MATHEMATICAL POPULATION STUDIES
https://doi.org/10.1080/08898480.2024.2412337

Routledge
Taylor & Francis Group

Polarization and hate speech based on fuzzy logic and transformers: the case of the 2023 Spanish general elections

Q1 Juan Antonio Guevara Gil[a,b], Belén Casas-Mas[c], and José Manuel Robles[d]

[a]Department of Statistics and Data Science, Faculty of Statistics, Complutense University of Madrid, Q2 Spain; [b]The University Institute of Statistics and Data Science, Complutense University of Madrid; [c]Department of Sociology: Methodology and Theory, Faculty of Information Sciences, Complutense University of Madrid, Spain; [d]Department of Applied Sociology, Faculty of Sociology, Complutense University of Madrid, Spain

ABSTRACT

Affective polarization in the digital debate of the Spanish presidential election campaign (2023), following the sudden call of the Spanish president on July 23, was measured. Using transformers, topics were detected, and sentiment analysis techniques were applied in the political debate during the elections to measure the emotional valence of the debate. The topics that dominate most of the debate are Candidates (n1 = 17170) and Opposition (n3 = 15327). These topics also show the highest typical polarization deviances. Based on affective polarization, a polarization measure (JDJ) grounded in the fuzzy sets was applied. The topic activism has the highest polarization value, while the topic of voting has the lowest. This analysis highlights a dichotomy that defines the Spanish political reality: the positive image of conventional political participation in the face of the rejection of collective action processes.

KEYWORDS

Affective polarization; fuzzy sets; hate speech; sentiment analysis; topic modeling; transformers Q4

1. Introduction

Over the last decades, especially in the context of electoral campaigns and in debates through social media, a growing process of political polarization has been observed. This process has been linked to two interrelated trends. On the one hand, the growing use of discourses based on emotions and, on the other, on frameworks of interpretation of socio-political reality designed to strengthen citizens' starting points. The term affective polarization (Tucker et al., 2018) refers to the growing trend toward the constitution of poles of political opinion based, no longer on reasons and arguments, but on the emotional alignment between citizens and representatives. These interactions are based on feelings such as hatred or treason (negative emotions) and positive feelings such as unconditional adhesion or group identity cohesion (Iyengar et al., 2019). Emotive speeches are a mechanism that favors the

CONTACT Juan Antonio Guevara Gil juanguev@ucm.es Department of Statistics and Data Science, Faculty of Q3 Statistics, Complutense University of Madrid, Madrid, Spain.

formation of homogeneous opinions (Shore, Baek, and Dellarocas, 2016) and generates a sense of collective identity. This type of messages is more clearly expressed in digital communication spaces and, especially, during electoral processes. The possibility of defining the adversary as "the other," "the opponent" and generating, from that identification, attributes based on emotions, has permeated the digitally mediated political debate.

Social networks have shown to be an excellent mean of filtering the information that is consumed by citizens, as well as of selecting those spaces and interlocutors with whom they interact (Sunstein, 2018). Selective exposure (Stroud, 2010) has become a polarizing mechanism because it favors those certain partial interpretations, those diagnostic frameworks of reality that we share with those who think like us, are strengthened and entrenched. This ideological polarization is a consequence of the very nature and options offered by digital social networks, but also the result of the strengthening of certain frameworks of analysis of reality. Another factor that might lead to polarization is radicalization over social values. Based on controversial issues, the public opinion may be polarized when it comes to express support or opposition to questions such as same-sex marriage or abortion, among others (Lee et al., 2014). Subsequently, the political ideologies of the party activists appear to further polarize the discussions surrounding specific issues that affect the community. However, the most relevant aspect of this process is its consequences. We must understand digital political polarization as a process that begins with selective exposure to socio-political content and interactions. This circumstance generates emotions and feelings of belonging that interpret the political adversary as "the other" (Papacharissi, 2015). The main consequence of this process is, on the one hand, a discourse based on incivility, flaming, or public attack (Sobieraj and Berry, 2011) and, on the other hand, a rupture of the communicative flow. We call the latter failed communication, and it is a key element that jeopardizes the very conception of democratic processes, in general, and their digital version, in particular. We consider a particularly relevant issue that should be a focus of attention for public and academic decision-makers.

1.1. Polarization measurement

Polarization measurement has been addressed by way of different disciplines, such as economy (Wolfson, 1994 or Esteban and Ray, 1994), sociology (Montalvo and Reynal-Querol, 2005) or fuzzy logic (Guevara et al., 2020), modifying the understanding of polarization measurement and addressing it from new perspectives.

From the presentation of the *JDJ* polarization measurement based on the fuzzy sets (Zadeh, 1965) by Guevara et al. (2020) it is possible to take into consideration nuances of people's attitudes within the measurement itself,

which enables us to address the polarization from a more realistic perspective. The measurement premise focuses on the polarization measurement by way of the quantification of the degree of radicalization of the individuals to both poles or extremes at the same time. This proposal enables the authors to avoid the belongs/does not belong dichotomy of an individual to a category, assuming the existence of graduation as regards this belonging. This position avoids the nonrealistic crispness that an element fully belongs to a category, but implies the existence of gradualism in their belonging. In this vein, a person may be a supporter of one political party, but may identify—to a certain degree—with other proposals made by other political parties.

The fuzzy logic has been applied to Social Sciences problems since the last decades as it provides resources to better represent the concepts in the human brain for perceiving, recognizing, and categorizing natural phenomena as they are often vague and imprecise (Abdullah, Abdullah, and Tap, 2004). Fuzzy sets allow for nuanced representation of categorical concepts by permitting degrees of membership instead of binary inclusion. Additionally, fuzzy sets help evaluate set-theoretic relationships like intersection, inclusion, necessity, and sufficiency, which are challenging to assess using conventional methods like the general linear model (Ragin and Pennings, 2005).

