Muñoz et al. *EPJ Data Science* (2024) 13:39
https://doi.org/10.1140/epjds/s13688-024-00480-3

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Quantifying polarization in online political discourse

Pau Muñoz[1], Alejandro Bellogín[1*], Raúl Barba-Rojas[2] and Fernando Díez[1]

*Correspondence:
alejandro.bellogin@uam.es
[1]Universidad Autónoma de Madrid, Madrid, Spain
Full list of author information is available at the end of the article

Abstract

In an era of increasing political polarization, its analysis becomes crucial for the understanding of democratic dynamics. This paper presents a comprehensive research on measuring political polarization on X (Twitter) during election cycles in Spain, from 2011 to 2019. A wide comparative analysis is performed on algorithms used to identify and measure polarization or controversy on microblogging platforms. This analysis is specifically tailored towards publications made by official political party accounts during pre-campaign, campaign, election day, and the week post-election. Guided by the findings of this comparative evaluation, we propose a novel algorithm better suited to capture polarization in the context of political events, which is validated with real data. As a consequence, our research contributes a significant advancement in the field of political science, social network analysis, and overall computational social science, by providing a realistic method to capture polarization from online political discourse.

Keywords: Polarization; Political discourse; Information theory

1 Introduction

Political polarization is a significant phenomenon in today's interconnected digital age. Defined by deep-rooted ideological divides and emotionally charged beliefs, it finds fertile ground in platforms like X, previously called Twitter (from now on, we refer to this social network as X). These platforms, serving as dominant channels for political discourse, reflect societal sentiments and magnify and distort them in many instances [1]. Challenges like the rapid spread of misinformation, the creation of echo chambers, and algorithm-driven content recommendations further compound the problem [2]. However, it is essential to recognize these platforms have a dual nature, given their vast user base and real-time data processing capabilities. While they can amplify divides, they also have the unparalleled potential to bridge gaps, encourage diverse dialogues, and shape global political sentiment [3].

These social media networks are steadily supplanting traditional media outlets as primary sources of information for many individuals. Whereas traditional media often operates on scheduled timelines and involves editorial oversight, social media provides real-time updates and democratizes content creation. Thus, this defining characteristic of

Springer

modern social media platforms empowers users to produce and share their content, moving beyond the confines of traditional, established media [4]. This democratization of information dissemination has transformed the media landscape. No longer are narratives solely shaped by politicians, journalists, and media houses; anyone with internet access can now contribute to the global conversation. This shift has led to a richer tapestry of voices and perspectives, fostering more vibrant discussions and debates. It is a double-edged sword, in any case. While this inclusivity encourages diverse participation [5], it also opens the door to misinformation and an increasingly polarized discourse [6]. Nevertheless, the ability for individuals to publish and circulate their content underscores a monumental shift in the dynamics of information exchange in the digital age [4].

In contemporary times, political polarization is not only persisting but is, alarmingly, on an upward trajectory [7]. Instead of evolving towards more centrist or unified perspectives, societies worldwide are witnessing a sharpening of ideological divides [8]. Factors such as the rise of populism [9], the influence of certain media outlets [10, 11], and the advent of algorithm-driven social media platforms have contributed to this heightened division [12, 13]. Individuals find themselves in ideological silos, often reinforced by selective exposure to information that aligns with their pre-existing beliefs [14]. This increasing chasm between opposing viewpoints can stifle constructive dialogue, lead to political stagnation, and exacerbate societal tensions. Addressing and reversing this trend is essential for sustaining healthy democracies and cohesive societies [15]. One of our main hypotheses is that electoral campaigns offer an ideal setting to analyze political polarization on social media. During these periods, parties and candidates ramp up their outreach to influence and galvanize followers [16]. The surge in news, commentary, and targeted messaging, combined with the platforms' dynamics, magnifies existing polarization, making campaigns a focal point for studying such divides [17].

However, there is not a single definition of polarization, since it depends on the political context and other variables. For example, affective polarization refers to how citizens feel sympathy towards partisan in-groups and antagonism towards partisan out-groups [18]. In multiparty systems, capturing the affect pattern towards multiple parties is more complex compared to two-party systems [18]. This conceptualization and measure of affective polarization in multiparty systems summarize the configuration of feelings towards political parties and their supporters [18]. Other definitions of political polarization focus on the underlying ideological divisions between voters. Capturing the summary ideological divisions between citizens, considering how the policy content of debates shifts [19]. They emphasize the shifting meaning of left and right and the efforts of "issue entrepreneurs" in shaping political polarization [19]. Another definition of political polarization in online social networks emphasizes ideological homophily, which refers to the tendency of individuals to connect and interact with others who share similar political beliefs [1, 20–22]. This definition underscores the role of social network dynamics and the formation of homogenous and segregated ideological groups in contributing to polarization [1, 20–22].

In fact, due to these multiple definitions, the current state-of-the-art algorithms employed to gauge polarization tend to hinge on three primary approaches: exploring network topology, content analysis of posts, or applying hybrid methods combining both. These techniques, equipped with their strengths, capture distinct facets of the polarization phenomenon. For instance, studying network topology can reveal clusters or echo chambers, while content analysis might shed light on the nature and intensity of polarized

rhetoric. However, these algorithms often fail to provide a comprehensive view despite their varying sophistication. One of their main challenges is the need for more adaptation to the particular political landscapes in which they are applied. A one-size-fits-all approach is less effective, especially in multi-party systems, where political dynamics can be more intricate and varied [18]. This emphasizes the necessity for developing algorithms and methodologies that are attuned to specific political contexts and can navigate the complexities inherent in different political systems.

Moreover, the vast majority of controversy/polarization detection algorithms available to date rely on the use of lattice structures that store information related to a specific moment. As described in Sect. 2, most of these algorithms can fall short because they are typically snapshot-based, capturing a single moment in time rather than the ongoing, dynamic processes that characterize real-world phenomena [1]. Such static approaches can miss the nuanced, evolving nature of social, political, or ideological divides, which are influenced by emerging events, shifting public opinions, and the complex interplay of various factors over time. In reality, polarization is not a fixed state but a fluid condition that can intensify, diminish, or change in nature in response to new information, societal changes, or interventions. By relying on snapshots, algorithms fail to account for these dynamics, potentially leading to oversimplified analyses that do not accurately reflect the complexities of real-world polarization [1]. This limitation underscores the need for more sophisticated, dynamic models that can capture the temporal aspects of polarization, offering a more accurate and insightful understanding of its causes, consequences, and potential remedies.

Furthermore, many of the proposed algorithms are limited to studying controversy between two groups or communities, precluding the possibility of considering more communities [18]. Similarly, these algorithms do not consider coherence in discourse. Polarization will be pronounced when each network maintains its own narrative, especially if it is negative towards the other network. Polarization will not be as pronounced when, despite the presence of well-defined communities, there exists a plurality or richness in the discourse within each one of them [23]. Hence, while current methodologies offer insights into the phenomenon of polarization, they capture only a fragment of its intricate reality [1, 18].

In light of these observations, our research undertakes several critical tasks. Firstly, we analyzed the state of the art, focusing explicitly on the evolving definitions and nuances of political polarization within the unique ecosystem of online social networks, and implemented the most significant and representative approaches. Building upon this foundation, we shift our lens to a detailed exploration and comparison of existing algorithms, delving into their methodologies, strengths, and weaknesses. Such a comparative study, while highlighting the current landscape, also uncovers gaps and opportunities for innovation. To this end, our research culminates in the proposal of *SPIN* (*Social-political Polarization analysis by INformation theory*), a novel algorithm rooted in the principles of information theory. This algorithm aims not just to measure but to shed light on the underlying complexities of political polarization in the digital realm, offering scholars and practitioners a new tool to understand and navigate this pressing issue. This benchmarking effort takes advantage of a novel dataset collected also as part of our contribution, where political discourse in Spain was gathered around national and local elections in the period of 2011-2019.

2 Background

2.1 Polarization dynamics

Understanding the dynamics of polarization during electoral processes is crucial for comprehending the broader implications on democratic health and political engagement. The electoral cycle, characterized by distinct phases from the pre-campaign period to post-election reflection, offers a unique lens through which to examine these phenomena.

In the initial phase of the electoral cycle, the pre-campaign stage, polarization begins to steadily increase [24]. This period is marked by the articulation of campaign agendas, the crystallization of party platforms, and the beginning of targeted outreach to potential voters. Political actors leverage these activities to delineate ideological boundaries, often amplifying differences to galvanize support and distinguish themselves from opponents. This strategic emphasis on differentiation contributes to the gradual intensification of polarization [25], as voters become more entrenched in their affiliations and perceptions of the political landscape. In reality, the pre-campaign phase could refer to many months, however, we are more interested in the last pre-campaign days, as they are closer to the electoral process to be analyzed.[1]

As the campaign progresses[2] towards the blackout period – a legally mandated cessation of campaign activities immediately preceding the election – a notable shift occurs. During this blackout period, polarization experiences a temporary decline, reaching what can be described as local minimum levels [26]. The absence of active campaigning reduces the immediate influx of polarizing messages, allowing voters a moment of respite from the heightened rhetoric [26]. This pause in the electoral fervor is thought to foster a more reflective environment, enabling individuals to consider their choices with less external pressure, potentially mitigating the sharpness of polarization felt during the peak campaign period.

However, the blackout period is short-lived. On election day and the following day, polarization surges, reaching again local maximum levels [27, 28]. This spike can be attributed to the culmination of campaign tensions and the immediate reactions to electoral outcomes. The realization of victory or defeat accentuates existing divisions, as stakeholders process the implications of the election results. Emotional investment in the political process and the stark contrast between winning and losing sides exacerbate feelings of division, driving polarization to its zenith.

In the aftermath of the election, a gradual decrease in polarization is observed, with levels eventually falling below those noted during the campaign and pre-campaign periods [29]. This post-election phase often ushers in calls for national unity and a collective focus on governance over campaigning, contributing to a de-escalation of polarized rhetoric. As the immediacy of the electoral contest fades, so too does the intensity of polarization, suggesting a return to a more moderate political discourse until the cycle inevitably renews.

Each of these phases underscores the fluid nature of polarization within the electoral context, highlighting the influence of campaign dynamics, legal frameworks, and collective psychological responses to political events. A nuanced understanding of these patterns is essential for developing strategies to mitigate excessive polarization, ensuring that

[1] The same principle applies for post-campaign.

[2] In Spain, the campaign lasts for 15 days and ends in the blackout period.

electoral processes contribute to the strengthening of democratic principles rather than their erosion.

Considering these observations, we identified several aspects that a polarization detection algorithm should be able to capture. These aspects define the benchmark that we use for validating our proposed algorithm, SPIN, and are agreed in the scientific community [24, 29–31], as it was previously discussed:

- Polarization increases steadily from the pre-campaign stage until the day(s) corresponding to the blackout period.
- During the blackout period, it drops, reaching local minimum levels.
- On election day and the day after, it rises, reaching a global maximum.
- Subsequently, it progressively decreases after the election, reaching lower levels than during the campaign and pre-campaign periods.

