	67	1238
@DGSaude	1012	222	181	1415
@francisco__rs	315	13	0	328
@GovernoMadeira	1157	131	136	1424
@govpt	487	12	8	507
@jcf_liberal	112	4	15	131
@OsVerdes	682	43	0	725
@Partido_PAN	268	46	3	317
@pcp_pt	2703	488	251	3442
@RuiRioPSD	344	33	14	391
@ruitavares	1365	141	82	1588
Spain	**86274**	**9511**	**4544**	**100329**
@anaponton	446	58	0	504
@ArnaldoOtegi	268	68	40	376
@CNNEE	63249	5985	2016	71250
@desdelamoncloa	4611	554	475	5640
@euronewses	10585	1547	1152	13284
@Europarl_ES	973	229	159	1361
@InesArrimadas	1108	170	144	1422
@iurkullu	271	37	19	327
@joanbaldovi	676	67	35	778
@junqueras	2	0	0	2
@LauraBorras	334	34	10	378
@pablocasado_	1550	216	164	1930
@PabloIglesias	53	218	204	475
@sanchezcastejon	1319	177	0	1496
@Santi_ABASCAL	829	151	126	1106
TOTAL	**185426**	**50027**	**11985**	**219078**

Source: own elaboration.

5. Variables and Labels in the aggregate data file

Once the texts were collected and the datasets were cleaned, and after automatically applying the aforementioned dictionaries, each of the units of analysis (paragraphs in the case of news outlets' opinion pieces and tweets in the case of texts from Twitter accounts) could be studied. The main variables used in the study, together with their explanation, are referred in Table 5.

Variable names, variable labels and value labels are all in English except when they refer to the proper nouns of the studied news outlets, the Twitter usernames of the studied account or to the entities studied in each country, which are maintained in their original language.

Table 5 List of Global Variables

Variable name	Variable type	Variable label	values
text	String	Analyzed text	A paragraph or a tweet
polright	Numeric	Level of polarization to the right	From 1 to 5, being 1 the lowest level and 5 the highest level of polarization
polleft	Numeric	Level of polarization to the left	From 1 to 5, being 1 the lowest level and 5 the highest level of polarization
negsent	Numeric	Level of negative sentiment	From 1 to 5, being 1 the lowest level and 5 the highest level of negativity
possent	Numeric	Level of positive sentiment	From 1 to 5, being 1 the lowest level and 5 the highest level of positivity
totsent	Numeric	Total sentiment (negsent+possent)	From -4 to 4, being -4 the most negative sentiment and 4 the most positive sentiment
date	Date	Date of the publication of the text	From 01.07.2020 to 06.01.2022
period	Numeric	Period of the publication of the text	1. 01.07.20-22.09.21; 2. 23.09.21-18.11.21; 3. 19.11.21-06.01.22
outlet_AR	Numeric	Argentinian news outlets	0. Others[1]; 1. Clarín; 2. La Nación; 3. Página 12; 4. Perfil; 5. Telam; 6. TN
twitter_AR	Numeric	Argentinean Twitter accounts	0. Others[1]; 1. @alferdez; 2. @alfredocornejo; 3. @AntonioBonfatti; 4. @CasaRosada; 5. @CFKArgentina; 6. @danielscioli; 7. @DiputadosAR; 8. @elisacarrio; 9. @GerardoMorales; 10. @horaciorlarreta; 11. @joseluisgioja; 12. @mauriciomacri; 13. @NicolasdelCano; 14. @PatoBullrich; 15. @SenadoArgentina
outlet_CH	Numeric	Chilean news outlets	0. Others[1]; 1. Cooperativa; 2. El mercurio; 3. El mostrador; 4. La tercera