The fuzzy logic has been demonstrated to also provide benefits in the polarization measurement from a dynamic perspective allowing to capture the risk of polarization for a pair of individuals in the near future (Guevara et al., 2022). Also, polarization measures based on the fuzzy sets outperform traditional measures applied to the case of graphs in Social Network Analysis to capture polarization from a relational perspective (Simón de Blas et al., 2022).

1.1.1. Fuzzy sets

Fuzzy sets have constituted one of the beginnings of a change of paradigm that has proven to be of great significance in the last century. Traditionally, Aristotelian two-valued logic like good-bad, high-low or a lot-a little have led researchers to contemplate their *crisp values* for variables that present a diffuse nature in themselves. The forced transformation of the nature of what is to be contemplated can lead to imprecision as regards that which is to be measured or modeled. Based on this principle, Zadeh (1965) proposed fuzzy sets. Fuzzy sets enable the representation of graduations in concepts that have traditionally been considered crisp, such as belonging to a category.

Definition 1.1 (Fuzzy sets) *An $\tilde{A}$ fuzzy set on the domain of X is defined as*:

$$\tilde{A} = \{(x, \mu_A(x)) \quad x \in X\} \tag{1}$$

where $\mu_A(x)$ represents the function of the membership degree: $\mu_A : X \rightarrow [0,1]$.

On the other hand, a determined object x may sustain determined membership degrees to more than one class c, where the total of its membership degrees is distributed along all the classes c, where $\mu_c(x) > 0$ if the membership degree of x to a class c is not null. The following must be fulfilled:

$$\sum_{c \in C} \mu_c(x) = 1 \tag{2}$$

where $\mu_c(x) \in [0,1]\ \forall x, \forall c$.

There is a natural relationship between the fuzzy sets and information aggregation operators. Information aggregation implies the dimensional simplification of the original information without losing its initial format (Montero et al., 2010). Originally, the aggregation operators were defined for the aggregation of values stemming from functions of membership degree associated with a fuzzy set. For example, in accordance with the aforementioned fuzzy classification where a determined object x can represent a certain membership degree to more than one class, this information may be aggregated for its combined consideration in one single value.

Definition 1.2 (Aggregate functions) *An* $A : [0,1]^n \rightarrow [0,1]$ function is an aggregate function in n provided that the following conditions are met:

- A is increasing in each argument, where for each $i \in \{1, 2, \ldots, n\}$ if $x_i \leq y$, $A(x_1, \ldots, x_n) \leq A(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$.
- A complies with the limits: $A(0, \ldots, 0) = 0$ y $A(1, \ldots, 1) = 1$.

Another concept of great significance within the aggregation operators is the *overlapping* and *grouping* functions. The concept of *overlapping* was introduced by Bustince et al. (2010) in order to measure the degree of overlapping presented by a specific object x to two different classes belonging to a fuzzy classification system.

Definition 1.3 (Overlapping function) *An* $O : [0,1]^2 \rightarrow [0,1]$ function is an overlap function provided that the following conditions are met:

- O is commutative.
- $O(x, y) = 0$ if and only if $xy = 0$.
- $O(x, y) = 1$ if and only if $xy = 1$.
- O is increasing in each argument.
- O is continuous.

As is the case with the overlapping functions, Bustince et al. (2010) proposed the *grouping* functions, with the aim of measuring the degree to which the combination of two classes, A and B, is supported.

Definition 1.4 (Grouping function) *A* $G : [0,1]^2 \to [0,1]$ function is a grouping function provided that the following conditions are met:

- G is commutative.
- $G(x,y) = 0$ if and only if $x = y = 0$.
- $G(x,y) = 1$ if and only if $x = 1$ o $y = 1$.
- G is increasing in each argument.
- G is continuous.

1.1.2. Polarization measurement with JDJ

Where $N = \{1, \ldots, n\}$ is a set of elements and X is a unidimensional variable where $x_i \in X$, which represents the attitudinal, opinion, affective, etc. axis of an individual $i \in N$ by which polarization will be measured. Likewise, with X_A and X_B as the poles—or extremes—of the variable X. With $\mu_A, \mu_B :\to [0,1]$ as the functions of the population N belonging to both poles, with $\mu_A(x_i)$ and $\mu_B(x_i)$ as the membership degrees of i to poles X_A and X_B. The risk of polarization between two individuals $\{i,j\}$ is considered as the possibility of the following two scenarios at the same time:

- The degree of radicalization of i to the pole X_A *vs* the degree of radicalization of j to the pole X_B.
- The degree of radicalization of i to the pole X_B *vs* the degree of radicalization of j to the pole X_A.

The *JDJ* measure is defined as:

$$JDJ(X) = \sum\nolimits_{i,j \in N, i \leq j} \varphi\left(\phi\left(\mu_{X_A}(i), \mu_{X_B}(j)\right), \left(\phi\left(\mu_{X_B}(i), \mu_{X_A}(j)\right)\right)\right) \quad (3)$$

Where $\phi : [0,1]^2 \to [0,1]$ is a overlapping operator and $\varphi : [0,1]^2 \to [0,1]$ is a grouping function. The *JDJ* measurement offers its maximum values in those cases in which 50% of the population presents maximum radicalization to X_A and the other 50% presents maximum radicalization to X_B. The minimum values are found not only when 100% of the population presents the same value in X, but when this value is an extreme, whether that be in X_A or in X_B.

This polarization measurement demonstrates better performance than other classical measurements by way of dynamic simulation models (Guevara et al., 2022) or when used in order to improve community detection issues in social media network analysis (Gutiérrez et al., 2021).