2.2 Algorithms for measuring polarization

To establish a fair comparison and to perform a deeper benchmark of our solution, we explored and implemented some of the well-known polarization detection algorithms in the literature. Indeed, the previous claims about electoral processes are considered as a ground truth for this benchmark, which allowed us to analyze the results of our proposed solution.

Measuring polarization, also called controversy by some authors [32], is a complex task, primarily because the term can be interpreted in various ways. As we have discussed previously, depending on context and perspective, definitions of polarization may differ, leading to inconsistencies in data and analysis. Accurately quantifying and comparing polarization across different mediums or regions with a universally accepted metric or definition becomes easier. Nevertheless, most existing measuring algorithms agree on a set of basic terms.

However, most of these algorithms capture the phenomenon only partially as most of them only rely on the structure (or the content) of social networks to provide a measurement of polarization, as hybrid algorithms are not as common as topology-based or content-based algorithms [33]. Also, most of the existing algorithms focus on general domain polarization, instead of exploring this phenomenon in the political context [34].

Indeed, while polarization is often associated with a two-party system, where divides appear distinctly binary, it is not exclusive to such structures. Despite having a broader array of political stances and entities, multi-party systems can also exhibit pronounced polarization [24]. Fragmentation can occur among various parties or ideological groups in these systems, leading to multiple echo chambers. Each group can become insular, intensifying its internal consensus while growing increasingly distant from or antagonistic toward other groups. Hence, polarization is not inherently bipartite or bipolar; it can manifest differently across different political landscapes, underscoring its complexity and the need for nuanced approaches in its study and mitigation.

In the subsequent sections, we present a concise overview of the state-of-the-art algorithms to measure polarization. These topology-based, content-based, and hybrid algorithms offer insights into various facets of the complex landscape of digital polarization. After discussing their dynamics, strengths, and potential limitations, in the rest of the paper we introduce our method based on Information Theory. Drawing from the lessons of existing tools and addressing their gaps, this proposed approach aims to provide a more

comprehensive and nuanced understanding of polarization dynamics in digital spaces. See Additional file 1 for more information about these algorithms.

2.2.1 Topology-based algorithms

Topology-based algorithms hinge primarily on the network structure to discern polarization patterns. By analyzing user connections and interactions, these algorithms map out the network's layout to identify clusters, bridges, or isolated nodes. Such clusters or echo chambers represent groups of like-minded individuals who frequently interact with one another, often reinforcing shared beliefs.

In these topology-based networks, individual users are denoted as nodes, with their interactions. Considering X,[3] user-to-user interactions such as retweets, likes, replies, and quotes can be used to represent the links or edges that connect them. For instance, if user A retweets user B (to provide an example within the X platform), this establishes a directional link from A to B. Similarly, a 'like' or 'reply' would form another type of connection. As more interactions accumulate, a more intricate web of connections emerges, vividly illustrating the flow of information and the nature of interactions among users. Over time, distinct clusters or communities become apparent, often signifying groups with shared beliefs or interests. Topology-based algorithms can infer the degree and patterns of polarization within the network by analyzing the pattern and density of these links. The granularity of these interactions provides a detailed map of the digital landscape, indicating areas of consensus, contention, and isolation.

In the literature, there are plenty of algorithms that are based on these ideas. One of the most basic algorithms that has served as a basis for other works is the Random Walk Controversy (RWC) [32]. This algorithm carries out polarization detection in a network that is divided into two groups of nodes. In the political scenario of the United States, this algorithm would be useful as it could be used to measure network polarization between Republicans and Democrats. The intuition behind the algorithm is simple: if we perform random walks from a random starting point in the network, then if the probability of beginning in a given group clearly affects the probability of reaching the other group, one could determine the polarization of the network. For instance, let us assume that after N random walks starting in a node with Republican political leaning (which could be different from random walk to random walk), then if a big number of these walks ended in another node with the same political leaning, one could argue that it is more likely to begin and end in the same partition (*community*) than it is to go from one partition (*community*) to the other, leading to a higher network polarization, as perceived from its structure. Some other algorithms extend this basic intuition to improve the accuracy of the polarization index, as is the case with Authoritative Random Walk Controversy (ARWC) [35] and Displacement Random Walk Controversy (DRWC) [35]. ARWC follows the exact same rules as RWC, however, a random walk is set to end whenever the walk reaches an influential node of any of the two partitions. Node influence can then be measured using network metrics such as the PageRank [36]. DRWC is based on a similar intuition. It considers random walks of a fixed length and it is focused on the number of community changes per random walk rather than the actual starting and ending points of such a walk. Followings its intuition, a higher average number of community changes per walk

[3] Any other social network could have been used in its place.

indicates a lower network polarization, however, a lower average number of community changes indicates the presence of polarization in the given network..

Some other algorithms that we can find in the literature are also topology-based, however, they are based on different ideas. For example, in Ref. [32] the authors discuss the Betweenness Centrality Controversy (BCC), an algorithm capable of measuring polarization in a social network focused on its connections. Its intuition is simple: if the network is polarized, the edges connecting nodes from one partition to another should have a very high betweenness centrality (since many of the shortest paths from a node in partition X to another node in partition Y will necessarily pass through them), as opposed to the low betweenness centrality of edges connecting nodes from the same partition. In a different work [32], authors discuss a polarization index, Embedding Controversy (EC), built upon the assumption that polarized networks have a high modularity. Thus, this algorithm uses layout algorithms that maximize modularity, such as Force Atlas 2 [37], in order to calculate an embedded representation of each node (in a bi-dimensional representation space), that is then used to compute distances between the nodes, so as to give a polarization measurement based on average distances between nodes of the same and different partitions. Indeed, a higher distance between nodes of different partitions, together with a smaller distance between nodes of the same partition gives the intuition of the network polarization that the index is capable of measuring.

Following other approaches, authors in Ref. [38] propose a polarization measurement algorithm built upon well-known physics concepts. Its intuition is straightforward: if a network is polarized, then the users of each community should tend to be positioned close to the extremes (as they "repel" each other), as opposed to a non-polarized network where the users should tend to be positioned closer to the center of the network. In Ref. [32], authors discuss the Boundary Connectivity controversy index, built upon the concepts of boundary and internal nodes. A boundary node is, in essence, a node that is connected to at least one node of the opposite community (the algorithm supports two communities only), whereas internal nodes are restricted to be connected only to nodes of the same partition [32]. Then, the intuition behind this algorithm is simple, as it is based on the fact that, when the network is polarized, its boundary nodes should have more connections to their corresponding internal nodes and fewer connections with the boundary nodes of the opposite partition. Then, the number of edges from boundary nodes to both internal and boundary nodes underscores the presence of polarization in the network in this algorithm.

As it can be seen, the topology's structure is insightful, working under the assumption that highly polarized networks exhibit fewer interconnections between differing groups and denser internal connections within like-minded clusters. This pattern reflects the echo chamber effect, where users mainly interact with those sharing similar beliefs, creating clear divisions within the more extensive network. However, all the previous algorithms have two important restrictions.

On the one hand, most of these algorithms only support social networks divided into two groups or communities, which may not necessarily cover the needs of those cases in which polarization must be studied between more than two groups, as is the case with the Spanish political system.[4] Algorithms for detecting and quantifying polarization for social

[4]The Spanish political system involves several (not necessarily two) parties and it is common to other western democracies.

networks divided into two or more groups do exist, but they are more scarce in the literature [18]. One algorithm that carries out polarization detection considering more than two groups is ERIS [39]. Such an algorithm is based on the creation of two matrices that provide an intuition of two concepts that the authors link to polarization: antagonism (how opposite one community is as perceived from another community - in pairwise fashion) and porosity (how frequent information flows happen between each pair of communities).

On the other hand, there is a second restriction to many of the algorithms that are present in the literature, just as the ones described above. This second restriction is the fact that most of these algorithms work with snapshots of a social network in a specific moment [32, 35, 39], without considering the dynamics of the information that flows from one user to another (i.e., which user introduced new information, which users amplified it, ...).

For the case of our proposed algorithm, SPIN, the dynamics of the information flow are considered and, furthermore, the algorithm supports multiple communities. Thus, it is, to our knowledge, the first algorithm in the literature to bring together all these aspects (including some others, such as its hybrid nature) to try to improve existing polarization detection and quantification algorithms.

2.2.2 Content-based algorithms

While topology-based algorithms provide insight into the structural patterns of networks, they overlook the content of posts, a vital component in assessing polarization. The very essence of a post, especially when laden with negative sentiments or direct opposition to contrasting views, amplifies polarization. It is not just about how clustered or isolated networks are, but also about the intensity of sentiments within those clusters. Real polarization is underscored when isolated networks echo shared beliefs and hostility towards differing perspectives. Content-based algorithms focus on analyzing the text within posts to gauge polarization. By examining language use, sentiment, and thematic content, these algorithms can discern the tone, intensity, and nature of the discussions, allowing for a deeper understanding of underlying beliefs and attitudes. Such algorithms can detect patterns of extreme views, recurring divisive topics, and the frequency of negative or adversarial language, offering a nuanced picture of polarization beyond mere network structures.

In the literature, we can find many polarization detection algorithms based on different techniques. Some of them are purely based on Natural Language Processing (NLP), such as [40–43]. All these algorithms apply NLP concepts to extract information from the content of the posts to obtain a measurement of polarization. However, such a measurement does not consider important aspects such as the topology of the network or the temporality of the posts (information dynamics), so they are somehow limited (just as topology-based algorithms do not consider the content of the posts, missing an important piece of information for measuring polarization).

Furthermore, some of the content-based algorithms that we can find in the literature are heavily based on Deep Learning techniques. An example of these algorithms can be found in Ref. [44], where the authors propose a Deep Learning based approach to carry out ideology detection and polarization detection using the sentiment analysis from tweets, in the context of the COVID-19 pandemic. However, it is possible to find other solutions to the problem of polarization detection and quantification using content-based algorithms.

In Ref. [45], the authors propose an algorithm based on the application of NLP to separate positive, neutral, and negative posts, to then formulate an index to measure polarization.

These methods primarily offer insights into the overarching sentiment or the general polarization levels related to specific topics, rather than the interactions and divisions between distinct user groups or communities. As a result, they may miss the subtleties of how polarization manifests and propagates across different segments of the network. In fact, these algorithms may overlook inter-community polarization dynamics, as they do not take into consideration the user nor the relationships between them. Moreover, since they tend to lean heavily on Machine Learning and Deep Learning techniques for classifying and understanding labeled posts, this brings inherent challenges. On the one hand, there is a consistent need to train and retrain these models to maintain their accuracy. Additionally, they often operate as "black boxes", making it challenging to discern how they arrive at specific classifications. This lack of transparency, known as the *explainability problem*, can hinder the broader acceptance and trust in these algorithms, especially in contexts where understanding the reasoning behind classifications is crucial [46]. These complexities, combined with their need for continuous retraining and their limited explainability, have made them more challenging to deploy effectively. These hurdles can impede widespread adoption, especially in contexts where stakeholders value transparency, understandability, and adaptability in the tools they utilize.