Variable name	Variable type	Variable label	values
twitter_CH	Numeric	Chilean Twitter accounts	0. Others[1]; 1. @Camara_cl; 2. @danieljadue; 3. @desbordes; 4. @evelynmatthei; 5. @gabrielboric; 6. @GobiernodeChile; 7. @HeraldoMunoz; 8. @ignaciobriones_; 9. @joseantoniokast; 10. @LavinJoaquin; 11. @MaldonadoCurti; 12. @PamJiles; 13. @paulanarvaezo; 14. @ProvosteYasna; 15. @sebastiansichel; 16. @Senado_Chile; 17. @ximerincon
outlet_IT	Numeric	Italian news outlets	0. Others[1]; 1. Corriere della Sera; 2. Il Sole 24 Ore; 3. La Repubblica; 4. La Stampa; 5. Libero
twitter_IT	Numeric	Italian Twitter accounts	0. Others[1]; 1. @berlusconi; 2. @EnricoLetta; 3. Euronewsit; 4. @Europarl_IT; 5. @GiorgiaMeloni; 6. @GiuseppeConteIT; 7. @matteorenzi; 8. @matteosalvinimi; 9. @Montecitorio; 10. @Palazzo_Chigi; 11. @robersperanza
outlet_PT	Numeric	Portuguese news outlets	0. Others[1]; 1. Correio da Manha; 2. Diario de Noticias; 3. Expresso; 4. Jornal de Noticias; 5. Observador; 6. Publico; 7. Sapo
twitter_PT	Numeric	Portuguese Twitter accounts	0. Others[1]; 1. @AndreCVentura; 2. @catarina_mart; 3. @DGSaude; 4. @francisco__rs; 5. @GovernoMadeira; 6. @govpt; 7. @jcf_liberal; 8. @OsVerdes; 9. @Partido_PAN; 10. @pcp_pt; 11 @RuiRioPSD; 12. @ruitavares
outlet_ES	Numeric	Spanish news outlets	0. Others[1]; 1. El Español; 2. El Confidencial; 3. El Periódico; 4. Público.es; 5. El Mundo; 6. La Razón; 7. ABC; 8. El País; 9. La Vanguardia
twitter_ES	Numeric	Spanish Twitter accounts	0. Others[1]; 1. @anaponton; 2. @ArnaldoOtegi; 3. @CNNEE; 4. @desdelamoncloa; 5. @euronewses; 6. @Europarl_ES; 7. @InesArrimadas; 8. @iurkullu; 9. @joanbaldovi; 10. @junqueras; 11. @LauraBorras; 12. @pablocasado; 13. @PabloIglesias; 14. @sanchezcastejon; 15. @Santi_ABASCAL
country	Numeric	Country in which the text was published	1. Argentina; 2. Chile; 3. Italy; 4. Portugal; 5. Spain
type	Numeric	Type of the text (paragraph or tweet)	1. News outlet; 2. Twitter

Source: own elaboration

[1]: All the cases not belonging to the country and type of content in question were classified in this category

in these variables.

The variables described in Table 5 are the ones that can be found in the joint dataset in which all the variables are included. Variables such as the presence of the detected political entities were not included here given that they were different for each country, which would have led to a very extent and unhandy dataset. Thus, given that these entities are not comparable between countries, they were only included in the specific datasets of each country. This way, each country had two datasets, one for the news outlets and one of the Twitter accounts, thus making a total of ten datasets. Tables 6 and 7 show the variables studied for each of the two types of content, that is, for news outlets and for tweets, respectively.

Table 6 List of Variables for news outlets' data

Variable name	**Variable type**	**Variable label**	**values**
text	String	Analyzed text	A paragraph collected from opinion pieces
polright	Numeric	Level of polarization to the right	From 1 to 5, being 1 the lowest level and 5 the highest level of polarization
polleft	Numeric	Level of polarization to the left	From 1 to 5, being 1 the lowest level and 5 the highest level of polarization
negsent	Numeric	Level of negative sentiment	From 1 to 5, being 1 the lowest level and 5 the highest level of negativity
possent	Numeric	Level of positive sentiment	From 1 to 5, being 1 the lowest level and 5 the highest level of positivity
totsent	Numeric	Total sentiment (negsent+possent)	From -4 to 4, being -4 the most negative sentiment and 4 the most positive sentiment
date	Date	Date of the publication of the text	From 01.07.2020 to 06.01.2022
period	Numeric	Period of the publication of the text	1. 01.07.20-22.09.21; 2. 23.09.21-18.11.21; 3. 19.11.21-06.01.22
media	Numeric	The news outlets of the country in question.	The news outlets in each country can be seen in Tables 2 and 3
reference	String	Name of the downloaded file from each outlet and from which the paragraph is extracted. Texts are classified as: 1. Editorials. 2. Opinion articles. 3. Columns	COUNTRY_ABBREVIATION OF THE NEWS OUTLET_DATE_TYPE OF ARTICLE
Entity[1]	Numeric	Number of occurences of each entity in the studied text	From 0, when the entity is not mentioned in the studied text, onwards.