1.2. Transformers

The machine learning algorithms and artificial intelligence have provided a significant number of resources that can be implemented broadly across social and computer sciences. Thanks to these resources, tasks centered upon Natural Language Processing (NLP) or Computer Vision (CP) have demonstrated significantly improved performance in recent years. Transformers are deep learning algorithms that make use of the embeddings in order to handle text in a more realistic way. Thanks to the embeddings, words are represented as vectors in accordance with their positions in a specific phrase. A vector that represents a word will include information regarding the context.

The focus of the transformer developed by Vaswani et al. (2017) integrates attention with encoders and decoders in order to enhance the extraction of relational contexts. By opting for the transferred learning paradigm, which focuses on obtaining knowledge from one task to apply it to other related tasks, the model overcomes the limitations of isolated learning paradigms, commonly built from scratch (Acheampong, Nunoo-Mensah, and Chen, 2021).

Specifically, of the transformers dedicated to NLP, Bidirectional Encoder Representations from Transformers (BERT) proposed by Devlin et al. (2019) is one of the most widely used, capable of performing question answering, text summarization or sentiment analysis tasks among others. One of BERT's strengths is that it can understand language by undergoing training on Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). BERT is able to *understand* the context of words in order to better understand their meaning, enabling it to undertake NLP tasks highly efficiently. There are many algorithms that have been based on BERT in order to target its performance to specific tasks such as its robust version RoBERTa (Liu et al., 2019), contextualization of emotion with EmoContext (Huang, Trabelsi, and Zaïane, 2019), topic detection with BERTopic (Grootendorst, 2022) or emotion and hate speech detection in specific languages such as RoBERTuito (Pérez et al., 2021).

2. Materials and methods

2.1. Case study and data source

The case study focuses on Spain's 2023 general elections. Suddenly, on May 29, Spain's Prime Minister, Pedro Sánchez, called the elections for July 23, as a consequence of the results of the autonomic elections held on May 28, The unexpected nature of this event led to significant commotion in both the political and general spheres, creating great

debate surrounding the elections. We focused on the posts created by verified accounts, public groups, or pages on Facebook as actors that generate polarization.

The data were downloaded from the Meta application CrowdTangle—formerly Facebook-. This application enables the download of the full history of social networks such as Facebook. For Facebook, there is access to verified profiles, public groups, and pages. The data downloaded was undertaken between May 29, 2023 and July 23, 2023, using the hashtag #23J, including only posts written in Spanish, where a total of 34, 968 posts were obtained.

2.2. Research aims

The research aim is to measure affective polarization on Facebook by way of the posts created by politicians involved in the 2023 elections in Spain. In order to achieve this aim, three specific objectives are proposed:

- Detect the different topics present in the digital debate.
- Apply sentiment analysis in order to measure the emotional valence of the debate.
- Measure the affective polarization by topic.

2.3. Topic detection

For this task, the topic modeling algorithm named BERTopic (Grootendorst, 2022) has been applied. It is an NLP model which uses pre-trained machine learning models—transformers—based on the BERT language model. This new model is based on the tf-idf text structure in order to subsequently represent it in embeddings, undertake cluster analysis, and return the topics that may underlie the text. BERTopic has demonstrated consistent results and good performance across different scenarios and languages. Embeddings have the virtue of analyzing the text and its units taking into consideration their context, to ensure that the same word, in different contexts, will imply different connotations. This technology facilitates the approach to text processing from a more realistic perspective.

Clustering analysis will be used in order to evaluate the performance of the model, in addition to distances and similarities between words and topics calculated by way of the computation of the embeddings.

2.4. Sentiment analysis

In regard to sentiment analysis, transformers are also used. The RoBERTuito model presented by Pérez et al. (2021) in their *pysentimiento* library is employed. It is a trained language model with over 500 million Spanish tweets

Q6 **Table 1.** Message with non-explicit negative connotations.

Message	Negative	Neutral	Positive
Why did Pedro Sánchez call elections on 23 July, falling during the summer holidays for many?	0.85	0.13	0.015

capable of detecting, for example, irony. This model uses the base architecture of the RoBERTa transformer with the parameters applied by Nguyen, Vu, and Nguyen (2020) in their well-known BERTweet model of sentiment analysis in English. Following the application of the model, percentages reflecting the presence of negative, positive, or neutral content are obtained for each message. This NLP model enables us to undertake automatic sentiment analysis taking into consideration the contextualization of words, their meaning, in addition to the use of sarcasm, irony, or hate, A real example of our database in which the good performance of the model is observed is in the following message:

The transformer not only presents the option to measure the positive-negative valence of the messages but it also enables the detection of hateful and aggressive in the messages. The combination of both tasks will facilitate knowledge of the intensity and negativity of the debate in order to be able to observe its affective polarization.

2.5. Polarization measurement

In accordance with what has been indicated, in order to calculate *JDJ* three different elements must be established. Firstly, the construction of $\mu_A(x_i)$ y $\mu_B(x_i)$ must be formalized, obtaining the membership degree functions or degree of radicalization of each element i to poles X_A and X_B.