2.2.3 Hybrid algorithms

Hybrid, or mixed algorithms, meld the strengths of both approaches: they incorporate the structural insights derived from network topology with the nuanced content analysis of posts. By doing so, they aim to provide a more holistic view of polarization, capturing both the overarching patterns of connectivity and the underlying sentiments and discourses prevalent within the network. This integration allows for a more comprehensive understanding of the multi-dimensional facets of polarization in digital spaces.

Although these techniques are scarce, there are useful polarization detection and quantification algorithms in the literature. For instance, Biased Random Walk (BRW) [33], is a random walk-based approach that introduces content-based components to improve the polarization metric. As a result, in such an algorithm the random walk starts with an "initial energy" that is consumed as the random walk traverses the social network (each user has an "energy loss" and content-based strategies can be used for their computation). Indeed, this algorithm combines the advantages of both approaches to detect and quantize polarization intelligently. Another example of hybrid algorithm is Diffpool [47], a hybrid approach for polarization detection based on Deep Learning. More specifically, this approach is capable of representing a graph through graph convolutional neural networks, as well as content-based information through embeddings. This network is coarsened thanks to the action of pooling layers, to finally provide a measurement of polarization. However, the same weaknesses described in Sect. 2.2.2 still apply to this algorithm, as it is based on the same techniques. Indeed, algorithm explainability and training efficiency become important consideration aspects to carry out polarization detection with this algorithm. Another hybrid algorithm that we can find in the literature is the Multi-Opinion based method for controversy detection [48], which partitions a network in a given number of communities (with algorithms such as METIS and Louvain partitioning [49, 50]). This algorithm does in fact support multiple communities and not only two, and it considers

both the topology and the content of a network, which are aspects that also characterize our proposed polarization algorithm, SPIN. In the same line of proposing hybrid algorithms, authors in Ref. [51] introduce the Generalized Euclidean (GE) algorithm, based on a generalization of the Euclidean Distance metric to measure the distance between nodes and offer a polarization index based on such distance (for users belonging to only two communities, thus this algorithm is also limited as some of the other aforementioned algorithms).

As it can be seen, a notable limitation of current hybrid (and non-hybrid) algorithms is their lack of consideration for temporal patterns. This means they might overlook the evolution of discussions, sentiments, and network structures over time. Understanding how and when polarization intensifies, ebbs, or shifts, especially in response to real-world events or online triggers, is crucial. Without this temporal dimension, we miss out on the dynamics of polarization, potentially leading to static or outdated interpretations of the digital landscape. It is this limitation that made us propose SPIN, a hybrid algorithm that uses the structure of the network, together with the information that flows through it (and its dynamics), to provide a measurement of its polarization. This proposed algorithm was designed to overcome some of the difficulties and limitations of existing algorithms, including the use of both structure and content to provide the measurement, the use of information dynamics and the support of multiple communities to provide the measurement of polarization.[5]

3 Data and methods

3.1 Data collection

For a rigorous and meaningful assessment of our research, it is imperative to use datasets encapsulating political discussions. Such datasets, rich in political discourse's nuances, complexities, and polarities, provide an ideal testing ground. By focusing on these politically charged conversations, we can later focus on the ability of the polarization detection algorithm to detect and effectively measure polarization, and its efficacy in parsing and understanding the varied facets of political dialogue. This context-specific evaluation ensures that such algorithms are robust and well-suited for the challenges of the real-world political landscape.

In light of this, we have compiled a dataset centered on Spanish electoral processes, utilizing data from X (Twitter) to chronicle these unique phases and the polarization dynamics they encompass. The dataset was generated starting from the main accounts of the parties that secured representation in each general electoral process. We then applied snowball sampling with three levels of depth. This involved collecting all the posts from the recursively identified accounts during 7 days of pre-campaign, 15 days of campaign,[6] both the reflection and election days, and the subsequent 7 days of post-campaign. It is worth noting that, although both the pre-campaign and the post-campaign can be extended in time, only the week closest to the election date was considered in both cases, so as to try to explain how polarization evolves during the electoral process, while keeping a resonable time window (of 31 days) to study per election. See Algorithm 1 for a pseu-

[5] This last condition is fundamental to create a polarization algorithm designed to study the polarization in the Spanish (and other similar Western democracies) electoral system.

[6] In the Spanish electoral process context, the campaign lasts for exactly 15 days.

Algorithm 1: Snowball Political Dataset Generation from X/Twitter

Data: X accounts of main political organizations: *accounts*
Election day: *election_day*
Number of publications: 8
Number of recursions: 2
Result: List of users influenced by political publications

common_list ← **empty list**;
start_date ← *election_day* – 3 **weeks**;
end_date ← *election_day* + 1 **week**;
for *account* ***in*** *accounts* **do**
  publications ← **download all publications between** *start_date* **and** *end_date* **from** *account*;
  for *i* ← 1 ***to*** *number of publications* **do**
    publication ← **randomly select from** *publications* **with at least one repost**;
    reposting_user ← **randomly select one user who reposted** *publication*;
    common_list.append(*reposting_user*);
  end
end
iteration ← 0;
while *iteration* < *number of recursions* **do**
  newly_retrieved ← **copy of** *common_list*;
  for *user* ***in*** *newly_retrieved* **do**
    for *i* ← 1 ***to*** *number of publications* **do**
      publication ← **randomly select from user's publications with at least one repost**;
      reposting_user ← **randomly select one user who reposted** *publication*;
      common_list.append(*reposting_user*);
    end
  end
  iteration ← *iteration* + 1;
end
return *common_list*;

Table 1 Datasets of Spanish general electoral processes from 2011 to 2019

Electoral Process	November 2019	April 2019	2016	2015	2011
Users	873	715	759	688	611
Posts	527,093	423,638	616,427	465,967	317,506
Negative Posts	118,951	83,708	92,881	74,338	55,754
Nodes	872	700	756	614	572
Edges	24,460	18,524	23,371	12,420	9,369
Degree	56.10	52.93	61.83	67.29	32.76
Avg Retweets	361.11	247.64	74.19	45.14	16.09
Avg replies	3.77	2.49	0.66	0.73	0.80
Avg likes	38.79	26.42	5.39	3.24	1.15
Avg quotes	1.29	0.85	0.19	0.09	0.00

docode of this process, and Tables 1, 2, and 3 for a dataset description considering both the overall electoral process, but also the process divided in the aforementioned phases.

In our data collection algorithm, it must be noted that the second part, where iteration is performed based on a number of recursions (snowball sampling), we randomly select a user post with at least one repost, but filtering its publication date between the specific start and end dates of each electoral process, thus achieving a dataset as described in Table 3. Additionally, all these tweets come from users that are involved in the context of the

Table 2 Datasets of Spanish local electoral processes from 2011 to 2019

Electoral Process	2019	2015
Users	708	722
Posts	398,689	535,306
Negative Posts	71,816	76,030
Nodes	698	719
Edges	16,550	19,397
Degree	47.42	53.96
Avg Retweets	184.03	44.59
Avg replies	1.80	0.68
Avg likes	20.64	3.58
Avg quotes	0.67	0.00

Table 3 Tweets downloaded during each phase of each electoral process studied

Election	Phase	Date Range	Posts
General (Nov) 2019	Pre-Campaign	2019-10-18 to 2019-10-24	119,233
	Campaign	2019-10-25 to 2019-11-08	282,605
	Reflection + Election	2019-11-09 to 2019-11-10	29,899
	Post-Campaign	2019-11-11 to 2019-11-17	95,356
General (Apr) 2019	Pre-Campaign	2019-04-05 to 2019-04-12	94,749
	Campaign	2019-04-12 to 2019-04-26	231,563
	Reflection + Election	2019-04-27 to 2019-04-28	23,482
	Post-Campaign	2019-04-29 to 2019-05-05	84,274
General 2016	Pre-Campaign	2016-06-03 to 2016-06-09	134,005
	Campaign	2016-06-10 to 2016-06-24	370,632
	Reflection + Election	2016-06-25 to 2016-06-26	31,243
	Post-Campaign	2016-06-27 to 2016-07-03	80,547
General 2015	Pre-Campaign	2015-11-27 to 2015-12-03	81,921
	Campaign	2015-12-04 to 2015-12-18	227,818
	Reflection + Election	2015-12-19 to 2015-12-20	93,328
	Post-Campaign	2015-12-21 to 2015-12-27	43,421
General 2011	Pre-Campaign	2011-10-28 to 2011-11-03	61,909
	Campaign	2011-11-04 to 2011-11-18	173,157
	Reflection + Election	2011-11-19 to 2011-11-20	23,027
	Post-Campaign	2011-11-21 to 2011-11-27	66,692
Local 2019	Pre-Campaign	2019-05-03 to 2019-05-09	87,262
	Campaign	2019-05-10 to 2019-05-24	233,448
	Reflection + Election	2019-05-25 to 2019-05-26	22,451
	Post-Campaign	2019-05-27 to 2019-06-02	63,730
Local 2015	Pre-Campaign	2015-05-01 to 2015-05-07	108,693
	Campaign	2015-05-08 to 2015-05-22	307,430
	Reflection + Election	2015-05-23 to 2015-05-24	30,348
	Post-Campaign	2015-05-25 to 2015-05-31	88,835

electoral process (relevant for polarization) as they were obtained from randomly selecting users who reposted content from official accounts of the main political parties in the country, thus we consider that our algorithm can effectively be used to gather quality data with which to carry out our experiments.

Last, it must be noted that, through the application of this data collection algorithm we were able to craft a dataset that is described in Tables 1, 2 and 3. The first two tables dive deep into several important aspects of the crafted dataset per electoral process, including the number of users, posts, and even additional information regarding the network that could be created through the usage of these datasets and the X interactions between users (retweets, replies, likes, and quotes). Moreover, average values for those relationships are

also provided, to contextualize the kind of usage that the social network was having during each electoral process. In this sense, we observe how retweets and likes are more common than quotes and replies for the different electoral processes studied. The main difference between Tables 1 and 2 is that the first one is only focused on Spanish general electoral processes, whereas the second table is focused on Spanish local electoral processes, from 2011 to 2019.

In Table 3, we provide information regarding the number of tweets downloaded per phase of each electoral process detected. As it was explained in Sect. 2, in Spain an electoral process tends to have different phases. This table provides relevant information for each phase, such as the number of posts, together with the date range considered, for each phase of each electoral process; thus, it allows to further understand and contextualize the previous two tables on Spanish general and local electoral processes.

