https://doi.org/10.1140/epjds/s13688-024-00493-y

EPJ.org

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

Analyzing user ideologies and shared news during the 2019 argentinian elections

Sofía M. del Pozo[1,2], Sebastián Pinto[1,2], Matteo Serafino[3], Lucio Garcia[1,2], Hernán A. Makse[3] and Pablo Balenzuela[1,2*]

[*]Correspondence: balen@df.uba.ar
[1]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física, Buenos Aires, Argentina
[2]CONICET - Universidad de Buenos Aires, Instituto de Física Interdisciplinaria y Aplicada (INFINA), Buenos Aires, Argentina
Full list of author information is available at the end of the article

Abstract

The extensive data generated on social media platforms allow us to gain insights over trending topics and public opinions. Additionally, it offers a window into user behavior, including their content engagement and news sharing habits. In this study, we analyze the relationship between users' political ideologies and the news they share during Argentina's 2019 election period. Our findings reveal that users predominantly share news that aligns with their political beliefs, despite accessing media outlets with diverse political leanings. Moreover, we observe a consistent pattern of users sharing articles related to topics biased to their preferred candidates, highlighting a deeper level of political alignment in online discussions. We believe that this systematic analysis framework can be applied to similar scenarios in different countries, especially those marked by significant political polarization, akin to Argentina.

Keywords: News sharing; Social media; Political polarization; Confirmation bias; News content analysis

1 Introduction

In 1998 Richard Feynman wrote *"The first principle is that you must not fool yourself, and you are the easiest person to fool."* [1]. How we perceive things and subsequently respond to them is a phenomena potentially influenced by personal biases.

The widespread use of social media platforms generates a large amount of data which, through careful interrogation and analysis, could reflect extensive and valuable information [2–4]. This data not only sheds light on, for instance, trending topics [5, 6] and public opinions [7–10] but also provides insights into the individual characteristics of users based on their behavior, such as their interactions and the news they share [11, 12]. In particular, news sharing behavior on social media is a phenomenon worthy of study [13–15], not only for its potential to infer users' information but also for its significant potential to influence society. Concerning the accuracy of shared news, the propagation of fake news could have serious implications, such as during elections [16, 17] and the COVID-19 pandemic, where misinformation heightened anxiety and psychological distress [18].

Numerous factors can influence the process of news sharing behavior [15, 19, 20]. For example, Osmundsen et al. [21] demonstrated in their study that partisan polarization is the primary psychological motivation behind the sharing of political fake news on Twitter. Westerwick et al. [22] examined the relationship between sources and content cues for confirmation bias, revealing that confirmation bias emerged irrespective of source quality [22]. In this sense, we can observe that user characteristics, in particular their political leaning or biases, can serve as both an explanation of news sharing behavior and as information resulting from this behavior.

Bias is defined as the tendency to favour or dislike a person or thing, especially as a result of a preconceived opinion [23]. While bias can manifest in different ways [24–27], the two types of biases that specifically concern us in this study are those influencing news consumption behavior and those affecting the media on social networks, known as confirmation bias and media bias, respectively. As Raymond Nickerson explained in [28], confirmation bias refers to the inclination to search for or interpret evidence in a manner that aligns with pre-existing beliefs, expectations, or a currently held hypothesis. In essence, it represents an unintentional shaping of facts to fit one's hypotheses or beliefs. In the context of our research, confirmation bias can be recognized as an instance of the *selective exposure theory*, as described by Stroud (2010) [29], which elucidates individuals' propensity to prefer information that conforms to their pre-existing beliefs, while consciously avoiding contradictory content. Regarding news consumption research, media bias takes on a prominent role. Media bias is defined as a deliberate and intentional tendency that favors a particular perspective, ideology, or desired outcome [27, 30].

As we mentioned above, social media serves as a channel for news consumption, where several factors influence the dynamics of how these news are shared. In particular, both confirmation bias and media bias can interplay when the news that social media users read are shared by others who possess their own ideological biases. Therefore, the study of this dynamic is of significant importance due to the impact of news consumption on people's opinions [31–34], for example the consumption of biased news can influence voters' decisions [35]. Additionally, the interaction between social media, political polarization, and political disinformation can significantly shape a society's future, affecting the quality of public policy and its democratic principles [36].

In this study, we examine the relationship between shared news and the ideologies of social media users who disseminate them during the 2019 general elections in Argentina. Specifically, we explore whether factors such as the news source, bias, or topics shared by users are associated with their political ideology. Our analysis incorporates data on the content of the shared news and the political affiliations of the users, previously categorized into Center-Left (CL) and Center-Right (CR) groups. The Twitter activity and partisan labels were obtained from the research conducted by Zhou et al. (2021) as referenced in [37].

The focal point of this study lies in examining the news shared by users within the existing dataset from [37]. To collect this data, we performed web scraping of the text from the links of news articles shared by users. Following this, we evaluated the bias of these news articles, along with the bias of the news media and the topics they cover. Subsequently, we analyzed the correlation between these factors and the users' bias towards the candidates from the two primary coalitions competing for the presidency.