Given the case study selected and the tools employed, the affective polarization will be calculated. According to Garrido, Rodríguez, and Rodríguez (2021):

> Unlike ideological polarization... affective polarization refers to a distance of an emotional type, the difference between the adhesion or affect generated within us by those who share our political ideas and the rejection or antipathy awoken in us by those who defend other ideas. (p. 23)

Unlike ideological polarization... affective polarization refers to a distance of an emotional type, the difference between the adhesion or affect generated within us by those who share our political ideas and the rejection or antipathy awoken in us by those who defend other ideas. (p. 23)

In order to do so, the work variable will be the emotional balance obtained by way of RoBERTuito, where, on the one hand, the relative frequency of appearance of negative sentiment $0 < n_n < 1$, in addition to the relative frequency of positive sentiment $0 < n_p < 1$ are obtained. Additionally, the relative

frequency of the appearance of hateful and aggressive is obtained, equally $0 < n_h < 1$ y $0 < n_a < 1$, all of which comply with the ownership of the membership functions $0 < \mu_{A,B}(x_i) < 1$. Given that the connotations of aggressive and hateful contribute significant importance, in order to determine the poles (negative and positive), the normalized sum of the frequencies of negative sentiment, hateful, and aggressive is defined as the membership functions to the negative pole, where:

$$N = n_n + n_h + n_a \tag{4}$$

$$\mu_N = \frac{N - min(N)}{max(N) - min(N)} \tag{5}$$

As regards the membership function to the positive pole, it is defined as the normalized subtraction of the frequencies of positive sentiment, hateful, and aggressive.

$$P = n_p - n_h - n_a \tag{6}$$

$$\mu_P = \frac{P - min(P)}{max(P) - min(P)} \tag{7}$$

Finally, we must define which overlapping function ϕ, which the product will use, and the aggregation function, φ which will be the maximum. The *JDJ* formula which will be obtained is as follows:

$$JDJ(X) = \sum_{i,j \in N, i \leq j} max\left(\left(\mu_N(i) * \mu_P(j)\right), \left(\left(\mu_P(i) * \mu_N(j)\right)\right)\right) \tag{8}$$

Likewise, Eq.~(6) can be expressed in index so that its values are between 0 and 1, according to Robles et al. (2022). $NC = \frac{N*(N-1)}{2}$ is the number of total comparisons made in the calculation of *JDJ*:

$$JDJ = \frac{JDJ(X)}{NC} * 2 \tag{9}$$

3. Results

3.1. Topic modeling

The text was first processed in order to proceed to the fine-tuning stage of the BERTopic transformer. Our dictionary initially presented a total of 49,316 different words. Stopwords were eliminated by adding the recurrent words *http*, *www*, *com*, *facebook*, *https*, *youtube*, *html*, and *org* as a result of including URLs. Our dictionary was reduced to 49,050 (a reduction of 0.54%). Words that only contain text, with a length greater than 1 character, were tokenized.

In addition, terms with a frequency of occurrence of less than 15 were excluded, and counted together with unigrams and bigrams, leaving a total of 6949 different words in our dictionary, representing the 14.1% of the original. The sentence-transformer model used to generate the embeddings was *all-MiniLM-L6-v2* (Reimers and Gurevych, 2019). Likewise, *Spanish* language was indicated with an automatic topics number. This model uses transformers and c-TF-IDF (Class Term Frequency—Inverse Document Frequency) for its performance. c-TF-IDF are adjusted TF-IDF matrices to work on a cluster/categorical/topic level rather than a document level which takes into account the frequency of a word in class *c*, the frequency of a word across all classes, and the average number of words per class.

A total of 37 topics were detected. Thanks to the embeddings, it is possible to calculate the similarity between them, in addition to representing their distances in a vector space. The similarity was calculated with the consensus between the embeddings for each topic, representing the distances in Figure 1, in addition to their hierarchical analysis.

The results shown in Figure 1 facilitate the observation of the possible redundancy of topics according to their distances and similarity (Figure 2). Additionally, as shown in Figure 3, the hierarchical structure offers a possible solution for the reduction of topics. The most similar topics were merged

Figure 1. Distance, similarity and hierarchy for topics.

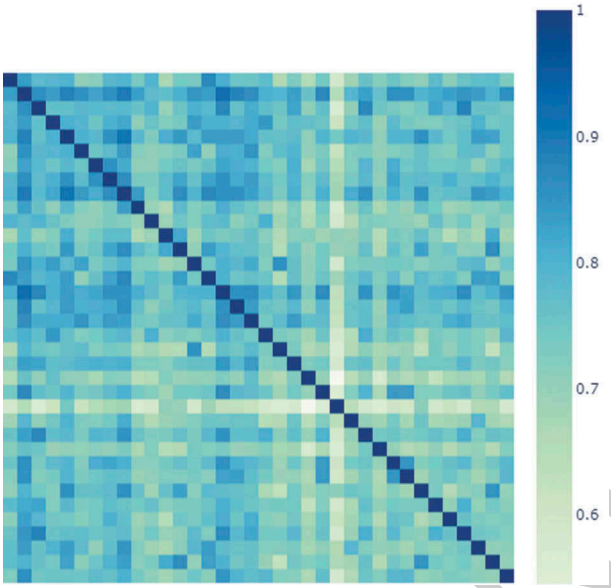

Figure 2. Similarity for topics.