3.2 SPIN algorithm

The proposed polarization metric is based on the idea that we can borrow the fundamental concepts from Information Theory to measure the flow of information between communities of different characteristics (in the context of social network polarization in politics, we refer to communities based on ideological positioning, associated with clearly identified political organizations), primarily through the concept of entropy (and the metrics, defined in the present literature, to estimate it). Our main hypothesis is that we consider a network is polarized when polarization exists:

1 *Between different communities*: The flow of (hostile) information between different communities is naturally a clear indicator of the existence of polarization in the network.
2 *Within each community*: Nevertheless, there is also the possibility that communities are highly polarized because (hostile) information (or information against other communities) circulates within each of them, without this information necessarily flowing to other communities. To model this possibility, it is essential to also consider the flow of information within each community, and not just between communities.

Given that entropy (and related measurements) indicates the flow of information between two entities, we can define an algorithm utilizing it to detect polarization. Our proposal goes as follows:

1 *Calculation of inter-community entropy*: By considering the connections between nodes (users) from different communities, through the calculation of entropy, we can quantify the amount of information that flows between communities.
2 *Calculation of intra-community entropy*: Considering those nodes (users) that are related, we can calculate the flow of information between them, estimating the amount of information in the network that flows within each specific community.
3 *Weighted contributions*: our algorithm uses the previous entropies to obtain a negativity ratio within each community (intra-community negativity ratio) and between communities (inter-community negativity ratio). A weighted average of both contributions results in the SPIN polarization index result. Indeed, for calculating this weighted average, two hyperparameters are introduced in the algorithm (α and β), which can be used for tuning the relevance of each contribution (increasing α gives more weight to the intra-community negativity

ratio, whereas increasing β gives more weight to the inter-community negativity ratio). For the benchmark of SPIN, we tested different combinations of weights, to understand the impact of each of these negativity ratios in the proposed polarization metric.

How information flow is estimated shall be described in Sect. 3.2.1, which is the basis to compute the inter- and intra-community entropies (see Additional file 2 for a detailed pseudocode of how these entropies are computed). However, these concepts alone would not allow us to calculate a proper polarization metric. For this reason, we restrict the computation of the entropy to hostile (or negative) information; this shall be described in Sect. 3.2.2. Because of this, it also requires the content of the nodes' (users') posts.

As summarized in Algorithm 2, the input network must be partitioned so that nodes with similar characteristics (ideologies, in the case of using polarization detection in social networks in politics) are located in the same partition and different partitions from nodes with different characteristics. We also devise a methodology to consider such domain knowledge, explained in Sect. 3.2.3.

In summary, the proposed polarization metric belongs to the family of hybrid algorithms, since it is based on a network structure representing the network on which polarization analysis (quantification) will be carried out (see Sect. 3.2.4 for more details about how to represent the network), the content of the nodes' posts, and the temporal moment

Algorithm 2: Calculation of the polarization metric in SPIN

Data: List of communities: *communities*
Intra-community entropy dictionary: *partition_intra*
Negative intra-community entropy dictionary: *partition_neg_intra*
Inter-community entropy matrix: *inter_matrix*
Negative inter-community entropy matrix: *neg_inter_matrix*
Adjustment coefficients α and β: α, β

Result: Network polarization index

$intra\_neg\_ratio \leftarrow 0$;
$inter\_neg\_ratio \leftarrow 0$;
for *comm* ***in*** *communities* **do**
  $intra\_neg\_ratio \leftarrow intra\_neg\_ratio + \frac{partition\_neg\_intra[comm]}{partition\_intra[comm]}$
end
$intra\_neg\_ratio \leftarrow \frac{intra\_neg\_ratio}{length(communities)}$
for *comm*1 ***in*** *communities* **do**
  for *comm*2 ***in*** *communities* **do**
    if *comm*1 ***not equal*** *comm*2 **then**
      $inter\_neg\_ratio \leftarrow inter\_neg\_ratio + \frac{neg\_inter\_matrix[comm1,comm2]}{inter\_matrix[comm1,comm2]}$
    end
  end
end
$inter\_neg\_ratio \leftarrow \frac{inter\_neg\_ratio}{length(communities)\cdot length(communities)-length(communities)}$
return $\alpha \cdot intra\_neg\_ratio + \beta \cdot inter\_neg\_ratio$;

these posts are published (timestamps), so that we could compare the estimated polarization (and the corresponding flow of information) at different moments of time.

3.2.1 Estimating information flow

In the literature, it is possible to find several entropy measurements to estimate the information flow between the users of a network. In [52], the authors show how, using the entropy rate estimator (h), originally proposed in [53], some other entropy measurements could be derived. The entropy rate is defined as follows:

$$h = \frac{N \log N}{\sum_{i=0}^{N} \Lambda_i} \tag{1}$$

where N refers to the length of the text whose entropy is being calculated, and Λ_i refers to the length of the longest non-contiguous subsequence from the target text that has appeared previously (in the previous i symbols) as a contiguous subsequence in the text.

From that basic information flow estimator, the authors propose the time-synchronized cross-entropy metric [52], which leverages the usage of entropy to calculate the information flow between two users taking into account the dynamics of the conversation, which can be considered useful in the context of social networks. Such a measurement is mathematically defined below:

$$h(T\|S) = \frac{N_T \log_2 N_S}{\sum_{i=1}^{N_T} \Lambda_i(T|S_{\leq t(T_i)})} \tag{2}$$

where T refers to the target, S refers to the source (for instance, target user and source user), N refers to the length of the text (where the sub-index determines whether it is the length of the source text or the length of the target text), and Λ_i refers to the longest subsequence in the target text that appears as contiguous in the source text, taking into consideration the time when the posts were published, if the metric is applied in the context of social networks' information flow measurement.

Although the time-synchronized cross-entropy is already a valid metric for measuring information flow, we decided to measure the information flowing through a network using an entropy metric derived from it, called *Neighbor Normalized Information Flow (NNIF)*, also defined in [52]. The reason behind this decision is simple: based on the benchmark carried out by the authors in [52], NNIF is able to measure a network's information flow in a much more reliable way than the time-synchronized cross-entropy, and other entropy measurements derived from it. In fact, such an entropy estimator has already been used in the literature (as it was aforementioned) achieving good results [54]. As a consequence, we decided to use NNIF as the metric to estimate the information flow in a network, when calculating our polarization metric, SPIN. Formally, the NNIF metric is defined as [52]:

$$NNIF(S,T) = \frac{h(T\|S)}{\sum_X h(T\|X)} - \frac{h(S\|T)}{\sum_X h(S\|X)} \tag{3}$$

where $h(Y\|Z)$ refers to the time-synchronized cross-entropy between a target Y and a source Z (Equation (2)), T refers to the target node, S refers to the source node, and X refers to any node in the (local) neighborhood.

3.2.2 Restricting entropy computation to negative information

Since the different entropy measurements allow capturing the flow of information generated in the network, the intuition behind the usage and restriction of entropy computation to negative information is simple: the more negative information is transmitted through a network, the more polarized it should be; hence, measuring entropy on negative information becomes the cornerstone upon which *SPIN* operates. As an abuse of language, we shall call *negative entropy* to those entropies (or related measurements) computed on negative information.

Specifically, to restrict the computation of entropy to negative information, the *LIWC 2007* framework [55] was used. This framework allows, through text analysis, to obtain the count of words from each specific category present in a particular text, for instance, in a tweet. These categories correspond to time, money, food, exclusion, inclusion, family, people, etc. Moreover, other categories of particular importance for calculating negative information are associated with sentiment analysis: number of words with positive emotion, negative emotion, number of words expressing sadness, anxiety, etc. Beyond all these categories, the framework also allows characterizing users' speech through other categories like the number of personal pronouns (in each personal form) and number of verbs (in different personal forms), allowing, for instance, to analyze if certain users consistently refer to supposed third parties.

Considering the above, we decided to use this framework to count the words of each category, filtering those tweets with any negative sentiment (not necessarily having a value for the category $EmoNeg > 0$, as more complex conditions could be used involving other categories related to speech, or certain sensitive topics like money or work). Thus, by filtering negative tweets (information), negative entropy (as described in the previous section) would be calculated, which provide an estimation on how much negative information is truly flowing through the network. Indeed, other LIWC categories could have been used to carry out negative emotion tweet detection, however, we considered the aforementioned manner of filtering tweets the most understandable one to keep tweets with a more or less (depending on the tweet) negative emotion.

While it is true that sentiment analysis techniques could also be used to detect the tweets containing negative information, we consider model explainability to play a fundamental role within the detection of negative posts for the application of polarization detection, as the polarization index to be computed can depend on the posts with negative sentiment detected. As a result, we opted to use the LIWC framework,[7] because it allows us to fully understand why a given post is considered to have (or not) negative emotion, while also being able to reliably capture such emotion (trade-off between prediction accuracy and explainability). On the other hand, this capability is often not provided by neural network-based models that could also be used for determining the sentiment of a post, as the model's interpretability often decreases with the model's predictive power [56].

While it is true that the LIWC tool is proprietary, its license has a low cost and it has been widely used across many similar works in the literature [57–59], thus its usage seemed to be ideal in a similar context as the one we perform in this research, where LIWC is used for detecting the negative information flowing through the network.

[7] Using LIWC also allows to introduce a more (or less) strict negative sentiment detection, to fine-tune the polarization index.

3.2.3 Community detection for political contexts

As already discussed, to apply many of the polarization algorithms detailed herein, a pre-partitioned network is required. However, since our focus, and in particular, SPIN, is on socio-political analysis, considering aspects such as the ideology of the users is fundamental, as it allows the partition of the graph into well-separated communities from a domain knowledge point of view. For instance, a user with a left-leaning political view should not be in the same partition as a user with right-leaning political view. Consequently, the usage of other graph partitioning algorithms, such as METIS [49] or Louvain [50], might not be as accurate (from a domain knowledge point of view) as SPIN requires, because those methods are typically based on maximizing network modularity [60], without considering important domain knowledge information, such as the political leaning of the users. In fact, this (important) distinction on how communities are identified is not always made in the polarization/controversy research, making this a relevant novelty of this work.

Thus, in order to generate a set of partitions more adequate to the study of the political conversation in Online Social Networks such as X, we propose a *label propagation* approach. This approach, as described in Algorithm 3, allows nodes to "influence" their neighboring nodes until converging on a final partition. The community generation algorithm takes a weighted directed (potentially multi-directed) graph as its main input, together with a list of already labeled "seed" nodes. In the graph, each node represents a user, each link represents a retweet relation (A retweeted B). The weight of the link corresponds to the number of retweets from one node to another.

By design, this algorithm has the following characteristics:

1 *Number of communities*: it supports the division into 2 or more communities, as its application in studying polarization, as obvious as it seems, may require more than two partitions in the conversation graph. For instance, when considering political data, it can be inferred that there are multiple communities or partitions of users, and not necessarily just two; there could be many more. At least as many as the number of political parties in the system.