Source: own elaboration

[1]: Each entity was a variable of its own. In order to avoid too long tables, the entities detected in each country can be seen in Table 8.

Table 7 List of Variables for Twitter data

Variable name	Variable type	Variable label	values
text	String	Analyzed text	A tweet
polright	Numeric	Level of polarization to the right	From 1 to 5, being 1 the lowest level and 5 the highest level of polarization
polleft	Numeric	Level of polarization to the left	From 1 to 5, being 1 the lowest level and 5 the highest level of polarization
negsent	Numeric	Level of negative sentiment	From 1 to 5, being 1 the lowest level and 5 the highest level of negativity
possent	Numeric	Level of positive sentiment	From 1 to 5, being 1 the lowest level and 5 the highest level of positivity
totsent	Numeric	Total sentiment (negsent+possent)	From -4 to 4, being -4 the most negative sentiment and 4 the most positive sentiment
date	Date	Date of the publication of the text	From 01.07.2020 to 06.01.2022
period	Numeric	Period of the publication of the text	1. 01.07.20-22.09.21; 2. 23.09.21-18.11.21; 3. 19.11.21-06.01.22
user	Numeric	The Twitter accounts of the country in question.	The Twitter accounts in each country can be seen in Tables 2 and 3
reference	Numeric	Unique ID of the tweet	Numeric value
author_id	Numeric	Twitter's ID of the author of the tweet	Numeric value
public_metrics.retweet_count	Numeric	Number of times the tweet has been retweeted	From 0, when the tweet has not been retweeted, onwards.
public_metrics.reply_count	Numeric	Number of replies the tweet has received	From 0, when the tweet has not been replied, onwards.
public_metrics.like_count	Numeric	Number of likes the tweet has received	From 0, when the tweet has not been liked, onwards.
public_metrics.quote_count	Numeric	Number of times the tweet has been quoted	From 0, when the tweet has not been quoted, onwards.
Entity[1]	Numeric	Number of occurences of each entity in the studied text	From 0, when the entity is not mentioned in the studied text, onwards.

Source: own elaboration

[1]: Each entity was a variable of its own. In order to avoid too long tables, the entities detected in each country can be seen in Table 6.