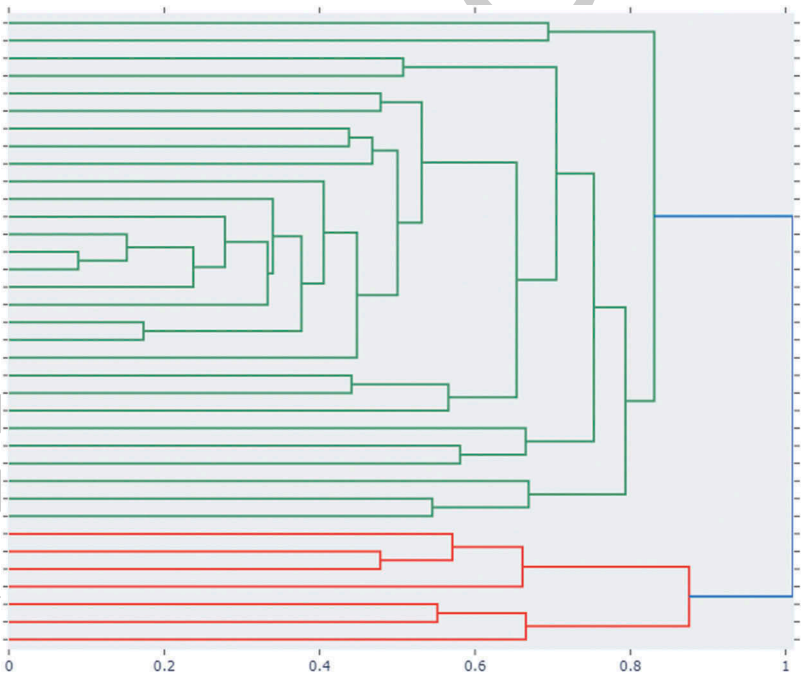

Figure 3. Hierarchical clustering.

together (Table 2), finally obtaining a total of six topics across the entire Facebook public debate, as shown in Table 3.

Table 2. Topics to be merged.

Original Topics	Topic
12, 22, 29, 15, 24, 16, 31	0
33, 23	1
19, 13, 28, 22, 15, 7, 17, 12, 9, 8, 3, 4, 1, 6, 16, 0, 2, 14, 24, 11, 29, 31, 35, 27	2
25, 21, 20	3
18, 10, 26, 34	4
5, 32, 30	5

Table 3. Number of messages by topic.

Topic	N
1. Candidates	17170
2. Activism	263
3. Opposition	15327
4. General debate	488
5. Voting	786
6. Electoral participation	934

The names of topics were interpreted by experts according to their most common words and wordclouds. In Table 3 we show the topics being Candidates ($n_1 = 17170$) and Opposition ($n_3 = 15327$) the most common ones, both occupying 93% of the debate.

4. Sentiment analysis

In regard to sentiment analysis, we employed the *pysentimiento* library for *Python*, as proposed by Pérez et al. (2021). In order to accomplish this task, two parallel analyses were applied: (1) sentiment analysis—positive, negative, or neutral-, and (2) hate speech analysis. While the first analysis allows us to know the presence of negative, positive, or neutral content in the message, the second one provides the percentage of occurrence of hateful, aggressive, and targeted nature of the message. The aim is to obtain both the emotional valence of the discourse, in addition to the presence of hateful. This not only enables us to know the emotional tone of the discourse that has been generated but also to segment it according to the previously detected topics. Figures 4 and 5 show the density functions for the emotions found in the general discourse both for the sentiment analysis and for the presence of hateful.

It can be observed that the *neutral* emotion predominates throughout the discourse, with a mean and standard deviation of relative occurrence of $\bar{x}_{Neu} = 0.54; sd_{Neu} = 0.236$, followed by negative emotion ($\bar{x}_{Neg} = 0.32; sd_{Neg} = 0.28$) and finally, by positive emotion ($\bar{x}_{Pos} = 0.14; sd_{Pos} = 0.167$). This result is expected given the nature of the available accounts on Facebook through CrowdTangle (verified profiles, public groups, and Facebook pages). The majority of these accounts represent public organizations and official entities, whose communication style is highly polite.

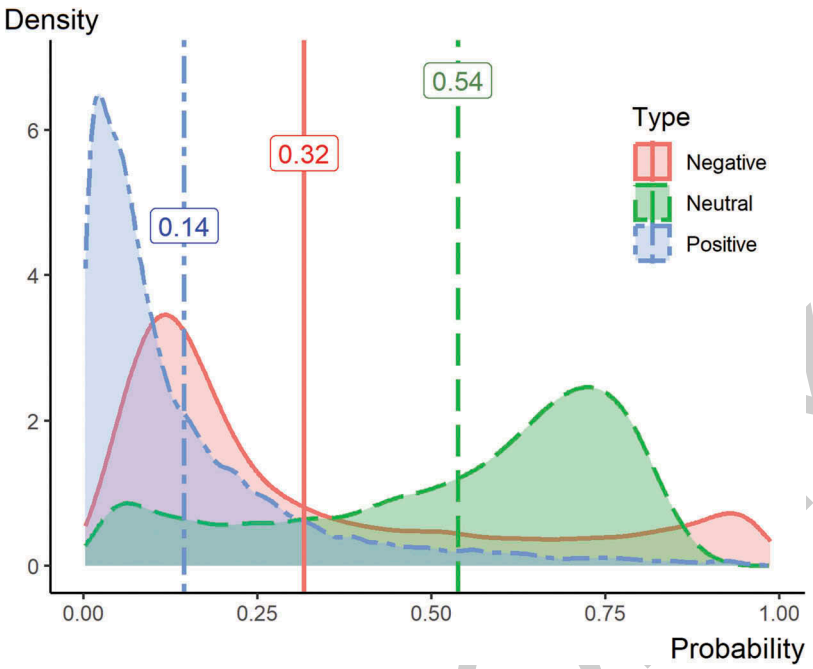

Figure 4. Density curves for negative, positive and neutral sentiments.

Nevertheless, it is noteworthy that the negative emotion is that which demonstrates a greater presence of high values >0.8 with respect to any of the other emotions. On the other hand, it can be observed that hateful, aggressive, and targeted have a very low presence in the discourse, with hateful showing the highest presence, with just 0.35% of average occurrence throughout the messages ($\bar{x}_H = 0.035; sd_H = 0.074$), followed by aggressive ($\bar{x}_A = 0.018; sd_A = 0.034$) and finally, targeted ($\bar{x}_T = 0.007; sd_T = 0.017$). It is worth mentioning that the density function for *hateful* presents slight prominence between the values 0.6 and 0.9, consequently skewing the mean upward. These values are not shown on the graph in order to facilitate the visualization given the low frequencies identified.