2 *Correct partitions from a semantic (domain knowledge) point of view*: this method encompasses, according to a certain semantic component, nodes that are similar within the same partition, and in a different partition compared to those that are semantically different. This is starkly different from other existing partitioning algorithms, such as *METIS* [49] or *Louvain* [50], where the graph's structure is used to create the partition. Additionally, it also utilizes the structural aspects of the network (e.g., connections) to assign communities in the best possible way, but always respecting the meaning of the partitions. For instance, in a political use-case, nodes with different ideologies should not be included in the same partition, even if that made sense from the topological point of view of the network.

Indeed, the described label propagation process only generates a list of potential members for the communities provided. Therefore, another step is required to determine which communities a node belongs to, based on a minimum community membership threshold that must be also provided as a parameter to this algorithm (τ). If a node has a value, for one or more communities, higher than the threshold, such a node will belong to all those communities meeting the condition. However, it could happen that a node does not have any community membership value higher than the minimum community membership threshold (τ). As a result, that node would not belong to any partition. To avoid this,

Algorithm 3: Community detection based on label propagation

Data: Directed graph: *network*
Initial dictionary of labeled nodes: *labeled_nodes*
Number of iterations: *max_iter*
Number of communities: N
Community membership threshold: τ
Creation of the "others" community: *create_others*
Result: Partition assignment to each node

final_labels ← **copy of** *labeled_nodes*;
unlabeled_nodes ← **set difference of** *network.nodes* **and** *keys*(*labeled_nodes*);
// Initialize the labels of the unlabeled nodes to $\frac{1}{N}$
for *node* ***in*** *unlabeled_nodes* **do**
  final_labels[*node*] ← *vector*(N, $\frac{1}{N}$); // vector of N $\frac{1}{N}$ values
end
for *iteration* ← 0 ***to*** *max_iteration* − 1 **do**
  unlabeled_nodes ← *shuffle*(*unlabeled_nodes*);
  for *node* ***in*** *unlabeled_nodes* **do**
    new_label ← *vector*(N, 0); // vector of N zeros
    weight_sum ← 0;
    for (*neighbor*, *node*, *weight*) ***in*** *network.in_edges*(*node*) **do**
      new_label ← *new_label* + *weight* · *final_labels*[*neighbor*];
      weight_sum ← *weight_sum* + *weight*;
    end
    for (*node*, *neighbor*, *weight*) ***in*** *network.out_edges*(*node*) **do**
      new_label ← *new_label* + *weight* · *final_labels*[*neighbor*];
      weight_sum ← *weight_sum* + *weight*;
    end
    new_label ← $\frac{new\_label}{weight\_sum}$;
    final_labels[*node*] ← *new_label*;
  end
end
// Assigning the list of partitions (communities) of each node
node_communities ← {};
for *node* ***in*** *network* **do**
  list_partitions ← **get** communities of *final_labels*[*node*] with value ≥ τ;
  if *create_others* ***and*** *list_partitions* ***is*** *empty* **then**
    node_communities[*node*] ← [N + 1];
  else
    node_communities ← *list_partitions*;
  end
end
return *node_communities*;

we allow the creation of a new community including all those nodes that could not be assigned to any other partition.

As it was mentioned, a weighted graph is required for this partitioning algorithm. This graph can potentially be a multi-directed graph if we considered multiple "agreement"[8] relationships (e.g., in X, we could consider retweets and likes). However, for this research we focused exclusively on the retweet relationships between the different nodes, as retweets tend to express more agreement than likes [61]. The (weighted and potentially multi-

[8]There is a need to use agreement relationships, so as to propagate the adequate labels between connected nodes.

directed) graph to be used for this community detection algorithm must have the same nodes as the graph to be used for the SPIN algorithm, but it could use a subset of the relationships used by the graph built for polarization detection with SPIN. For instance, SPIN can use reply and quote relationships between users, but these relationships would not be as useful as like or retweet relationships for the task of community detection, so they could be removed from the graph built for community detection between nodes. Thus, one can understand that in Alg. 3 line 11 the weight refers to the sum of weights of the different relationships present in the graph built for community detection. In our case, we used retweet relationships to carry out community detection for the aforementioned reasons.

3.2.4 Network representation

In our proposed algorithm, SPIN, the representation of Online Social Networks, such as X, is done through a multi-directed graph whose nodes represent users and whose edges represent connections between users. Those edges must, in fact, be weighted, as the edge weight represents the frequency of the interactions between users. The intuition behind this representation is quite simple: two users (nodes) could have one or more connections with each other in any direction, and these connections are related to user interactions that could have occurred multiple, potentially many, times due to the natural interactions and information exchange that characterizes microblogging Online Social Networks, such as X. However, as the representation involves a (multi)directed graph, it is fundamental to define both the possible interactions between users, as well as the direction of those interactions. In this section, X (Twitter) interactions are considered, although similar interactions could be defined for other social networks, thus using any other social network would not have an impact on the network representation required for the SPIN algorithm:

- *Retweets*: The connection (edge) between the user (user A) that creates the retweet and the retweeted user (user B) must go from the retweeted user to the user that creates the retweet, i.e., from user B to user A. The intuition behind this is simple: the information in the network flowed from user B (who created the original post) to user A (who read the post and retweeted it).
- *Replies*: The connection between the user (user A) that replies to the post created by another user (user B) must go in the direction of the replied user to the user that created the reply, following the same direction as the information flow in the network, i.e., from user B to user A.
- *Quotes*: Similarly to the replies, the connection between the user that creates the quote (user A) and the user whose post is quoted (user B) must go in the direction of the quoted user to the user that created the quote, i.e., from user B to user A.
- *Mentions*: In the case in which a user (user A) creates a post in which it mentions another user (user B), the edge must go in the direction of the user that creates the mention (user A) to the mentioned user (user B), as the intention of user A is to make user B read the post, thus information flows in that direction.

In the case of SPIN, it is convenient to consider the previous four possible connections between users instead of simply considering retweet networks, follow networks, mention networks, or hashtag networks separately, as it has already been done previously in the literature [32, 48], because by considering those four possible connections (simultaneously), it is possible to represent a bigger portion of the information flowing through a network,

which is what our algorithm, SPIN, tries to detect and quantify. Thus, a correct representation of the network is critical to the correct functioning of SPIN.

4 Results

4.1 Settings

In order to benchmark the utility of our algorithm, we applied Algorithm 1 to obtain datasets related to each of the Spanish electoral processes from 2011 to 2019, including both general and local electoral processes (see Tables 1 and 2 for a description of the statistics of these datasets).

We propose the usage of the political scenario of Spain to benchmark SPIN, as the country is characterized by a complex political scenario where the society is represented by several political parties (not only two), with different political leanings. We consider that using such a complex political scenario could show the true potential of SPIN to carry out reasonable and precise polarization detection and quantification, at the same time that it allows an in-depth sociopolitical analysis thanks to the high explainability of its results.

4.2 Analysis

4.2.1 SPIN as polarization metric

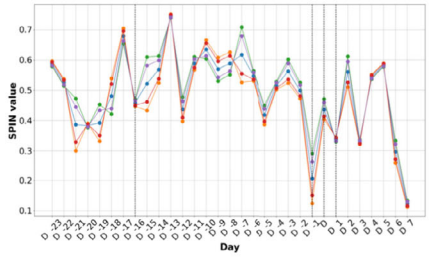
First, we compare the results of SPIN when using different hyperparameters, obtaining the results described in Fig. 1. As we observe from these visualizations, the SPIN algorithm adeptly captures the polarization dynamics outlined in Sect. 2. We see a subtle and incremental rise in polarization from the pre-election phase leading up to the electoral campaign. This is followed by a decrease on the day before the elections (blackout period) due to the inactivity of politicians and parties. There is a notable surge during the election event, which then gradually tapers off post-election.

Both the rise in polarization during the campaign and, notably, its decline during the blackout period serve as strong indicators of the role political parties and candidates play in the escalation of political polarization. Similarly, we also observe that polarization peaks on the days when electoral debates are held on the country's main television channel (RTVE), which take place 5 days prior to the general elections.

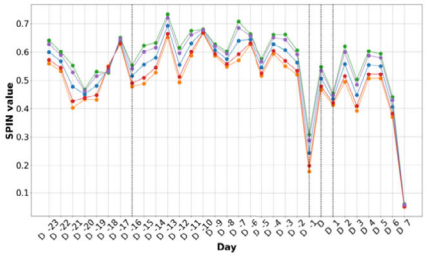
Regarding the evolution of polarization, we broadly observe that polarization has increased over time, displaying consistently higher and more sustained patterns throughout the entire process. This is especially evident during the general electoral processes, with less polarization and greater variability during local processes, although we will analyze this later in more detail. This is likely due to the diversity of cities and options, as well as the adoption of decentralized communication strategies by political parties.

These observations are more or less stable independently of the hyperparameters (α and β to account for intra- and inter-community entropies, and τ as the community membership threshold), although we found more consistent results when $\alpha = \beta = 0.5$, and this is the configuration we shall use in the rest of the work.

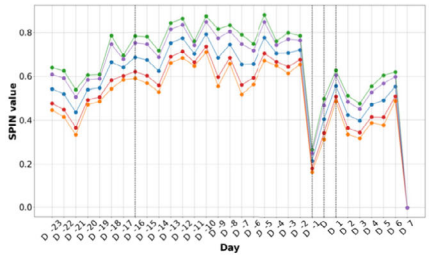
Last, we observe that SPIN suits well the polarization dynamics described in Sect. 2, as polarization increases during the pre-campaign stage, then experiences a drop during the blackout period and a rise during the election day, continuing a polarization trend that tends to decrease as the post-campaign advances. Indeed, the proposed hybrid polarization metric seems to adjust to the known polarization dynamics around electoral processes, while also providing a daily polarization score that is quite stable and does not experience very extreme changes from one day to another.



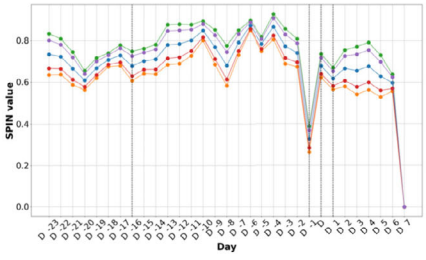
(a) Spanish elections of 2011 (general).



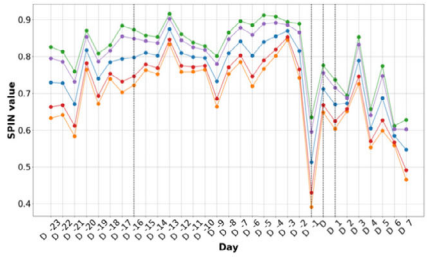
(b) Spanish elections of 2015 (general).



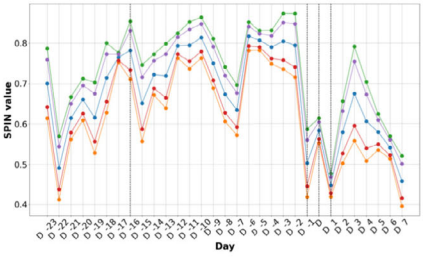
(c) Spanish elections of 2015 (local).