Table 8 List of political entities detected in each country

Country	Variable type
Argentina	Fernández; Macri; Bullrich; Cornejo; Morales; Carrió; Gioja; Scioli; Bonfatti; FdT; PRO; UCR; CCARI; PJ; PS; Kirchneristas; Macristas; Socialistas; AlbertoFernández; MauricioMacri; PatriciaBullrich; DelCaño; RodríguezLarreta; AlfredoCornejo; GerardoMorales; ElisaCarrió; DanielScioli; AntonioBonfatti; CoalicióCívida; PartidoJusticialista; PartidoSocialista; RadicalesArgentinos; VotantesdelPRO; VotantesdePJ; HoracioRodríguezLarreta; JoséLuisGioja; NicolásDelCaño; Frentedetodos; UniónCívicaRadical; VotantesdeFI; CristinaFernándezdeKirchner; VotantesdelaUCR; VotantesdelPartidoSocialista; VotantesdeFrentedeTodos; VotantesdelaCoalicióCívica
Chile	Kast; Lavín; Matthei; Desbordes; Briones; Sichel; Rincón; Provoste; Muñoz; Narvaez; Maldonado; PLR; UDI; Jadue; Jiles; Boric; RN; EVOPOLI; EVO; ChV; PDC; PPD; PC; PH; FA; DC; PS; PR; Socialistas; Radicales; Comunistas; Humanistas; JoaquínLavín; EvelynMatthei; MarioDesbordes; IgnacioBriones; SebastianSichel; XimenaRincón; YasnaProvoste; HeraldoMuñoz; PaulaNarvaez; CarlosMaldonado; DanielJadue; PamelaJiles; GabrielBoric; PartidoRepublicano; RenovaciónNacional; EvoluciónPolítica; DemocraciaCristiana; PartidoSocialista; PartidoComunista; PartidoHumanista; PartidoRadical; FrenteAmplio; DemócratasCristianos; JoséAntonioKast; UniónDemócrataIndependiente; IndependienteChileVamos; VotantesdeEVOPOLI; VotantesdeSichel; PartidoporlaDemocracia; VotantesdelPartidoRepublicano; VotantesdeRenovaciónNacional; VotantesdelPartidoSocialista; VotantesdelPartidoRadical; VotantesdelPartidoComunista; VotantesdelPartidoHumanista; VotantesdelFrenteAmplio; VotantesdelaUniónDemocróticaIndependiente; VotantesdelPartidoporlaDemocracia
Italy	Letta; Conte; Speranza; Renzi; Meloni; Berlusconi; Salvini; PD; M5S; LeU; IV; Fdi; FI; LEGA; Enrico.Letta; Giuseppe.Conte; Roberto.Speranza; Matteo.Renzi; Giorgia.Meloni; Silvio.Berlusconi; Matteo.Salvini; Partito.Democratico; Italia.Viva Fratelli.d.Italia; Forza.Italia; Movimento.5.Stelle; Liberi.e.Uguali; Gli.elettori.della.Lega; Gli.elettori.del.Partito.Democratico; Gli.elettori.di.Italia.Viva; Gli.elettori.di.Fratelli.d.Italia; Gli.elettori.di.Forza.Italia; gli.elettori.del.movimiento.5.stelle; gli.elettori.di.liberi.e.uguali
Portugal	Costa; Martins; Rio; Figueiredo; Ventura; Tavares; PS; BE; PSD; PCP; CDS; PP; PEV; PAN; IL; CH; Chega; Livre; Ecologistas; PessoasAnimaisNatureza; AntónioCosta; CatarinaMartins; RuiRio; AndréVentura; RuiTavares; PartidoEcologista; PartidoSocialista; IniciativaLiberal; Osverdes; JoãoCotrimFigueiredo; BlocodeEsquerda; PartidoSocialistaDemocrata; PartidoComunistaPortuguês; EleitoresdoPS; EleitoresdoBE; EleitoresdoPSD; militantesdoPSD; simpatizantesdoPSD; EleitoresdaCDU; EleitoresdoCDSPP; EleitoresdoPAN; EleitoresdaIL; EleitoresdoChega; EleitoresdoLivre
Spain	Sánchez; Iglesias; Arrimadas; Abascal; Casado; Junqueras; Baldoví; Borràs; Urkullu; Otegi; Clavijo; Pontón; PSOE; Ciudadanos; Cs; VOX; PP; ERC; CCPV; Compromís; Junts; PNV; CC; PNG; Socialistas; Sanchistas; Podemitas; Morados; Abascalistas; Peperos; Peperas; Casadistas; Junqueristas; Puigdemontistas; Abertzales; Pedro.Sánchez; Pablo.Iglesias; Inés.Arrimadas; Santiago.Abascal; Pablo.Casado; Oriol.Junqueras; Joan.BaldovÃ; Iñigo.Urkullu; Arnaldo.Otegi;

Country	Variable type
	Fernando.Clavijo; Ana.Pontón; Izquierda.Socialista; Partido.Socialista; Unidas.Podemos; Los.Populares; Fuerza.Naranja; Coalició.Compromís; EH.Bildu; Izquierda.Aberzale; Coalición.Canaria; nacionalistas.Vascos; Independentistas.Vascos; Nacionalistas.Gallegos; Partido.Popular.Español; Junts.per.Catalunya; Juntos.por.Cataluña; Partido.Nacionalista.Vasco; Euzko.Alderdi.Jeltzalea; Euskal.Herria.Bildu; Bloque.Nacionalista.Galego; Votantes.del.PSOE; Votantes.de.Podemos; Votantes.de.Ciudadanos; Votantes.de.VOX; Votantes.del.PP; Votantes.de.ERC; Votantes.de.Comrpomís; Votantes.de.Junst; Votantes.de.PNV; Votantes.de.EH.Bildu; Votantes.de.CC; Votantes.de.PNG; Partido.Socialista.Obrero.Español; Ciudadanos.Partido.de.la.Ciudadanía; Esquerra.Republicana.de.Catalunya; Izquierda.Republicana.de.Cataluña

Source: own elaboration