Once the emotional valence of the messages has been detected, it is possible to segment the results for each of the topics detected, the results of which can be seen in Table 4. In this case, the results related to the targeted nature of the message will be omitted given their low frequency, in addition to their loose link with the polarization calculation. The highest values of negativity are found in the topic *activism*, while the highest values of positivity are located in *electoral participation*, with the highest presence of neutral sentiment in the topic *Voting*. The highest values for hate speech are found in the topic *activism*, as well as those for aggressive.

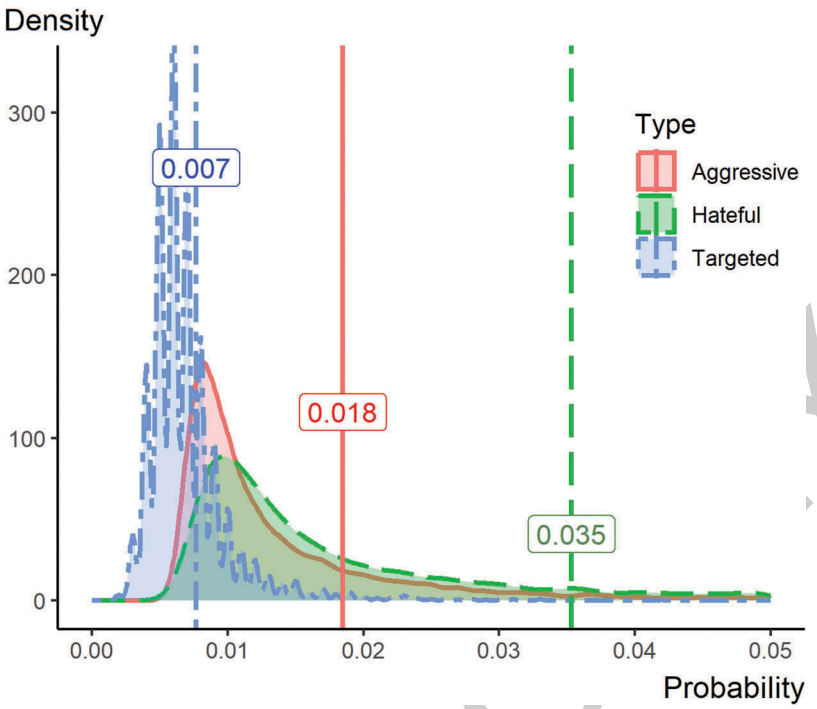

Figure 5. Density curves for hateful, aggressive and targeted content.

Table 4. Mean RoBERTuito punctuations by topic for each sentiment.

	Negative	Positive	Neutral	Hateful	Aggressive
Candidates	0.33	0.14	0.52	0.04	0.02
Activism	0.55	0.11	0.34	0.11	0.05
Opposition	0.30	0.15	0.55	0.03	0.02
General debate	0.34	0.15	0.51	0.02	0.01
Voting	0.24	0.10	0.66	0.02	0.01
Electoral participation	0.24	0.20	0.55	0.05	0.03
Global means	0.32	0.14	0.54	0.035	0.018

4.1. Polarization calculation

The scatter plot shown in Figure 6 facilitates a graphical representation as well as the relationship between the membership functions μ_N and μ_P. Given that the polarization values are determined by the membership functions, maximum polarization is expected when 50% of the cases have values close to 0 for the negative poles and close to 1 for the positive pole, and the other 50% vice versa. Finally, affective polarization is calculated, finding values $JDJ = 0.257$, where $JDJ \rightarrow [0, 1]$. The high concentration of values in Figure 6 around $x = 0.5 \text{and} y = 0.5$ and around $x = 0 \text{and} y = 1$ signifies a significant decrease in polarization, as the minimum values would occur in cases of maximum concentration at $x = 0$ and $y = 1$ or $x = 1$ and $y = 0$.

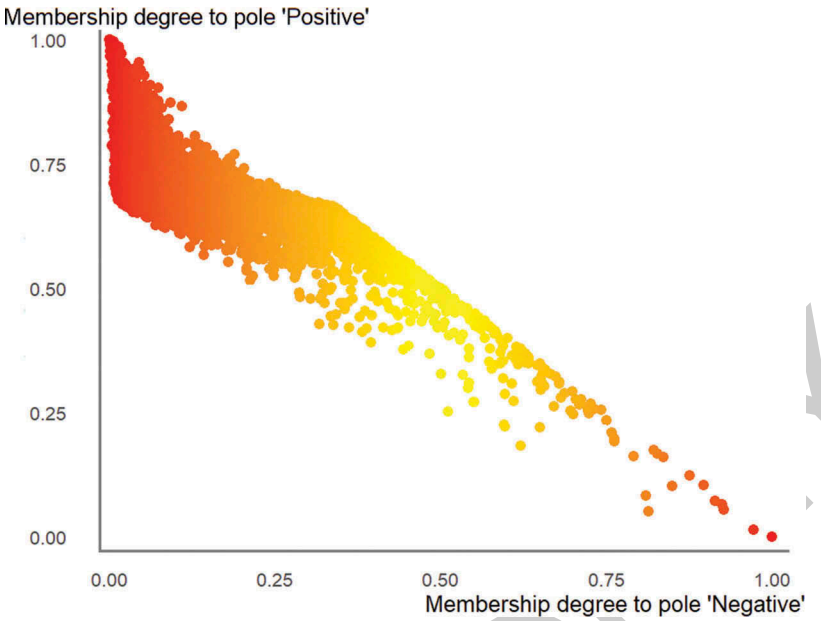

Figure 6. Scatter plot of μ_N y μ_P.