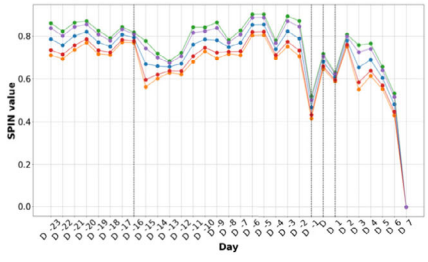
(d) Spanish elections of 2016 (general).



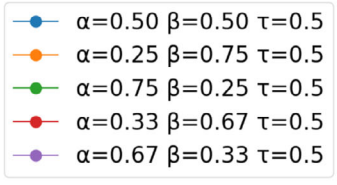
(e) Spanish elections of April 2019 (general).



(f) Spanish elections of 2019 (local).



Spanish elections of November 2019 (general).

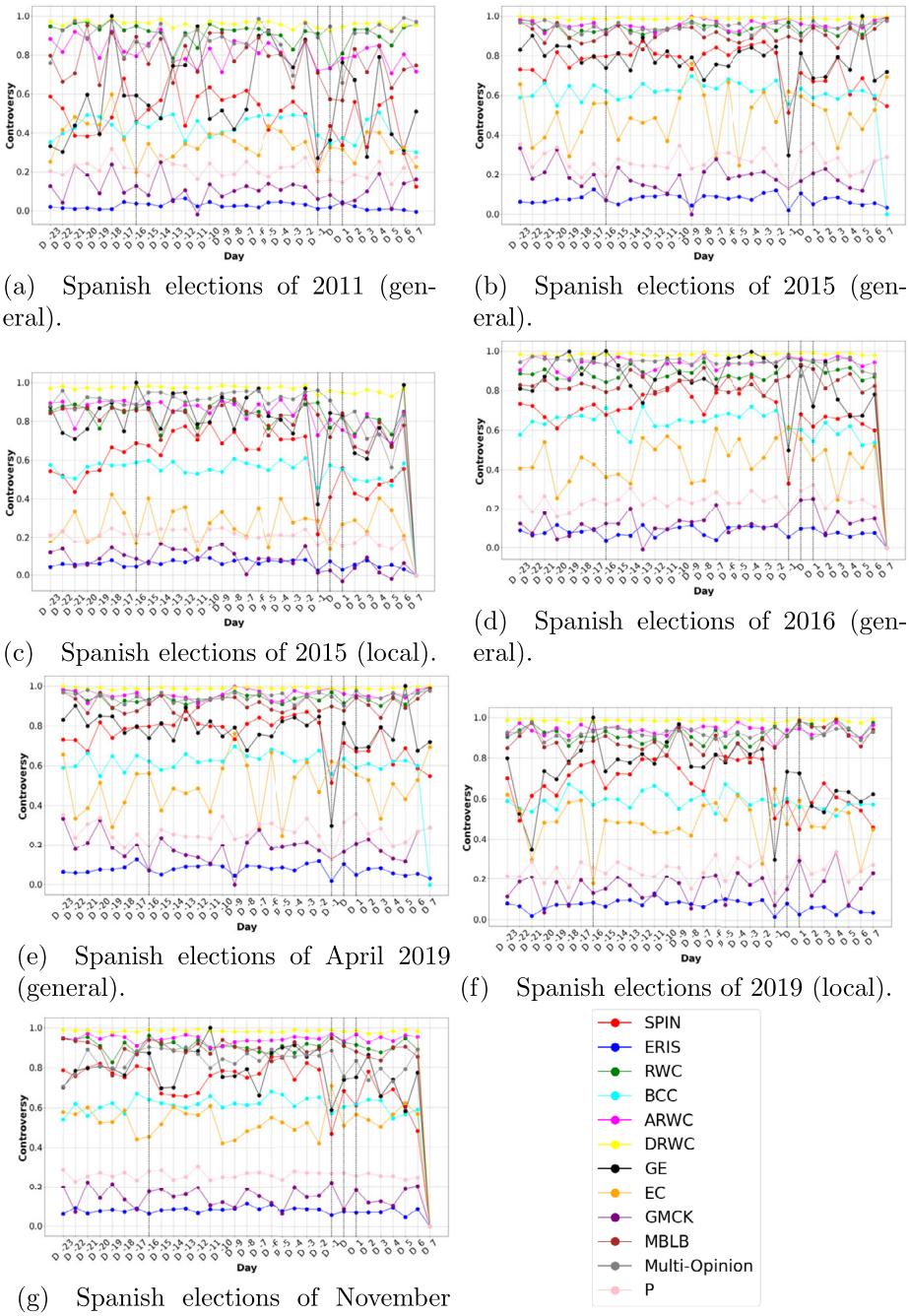


(g) Color legend used in the plots

Figure 1 Comparison of SPIN results with different parametrizations across Spanish electoral processes from 2011 to 2019. Seven plots are shown depicting the SPIN polarization values based on different hyperparameters of this algorithm when applied to the local and general Spanish elections from 2011 to 2019. No large differences are obtained in behavior depending on the hyperparameters considered in the SPIN algorithm

4.2.2 Benchmarking polarization approaches

In this section, we compare our proposed SPIN polarization metric against other algorithms of the literature, so as to determine whether, for the specific context of our use-case (political scenario), our algorithm proves to work better or worse than already existing algorithms. The comparison results are presented in Fig. 2.



(a) Spanish elections of 2011 (general).

(b) Spanish elections of 2015 (general).

(c) Spanish elections of 2015 (local).

(d) Spanish elections of 2016 (general).

(e) Spanish elections of April 2019 (general).

(f) Spanish elections of 2019 (local).

(g) Spanish elections of November 2019 (general).

(h) Color legend used in the plots

Figure 2 Comparison of SPIN results with different parametrizations across Spanish electoral processes from 2011 to 2019. Seven plots are shown depicting the SPIN polarization values compared against 11 polarization approaches from the literature when applied to the local and general Spanish elections from 2011 to 2019. The values obtained by each algorithm are varied, and cover almost the full spectrum of the Y axis. Whereas some methods tend to obtain stable values independently of the day being analyzed, other methods (SPIN included) show consistently polarization values that change throughout the election process: start in the middle, increase until the day of the election (except for the previous day, because of the blackout period), and rises again after the election

When comparing the SPIN algorithm with the rest of the studied algorithms, we find that almost all of them fail to capture polarization according to the well-known polarization dynamics described earlier in this work, particularly during the blackout period and the post-election week. Similarly, a majority of the studied algorithms exhibit particularly high and sustained values over time (RWC, MBLB, Multi-Opinion) or notably low values (ERIS, GMCK, P) that remain almost unchanged as the electoral process advances, thus failing to capture the discussed polarization dynamics in Sect. 2.

Among the most similar algorithms, capable of showing centered values throughout the entire spectrum (neither too high nor too low), we find our proposal alongside GE, EC, and BCC. However, both EC and BCC particularly fail to capture the post-election week and the blackout period. They also display unstable patterns that fluctuate throughout the process. Considering the GE polarization index, we observe how this algorithm follows the well-known characteristics of an electoral process in Spain (see Sect. 2), similarly to our proposed polarization algorithm SPIN. The only difference we observe in the results is that GE tends to have a higher variance in its expected polarization index when compared to SPIN, which tends to be more stable during the prediction of daily polarization from the pre-campaign to the post-campaign phases.

It is worth noting that this trend is consistent across all data sets, corresponding to either general or local elections.

4.2.3 Polarization in local vs general elections

In this section, we use our SPIN algorithm to gain a general understanding on the evolution of the (political) polarization in Online Social Networks. Now, our analysis is divided into two separate studies: the evolution of polarization across general electoral processes (see Fig. 3), and the evolution of polarization across local electoral processes (see Fig. 4), as this approach also allowed to compare the polarization between the two types of electoral processes.

From these figures, we have observed a gradual increase in political polarization over the years. This polarization has escalated similarly in both general and local electoral processes, although it is true that, while polarization has grown in both processes, it remains higher in general elections due to their central role in the country's political and media agenda.

Considering SPIN is well-correlated with the polarization dynamics previously described (see Sect. 2), we conclude the observed increase in polarization is a precise reflection of this phenomenon in the society. In fact, we show now in Fig. 5 the average polarization computed according to our approach, considering both general and local electoral processes. These measurements further emphasize that the polarization surrounding political discourse on X (Twitter) in Spain has experienced an increasing trend from 2011 to 2019. This trend has been particularly pronounced in national politics as opposed to local politics.

5 Discussion

In this work, we propose a novel approach to measure polarization – named *SPIN*, from Social-political Polarization analysis by INformation theory –, based on novel or recently proposed approaches to estimate the information flow, account for negative information, and detect communities in political contexts. The uniqueness of the SPIN algorithm lies

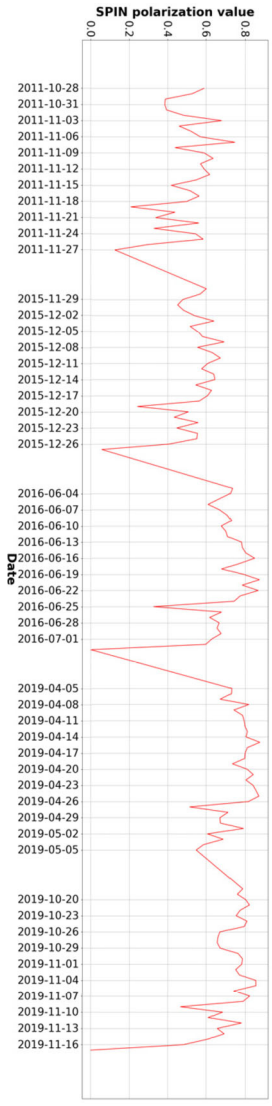


Figure 3 Evolution of polarization during Spanish general electoral processes from 2011 to 2019. The polarization values are plotted throughout all the national election processes analyzed in this study. The trend that can be observed is that the overall value tends to increase in each new election

in its foundation in Information Theory, offering a different perspective on measuring polarization. This approach allows us to encapsulate the core tenets of polarization as articulated by predominant definitions in the field. While other algorithms might offer insights based on surface interactions or apparent divides, SPIN delves into the inherent informational structures and patterns. This deep-rooted analysis ensures a more nuanced and precise understanding of polarization, making SPIN stand out from the other approaches.

A remarkable feature is its incorporation of the temporal dimension, tracking the evolution and flow of information over time. This temporal consideration is crucial as polarization is not static; it evolves, intensifies, or diminishes in response to real-time events and discourses. Additionally, SPIN emphasizes coherence in content, ensuring that consistent themes and narratives within a community are recognized and factored into the analysis. Considering the community's structural intricacies, it keeps sight of the network's topology. By weaving together the temporal, content coherence, and structural aspects, SPIN offers a comprehensive and nuanced insight into the multifaceted nature of polarization.

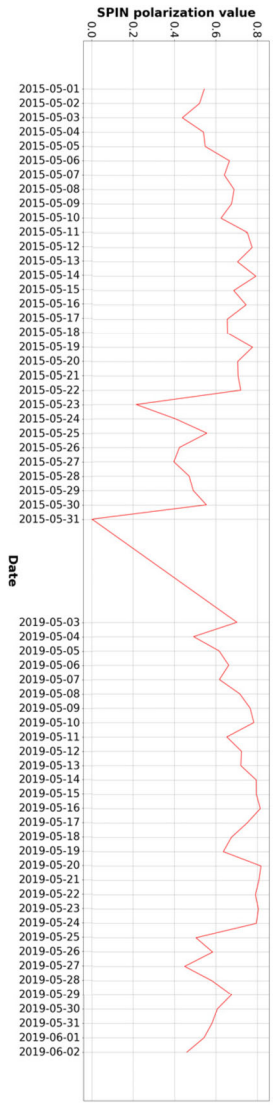


Figure 4 Evolution of polarization during Spanish local electoral processes from 2011 to 2019. The polarization values are plotted throughout all the local election processes analyzed in this study. The trend that can be observed is that the overall value tends to increase in each new election, although slightly less than in Fig. 3

Consequently, our algorithm accurately assesses the polarization phenomenon, can produce results within a reasonable time frame, and the outcomes it generates align with the postulated axioms in the literature regarding what polarization should be during an electoral campaign, based on the discussion in Sect. 2.

All in all, our algorithm takes into consideration both the topology of the network, as well as the information flowing through it, hence utilizing as many resources as possible to carry out polarization detection and quantification. Indeed, other hybrid algorithms have already been proposed and used for the same purpose, nevertheless, none of these mixes a hybrid approach with the analysis of the dynamics of the network, i.e., considering the information that is flowing through the network at every moment of time, which could have a significant influence on the other posts and information exchanges between the users in the network. As a result, our algorithm, SPIN, blends the best of both Information Theory and Social Network Analysis to carry out precise polarization detection and quantification, thanks to its support to multiple communities and its ability to model the

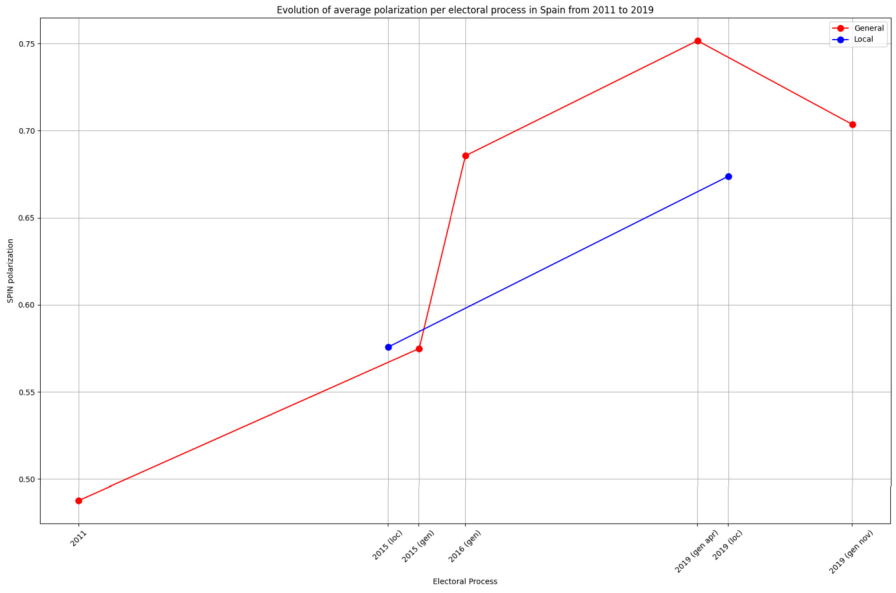


Figure 5 Average polarization per electoral process in Spanish electoral processes from 2011 to 2019. Two lines depict the average polarization value computed through each entire electoral process to compare general and local elections. It is observed that the polarization increases steadily, but the overall polarization is larger for the general case

relationships between them. Based on the presented benchmark results, we consider SPIN to be a useful contribution to the sociopolitical analysis, as it was able to model polarization along different Spanish electoral processes, from 2011 to 2019, in a complex political scenario, where old political parties are still supported by a vast amount of people (PP and PSOE, right- and left-wing, respectively), at the same time that new political parties are emerging and gaining more and more traction in the political arena (such as Ciudadanos, Vox, or Podemos), and divides between supporters of the different political parties tend to broaden more and more.

It is worth noting the high degree of explainability of the proposed polarization metric, in particular, in comparison against other approaches. As we have shown, it has enough variability to capture ups and downs in polarization, especially considering the temporal dimension. This, in particular, matched quite accurately with the main assumptions collected from political polarization literature. Indeed, we argue that these characteristics make the SPIN algorithm a suitable option, not only for polarization detection and quantification, but also for gaining a deep understanding of the sociopolitical elements that most contribute to the polarization of the network to potentially prevent it, as polarization detection and quantification can be used as the tools of today's society for closing divides between the different communities that we can find in social networks, which are no more than a digital reflection of our society.

Our study has two main drawbacks. First, computational efficiency is a concern. There are several factors that can impact the efficiency of the proposed algorithm. While the algorithm has proven its capability and utility in analyzing multiple electoral processes in Spain, it is clear that for especially large data sets (i.e., graphs), the process could take an excessive amount of time. However, this is a challenge common to almost all of the algorithms studied. More specifically, in our particular case, if the communities are large, the

number of times an entropy measurement needs to be calculated increases significantly (since it scales pairwise based on connections between nodes), which can notably increase the execution time. Similarly, nodes with many publications also affect the efficiency of the algorithm, since entropy takes longer to calculate in these cases.

Second, the accuracy of the partitioning can also influence the result. The proposed algorithm is heavily dependent on the network partitioning algorithm (label propagation). If a good partitioning is used, the results will be accurate. However, if the partitioning is not appropriate (semantically, not structurally), then the results can deviate significantly from reality. Specifically, in the case of the label propagation algorithm, if the "seed" nodes are not selected in a way that truly represents the political system, the effectiveness of the algorithm will not be accurate and the semantics of the polarization results will not be valid.

Prospective research trajectories for the SPIN algorithm present a plethora of opportunities. There is potential in refining and optimizing its mechanics for even more accurate results. Venturing beyond the Spanish-language datasets, testing its applicability across diverse linguistic and cultural contexts can further validate its universal relevance. Moreover, while its current focus is on political polarization, SPIN's underlying principles hold promise for broader applications. Areas like marketing can benefit from understanding consumer polarities, preferences, or brand loyalties. Similarly, in misinformation, such an algorithm can be pivotal in discerning echo chambers, biased information flows, and the intensity of misleading narratives. Such expansive applications could position SPIN as a versatile tool for various analytical challenges.

Indeed, the implications of the SPIN algorithm extend beyond mere measurement. One of its most profound utilities lies in its potential to inform and shape strategies to counter polarization. By offering a detailed insight into the intricate webs of polarization — including its temporal evolution, content coherence, and community structure — SPIN provides policymakers, platform developers, and community leaders with a granular understanding of where and how divisions occur. With this knowledge, targeted interventions such as developing custom recommendation strategies can be formulated to bridge divides, foster understanding, and promote more cohesive dialogues. The SPIN algorithm is not just a diagnostic tool; it is a foundation upon which effective solutions to the challenges of polarization can be built.

Although in this research we focused on the utilisation of SPIN to get a daily polarization score, one can understand that it may naturally be used for studying the polarization evolution within each of the communities detected by the label-propagation based community detection algorithm during a given period of time. Indeed, our algorithm is not only capable of providing a daily polarization metric, but it can also be used for the purpose of studying polarization within the detected communities, which allows a more nuanced polarization analysis around an electoral process, providing the capability of answering questions such as "Which community of users is linked to a higher negative information flow?", "To which extent two communities are polarized between each other?", or even "Which user is responsible for the highest negative information flow output?". Indeed, we consider that SPIN implications go far beyond from just a daily measurement, as it can potentially be used for offering detailed insights regarding polarization analysis.

Abbreviations
ARWC, Authoritative Random Walk Controversy; BCC, Betweenness Centrality Controversy; BRW, Biased Random Walk; DRWC, Displacement Random Walk Controversy; EC, Embedding Controversy; GE, Generalized Euclidean; NLP, Natural Language Processing; NNIF, Neighbor Normalized Information Flow; RWC, Random Walk Controversy; SPIN, Social-political Polarization analysis by INformation theory.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00480-3.

Additional file 1. Polarization algorithms information. The attached supporting document contains a detailed description of the polarization algorithms discussed in Sect. 2. (PDF 202 kB)
Additional file 2. Algorithms to compute intra- and inter-community entropy. The attached supporting document contains a detailed description of the algorithms discussed in Sect. 3.2. (PDF 229 kB)

Acknowledgements
The authors thank the reviewers for their thoughtful comments and suggestions.

Author contributions
PM designed the study, collected the data, implemented the algorithms from the state of the art, implemented the proposed polarization approach, and performed the analysis. AB and FD supervised the methodology and obtained results, performed the analysis. RB implemented the algorithms from the state of the art, contributed towards the final version of SPIN. All authors wrote the paper and approved the final manuscript.

Funding
This work was supported by Grant PID2022-139131NB-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe."

Data availability
The datasets generated and analyzed during the current study are not publicly available to comply with Twitter/X Developer Agreement but are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Universidad Autónoma de Madrid, Madrid, Spain. [2]University of Castilla-La Mancha, Ciudad Real, Spain.