We then proceeded to calculate the polarization per topic. As can be observed in Figure 7, the highest polarization values are found in the topic *activism*, with a value of $JDJ_A = 0.454$, while the lowest values are found in the topic *voting*, with $JDJ_V = 0.169$. As regards the errors shown in the graph, the highest standard deviations are found in the topics *candidates* and *opposition*.

5. Conclusion

The method developed here allows us to offer some relevant conclusions. Firstly, we have identified a set of topics that are central to understanding the online political debate in Spain. The topics that occupy the most debate are Candidates ($n_1 = 17170$) and Opposition ($n_3 = 15327$) and, the least recurring, General debate ($n_4 = 488$) and Activism ($n_4 = 263$). The emotional tendency is negative. However, this negativity is not equally distributed among the topics that define the debate agenda. On the contrary, we identified a topic, Activism, that becomes a hate driver, while the rest of the topics present a moderate, if not positive, emotional valence. This positivity focuses, especially, on Electoral Participation. When we analyze polarization, in general, and affective polarization, in particular, this trend is confirmed. Beyond, the high degree of polarization in the Spanish elections, the topic Activism is the center of affective polarization. Meanwhile, Vote and Electoral Participation are the topics that generate the maximum consensus or a lower degree of polarization. This is the first substantive finding that is the result of

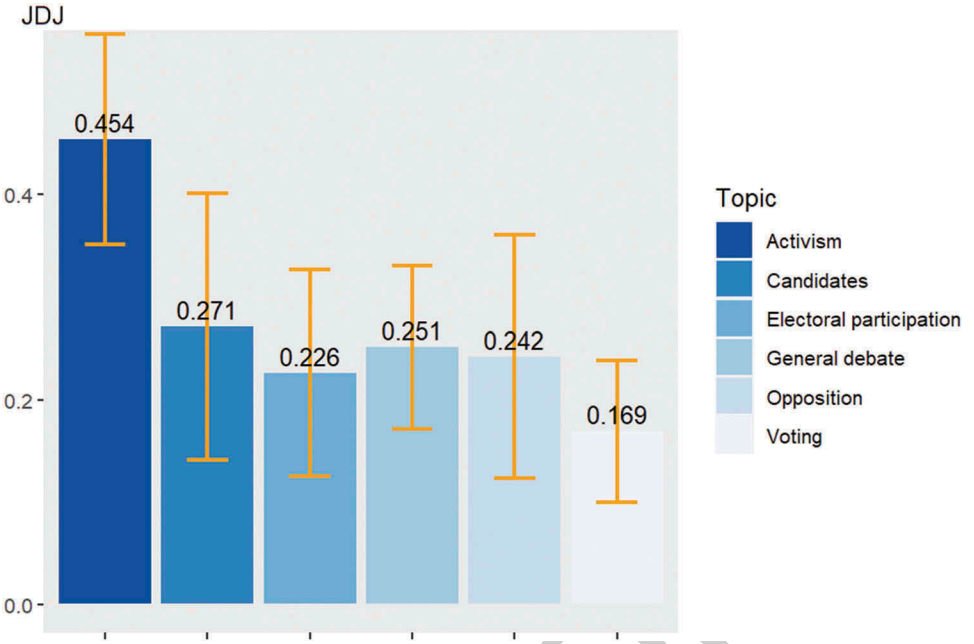

Figure 7. Values of *JDJ* for each topic.

the applied techniques: the dichotomy between conventional political participation and non-conventional political participation. Conventional political participation, such as citizen interest in electoral processes or voting, has, at least in Spain, a great reputation because of the historical cost of the transition to democracy in the 1980s. However, and especially after the events of the collective action process of 15-M, the Catalan demands, or the recent revolts against the amnesty of leaders of the "Procés Català,"[1] activism is pointed out as a strategy for questioning the constitutional order. Thanks to our empirical analysis, this process typical of the Spanish political context is identified as central and, most importantly, analyzed with valences polarized around positive and negative. Finally, we believe it is important to point out two issues. Firstly, in line with other processes around the world, we found in our case study a high incidence of affective polarization focused on issues (Lee et al., 2014). Secondly, given the difficulty and complexity of analyzing electoral processes, the basic interpretative keys are at the local level. To go beyond general trends (affective polarization by topics) it is necessary to include the specific keys and trends of the historical culture of the country being analyzed.

[1]The "15-M," in clear alignment with the Tahrir Square protests in Egypt or Occupy Wall Street in New York, was a central collective action process for public opinion and Spanish society against the cuts due to the economic crisis of 2011. In the cycle of protests in Spain, another key point is the Catalonian mobilizations that started in October 1, 2017 to express the request for independence from the Spanish state. Finally, in the first weeks of November 2023, important mobilizations took place in the streets of many Spanish cities (especially Madrid) against the decision of the current government to offer an amnesty to the Catalan independence leaders.

Regarding the methodological limitations, model bias and data representativeness are significant concerns, as pre-trained models and the nature of CrowdTangle's data can skew results. These models, trained on extensive datasets, may carry inherent biases that can influence sentiment and topic analysis, reflecting the model's training data more than the actual content analyzed. Additionally, the reliance on data from CrowdTangle, which predominantly includes verified profiles and public groups, limits the generalizability of the findings to a broader social media audience.

Disclosure statement

Q7 No potential conflict of interest was reported by the author(s).

Funding

Authors acknowledge the financial support of MINECO [PID2019-106254RB-100] (Duration: 2020-2024) and [PID2021-122905NB-C21].

References

Abdullah, M. L., Abdullah, W. S. W., and Tap, A. O. M. (2004). Fuzzy sets in the social sciences: an overview of related researches. *Jurnal Teknologi*, *41*(1): 43–54.

Acheampong, F. A., Nunoo-Mensah, H., and Chen, W. (2021). Transformer models for text-based emotion detection: a review of Bert-based approaches. *Artificial Intelligence Review*, *54*(8), 5789–5829. doi: 10.1007/s10462-021-09958-2

Bustince, H., Fernandez, J., Mesiar, R., Montero, J., and Orduna, R. (2010). Overlap functions. *Nonlinear Analysis, Theory, Methods & Applications*, *72*(34): 1488–1499. doi: 10.1016/j.na.2009.08.033

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: pre-training of deep
Q8 bidirectional transformers for language understanding. *arXiv Preprint arXiv: 181004805*.

Esteban, J.-M. and Ray, D. (1994). On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, *62*(4): 819–851. doi: 10.2307/2951734

Garrido, A., Rodríguez, M. A. M., and Rodríguez, A. M. (2021). Polarización afectiva en España. *Más Poder Local*, *45*: 21–40.

Grootendorst, M. (2022). Bertopic: neural topic modeling with a class-based tf-idf procedure.
Q9 *arXiv Preprint arXiv: 2203.05794*.

Guevara, J. A., Gómez, D., Castro, J., Gutiérrez, I., and Robles, J. M. (2022). A new approach to polarization modeling using markov chains, In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Cham: Springer International Publishing,
Q10 151–162.

Guevara, J. A., Gómez, D., Robles, J. M., and Montero, J. (2020). Measuring polarization: A fuzzy set theoretical approach, In *Information Processing and Management of Uncertainty*
Q11 *in Knowledge-Based Systems*. Cham: Springer International Publishing, 510–522.

Gutiérrez, I., Guevara, J. A., Gómez, D., Castro, J., and Espínola, R. (2021). Community detection problem based on polarization measures: an application to twitter: the COVID-19 case in Spain. *Mathematics*, *9*(4): 443. doi: 10.3390/math9040443

Huang, C., Trabelsi, A., and Zaïane, O. R. (2019). Ana at semeval-2019 task 3: contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv Preprint arXiv: 1904.00132*.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, *22*(1): 129–146. doi: 10.1146/annurev-polisci-051117-073034

Lee, J. K., Choi, J., Kim, C., and Kim, Y. (2014). Social media, network heterogeneity, and opinion polarization. *Journal of Communication*, *64*(4): 702–722. doi: 10.1111/jcom.12077

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: a robustly optimized Bert pretraining approach. *arXiv Preprint arXiv: 1907.11692*.

Montalvo, J. G. and Reynal-Querol, M. (2005). Ethnic polarization, potential conflict, and civil wars. *The American Economic Review*, *95*(3): 796–816. doi: 10.1257/0002828054201468

Montero, J., Gómez, D., López, V., Rodríguez, S., and Vitoriano, B. (2010). Sobre las funciones y reglas de agregación. In *XV Congreso Español Sobre Tecnologías y Lógica Fuzzy*. Universidad de Huelva, Servicio de Publicaciones. 303–308.

Nguyen, D. Q., Vu, T., and Nguyen, A. T. (2020). Bertweet: a pre-trained language model for English tweets. *arXiv Preprint arXiv: 2005.10200*.

Papacharissi, Z. (2015). *Affective Publics: Sentiment, Technology, and Politics*. Oxford University Press.

Pérez, J. M., Furman, D. A., Alemany, L. A., and Luque, F. (2021). Robertuito: a pre-trained language model for social media text in Spanish. *arXiv Preprint arXiv: 2111.09453*.

Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alemany, L. A., and Martínez, M. V. (2023). Pysentimiento: a python toolkit for opinion mining and social nlp tasks. *arXiv Preprint arXiv: 2106.09462*.

Ragin, C. C. and Pennings, P. (2005). Fuzzy sets and social research. *Sociological Methods & Research*, *33*(4), 423–430. doi: 10.1177/0049124105274499

Reimers, N. and Gurevych, I. (2019). Sentence-Bert: sentence embeddings using siamese Bert-networks. *arXiv Preprint arXiv: 1908.10084*.

Robles, J.-M., Guevara, J.-A., Casas-Mas, B., and Go´mez, D. (2022). Cuando la negatividad es el combustible. bots y polarización política en el debate sobre el covid-19. *Comunicar Revista Científica de Comunicación y Educación*, *30*(71): 63–75. doi: 10.3916/C71-2022-05

Shore, J., Baek, J., and Dellarocas, C. (2016). Network structure and patterns of information diversity on twitter. *arXiv Preprint arXiv: 1607.06795*.

Simón de Blas, C., Guevara, J. A., Morillo, J., and Gómez González, D. (2022). Polarization measures in bi-partition networks based on fuzzy graphs. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* 398–409: Springer.

Sobieraj, S. and Berry, J. M. (2011). From incivility to outrage: political discourse in blogs, talk radio, and cable news. *Political Communication*, *28*(1): 19–41. doi: 10.1080/10584609.2010.542360

Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, *60*(3): 556–576. doi: 10.1111/j.1460-2466.2010.01497.x

Sunstein, C. (2018). *# Republic: Divided Democracy in the Age of Social Media*. Princeton university press.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., and Nyhan, B. 2018. Social media, political polarization, and political disinformation: a review of the scientific literature. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3144139

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wolfson, M. C. (1994). When inequalities diverge. *The American Economic Review*, *84*(2): 353–358.

Zadeh, L. A. (1965). Fuzzy sets. *Information & Control*, *8*(3): 338–353. doi: 10.1016/S0019-9958(65)90241-X

Q24