Received: 30 September 2023 Accepted: 20 May 2024 Published online: 05 June 2024

References

1. Garcia D, Abisheva A, Schweighofer S, Serdült U, Schweitzer F (2015) Ideological and temporal components of network polarization in online political participatory media. Policy Internet 7(1):46–79. https://doi.org/10.1002/poi3.82
2. Chen K, Luo Y, Hu A, Zhao J, Zhang L (2021) Characteristics of misinformation spreading on social media during the covid-19 outbreak in China: a descriptive analysis. Risk Manag Healthc Policy 14:1869–1879. https://doi.org/10.2147/rmhp.s312327
3. Tufekci Z (2017) Twitter and tear gas: the power and fragility of networked protest. Yale University Press, New Haven
4. Boulianne S (2020) Twenty years of digital media effects on civic and political participation. Commun Res 47(7):947–966
5. Wojcieszak M, Price V (2013) The impact of candidate communication strategies on citizens' attitudes and behavior: a social identity framework. Polit Psychol 34(3):337–361. https://doi.org/10.1111/pops.12007
6. Kubin E, von Sikorski C (2021) The role of (social) media in political polarization: a systematic review. Ann Int Commun Assoc 45(3):188–206
7. Torcal M, Comellas JM (2022) Affective polarisation in times of political instability and conflict. Spain from a comparative perspective. South Eur Soc Polit 27(1):1–26. https://doi.org/10.1080/13608746.2022.2044236
8. Ford R, Jennings W (2020) The changing cleavage politics of western Europe. Annu Rev Pol Sci 23:295–314
9. Algan Y, Guriev S, Papaioannou E, Passari E (2017) The European trust crisis and the rise of populism. Brookings Pap Econ Act 2017(2):309–400
10. Padró-Solanet A, Balcells J (2022) Media diet and polarisation: evidence from Spain. South Eur Soc Polit 27(1):75–95
11. Ramírez-Dueñas JM, Vinuesa-Tejero ML (2021) How does selective exposure affect partisan polarisation? Media consumption on electoral campaigns. J Int Commun 27(2):258–282. https://doi.org/10.1080/13216597.2021.1899957
12. Halu A, Zhao K, Baronchelli A, Bianconi G (2013) Connect and win: the role of social networks in political elections. Europhys Lett 102(1):16002. https://doi.org/10.1209/0295-5075/102/16002
13. Yen DA, Dey B (2019) Acculturation in the social media: myth or reality? Analysing social-media-led integration and polarisation
14. Barberá P (2020) Social media, echo chambers, and political polarization. Soc Media Democ 34

15. Overgaard CSB, Dudo A, Lease M, Masullo GM, Stroud NJ, Stroud SR, Woolley SC (2021) Building connective democracy: interdisciplinary solutions to the problem of polarisation. In: The Routledge companion to media disinformation and populism. Routledge, London, pp 559–568
16. Vergeer M, Hermans L, Sams S (2013) Online social networks and micro-blogging in political campaigning: the exploration of a new campaign tool and a new campaign style. Party Polit 19(3):477–501
17. Marozzo F, Bessi A (2018) Analyzing polarization of social media users and news sites during political campaigns. Soc Netw Anal Min 8:1–13
18. Wagner M (2021) Affective polarization in multiparty systems. Elect Stud 69:102199. https://doi.org/10.1016/j.electstud.2020.102199
19. Bischof D, Wagner M (2019) Do voters polarize when radical parties enter Parliament? Am J Polit Sci 63(4):888–904. https://doi.org/10.1111/ajps.12449
20. Valle MED, Broersma M, Ponsioen A (2021) Political interaction beyond party lines: communication ties and party polarization in parliamentary Twitter networks. Soc Sci Comput Rev 40(3):736–755. https://doi.org/10.1177/0894439320987569
21. Alsinet T, Argelich J, Béjar R, Martínez S (2021) Measuring polarization in online debates. Appl Sci 11(24):11879. https://doi.org/10.3390/app112411879
22. Ajovalasit S, Dorgali V, Mazza A, D'Onofrio A, Manfredi P (2021) Evidence of disorientation towards immunization on online social media after contrasting political communication on vaccines. Results from an analysis of Twitter data in Italy. PLoS ONE 16(7):0253569. https://doi.org/10.1371/journal.pone.0253569
23. Arora SD, Singh GP, Chakraborty A, Maity M (2022) Polarization and social media: a systematic review and research agenda. Technol Forecast Soc Change 183:121942. https://doi.org/10.1016/j.techfore.2022.121942
24. Hansen KM, Kosiara-Pedersen K (2017) How campaigns polarize the electorate: political polarization as an effect of the minimal effect theory within a multi-party system. Party Polit 23(3):181–192
25. Johnston R, Lachance S (2022) Polarization and campaign dynamics in canada, 1988–2021
26. Aragón P, Kappler KE, Kaltenbrunner A, Laniado D, Volkovich Y (2013) Communication dynamics in twitter during political campaigns: the case of the 2011 spanish national election. Policy Internet 5. https://doi.org/10.1002/1944-2866.POI327
27. Bruns A, Burgess J (2011) # ausvotes: how Twitter covered the 2010 Australian federal election. Commun Polit Cult 44(2):37–56
28. Gunnarsson Lorentzen D (2014) Polarisation in political Twitter conversations. Aslib J Inf Manag 66(3):329–341
29. Hernández E, Anduiza E, Rico G (2021) Affective polarization and the salience of elections. Elect Stud 69:102203
30. Gagrčin E et al (2023) Datafication markers: curation and user network effects on mobilization and polarization during elections. Media Commun 11(3)
31. Olivares G, Cárdenas JP, Losada JC, Borondo J (2019) Opinion polarization during a dichotomous electoral process. Complexity 2019
32. Garimella K, Morales GDF, Gionis A, Mathioudakis M (2018) Quantifying controversy on social media. ACM Trans Soc Comput 1(1):3–1327. https://doi.org/10.1145/3140565
33. Emamgholizadeh H, Nourizade M, Tajbakhsh MS, Hashminezhad M, Esfahani FN (2020) A framework for quantifying controversy of social network debates using attributed networks: biased random walk (BRW). Soc Netw Anal Min 10(1). https://doi.org/10.1007/s13278-020-00703-1
34. Garimella K, Morales GDF, Gionis A, Mathioudakis M (2017) Quantifying Controversy in Social Media. arXiv:1507.05224
35. Villa G, Pasi G, Viviani M (2021) Echo chamber detection and analysis: a topology- and content-based approach in the covid-19 scenario. Soc Netw Anal Min 11. https://doi.org/10.1007/s13278-021-00779-3
36. Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw 30:107–117
37. Jacomy M, Venturini T, Heymann S, Bastian M (2014) Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. PLoS ONE 9:98679. https://doi.org/10.1371/journal.pone.0098679
38. Morales AJ, Borondo J, Losada JC, Benito RM (2015) Measuring political polarization: Twitter shows the two sides of venezuela. Chaos, Interdiscip J Nonlinear Sci 25(3). https://doi.org/10.1063/1.4913758
39. Guyot A, Gillet A, Leclercq E, Cullot N (2022) ERIS: an approach based on community boundaries to assess polarization in online social networks, pp 88–104. https://doi.org/10.1007/978-3-031-05760-1_6
40. DiMaggio P, Evans J, Bryson B (1996) Have American's social attitudes become more polarized? Am J Sociol 102(3):690–755. Accessed 2023-09-03
41. Chen X, Lijffijt J, De Bie T (2018) Quantifying and minimizing risk of conflict in social networks. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '18. Assoc. Comput. Mach., New York, pp 1197–1205. https://doi.org/10.1145/3219819.3220074
42. Matakos A (2017) Measuring and moderating opinion polarization in online social networks. https://api.semanticscholar.org/CorpusID:19844211
43. Wojatzki M, Mohammad SM, Zesch T, Kiritchenko S (2018) Quantifying qualitative data for understanding controversial issues. In: International conference on language resources and evaluation. https://api.semanticscholar.org/CorpusID:21709904
44. Kabir MY, Madria S (2022) A deep learning approach for ideology detection and polarization analysis using covid-19 tweets. In: Ralyté J, Chakravarthy S, Mohania M, Jeusfeld MA, Karlapalem K (eds) Conceptual modeling. Springer, Cham, pp 209–223
45. Yang M, Wen X, Lin Y-R, Deng L (2017) Quantifying content polarization on Twitter. In: 2017 IEEE 3rd international conference on Collaboration and Internet Computing (CIC), pp 299–308. https://doi.org/10.1109/CIC.2017.00047
46. Bell A, Solano-Kamaiko I, Nov O, Stoyanovich J (2022) It's just not that simple: an empirical study of the accuracy-explainability trade-off in machine learning for public policy. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, pp 248–266
47. Benslimane S, Azé J, Bringay S, Servajean M, Mollevi C (2021) Controversy detection: a text and graph neural network based approach. In: Zhang W, Zou L, Maamar Z, Chen L (eds) Web information systems engineering - WISE 2021 - 22nd international conference on web information systems engineering, WISE 2021, proceedings, part I, Melbourne, VIC, Australia, October 26-29, 2021. Lecture notes in computer science, vol 13080. Springer, Berlin, pp 339–354. https://doi.org/10.1007/978-3-030-90888-1_26

48. Singh M, Iyengar SRS, Kaur R (2022) A multi-opinion based method for quantifying polarization on social networks. arXiv:2204.08697
49. Karypis G, Kumar V (1997) Metis—a software package for partitioning unstructured graphs, partitioning meshes and computing fill-reducing ordering of sparse matrices
50. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10008. https://doi.org/10.1088/1742-5468/2008/10/p10008
51. Hohmann M, Devriendt K, Coscia M (2023) Quantifying ideological polarization on a network using generalized Euclidean distance. Sci Adv. https://doi.org/10.1126/sciadv.abq2044
52. South T, Smart B, Roughan M, Mitchell L (2022) Information flow estimation: a study of news on Twitter. Online Soc Netw Media 31:100231. https://doi.org/10.1016/j.osnem.2022.100231
53. Kontoyiannis I, Algoet P, Suhov Y, Wyner A (1998) Nonparametric entropy estimation for stationary processesand random fields, with applications to English text. IEEE Trans Inf Theory 44:1319–1327. https://doi.org/10.1109/18.669425
54. Smart B, Watt J, Benedetti S, Mitchell L, Roughan M (2022) #IStandWithPutin versus #IStandWithUkraine: the interaction of bots and humans in discussion of the Russia/Ukraine war. In: Lecture notes in computer science. Springer, Berlin, pp 34–53. https://doi.org/10.1007/978-3-031-19097-1_3
55. Pennebaker J, Chung C, Ireland M, Gonzales A, Booth R (2007) The development and psychometric properties of liwc2007
56. Diwali A, Saeedi K, Dashtipour K, Gogate M, Cambria E, Hussain A (2023) Sentiment analysis meets explainable artificial intelligence: a survey on explainable sentiment analysis. IEEE Trans Affect Comput, 1–12. https://doi.org/10.1109/TAFFC.2023.3296373
57. Giuntini FT, Cazzolato MT, dos Reis MdJD, Campbell AT, Traina AJ, Ueyama J (2020) A review on recognizing depression in social networks: challenges and opportunities. J Ambient Intell Humaniz Comput 11:4713–4729
58. Himelboim I, Xiao X, Lee DKL, Wang MY, Borah P (2020) A social networks approach to understanding vaccine conversations on Twitter: network clusters, sentiment, and certainty in hpv social networks. Health Commun 35(5):607–615
59. Kivran-Swaine F, Naaman M (2011) Network properties and social sharing of emotions in social awareness streams. In: Proceedings of the ACM 2011 conference on computer supported cooperative work. CSCW '11. Assoc. Comput. Mach., New York, pp 379–382. https://doi.org/10.1145/1958824.1958882
60. Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582. https://doi.org/10.1073/pnas.0601602103
61. Boyd D, Golder S, Lotan G (2010) Tweet, tweet, retweet: conversational aspects of retweeting on Twitter. In: 2010 43rd Hawaii international conference on system sciences. IEEE, Los Alamitos

Publisher's Note