2 Background

2.1 Argentinian context

The systematic framework introduced in this work, aimed at quantifying both user and media outlet preferences, fills a significant gap in understanding, especially within the Argentine context. While some of the main Argentine media outlets are listed on Media Bias/Fact Check organization [38], currently, there is no centralized source for evaluating the media bias of all outlets in the country. As we apply this framework to Argentina during the 2019 presidential election campaign, this section offers an overview of the political and media landscape during this period to provide contextualization.

Over the past decade, Argentina's political scene has been characterized by the predominance of two major coalitions: one, a center-Checkleft coalition (CL) led by Cristina Fernández de Kirchner, known as *Frente de Todos*, and the other, a center-right coalition (CR) led by Mauricio Macri, referred to as *Juntos por el Cambio*. Cristina Kirchner held the presidency in Argentina during the periods of 2007 – 2011 and 2011 – 2015, while Mauricio Macri served as president from 2015 to 2019, as documented in [39]. During the 2019 elections, the center-left coalition presented Alberto Fernández and Cristina Fernández de Kirchner as their candidates. Meanwhile, the center-right coalition sought a second term for Mauricio Macri as president, with Miguel Ángel Pichetto as his vice-presidential candidate. National elections in Argentina comprise two obligatory phases: the primary election, known as PASO (which stands for *Primarias, Abiertas, Simultáneas y Obligatorias* in Spanish, translating to Open, Simultaneous, and Obligatory Primaries in English), and the general election. In the year 2019, these events occurred on August 11th and October 27th, respectively. Additionally, if the results of the general election necessitate it, a third round, referred to as a *ballotage*, may also be conducted.

Regarding the media landscape, the digital media scene in Argentina is primarily characterized by three major players: *Infobae*, *Clarín* and *La Nación*, each boasting approximately 20 million unique users in 2020, as reported by Comscore data [40]. Following closely are a second tier of media outlets with audience numbers ranging from 6 to 13 million unique visitors. Prominent among this group are *Página 12*, *Ámbito Financiero*, *TN Noticias* and *El Destape Web*.

In Argentina, a pronounced polarization has been reported through the distinct ideological orientations of the country's primary media outlets [4, 41]. For instance, Página 12 is recognized as a left-of-center broadsheet newspaper, while Clarín is considered a centrist tabloid and La Nación is characterized as a center-right newspaper [42]. Between 2008 and 2014, a confrontation occurred between the government of Cristina Fernández de Kirchner (Center-Left) and major media corporations [43]. During this period, a conflict arose, leading to the establishment of a set of newspapers aligned with the policies of the Kirchner government (e.g., Página 12). Simultaneously, another cluster of newspapers emerged, known for their vehement editorial criticism of the government's actions during this era (e.g., Clarín and La Nación, among others) [43–45].

While the examination of media outlet bias is increasing, particularly among English-based outlets, the scenario is different in countries like Argentina. Notably, only three of the main outlets in Argentina have definitive bias classifications provided by Media Bias/Fact Check [46–48]. In this context, our study not only classified Argentine news outlets but also introduced a versatile bias index for situations where specific classifications are lacking.

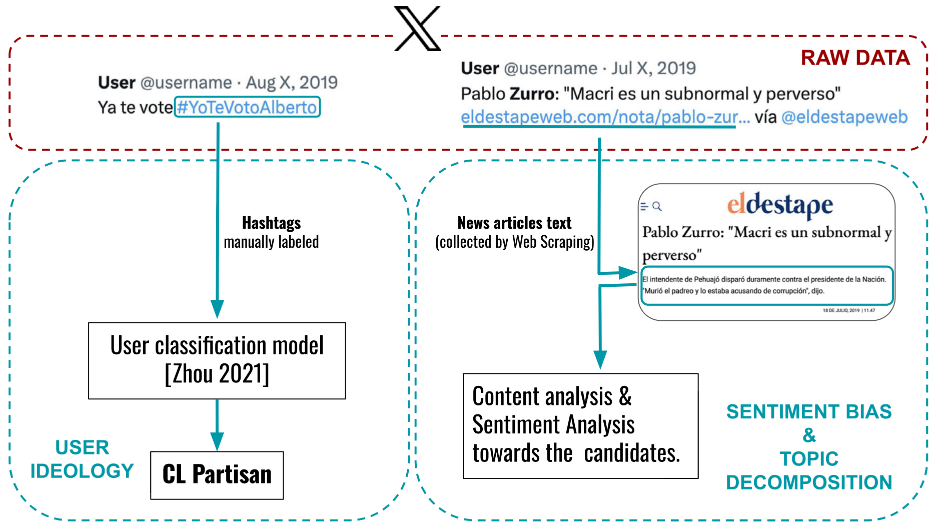
3 Material and methods

This section provides an overview of the data and methods utilized in this study. Figure 1 illustrates the progression of our pipelines, starting with the raw tweets (top panel of Fig. 1). Users are classified as supporters of a particular candidate based on the content of their tweets [37]. Additionally, tweets containing URLs to external media outlets undergo scraping (right panel of Fig. 1), allowing for the analysis of news outlet bias, news, and topic bias based on the text of the news, rather than the text of the tweets themselves. Below is a detailed description of our methods.

3.1 Users' classification

The users classification process begins with a manually classified set of hashtags, collecting and categorizing the most frequent hashtags as Pro-Fernandez, Pro-Macri, or Neutral. Tweets containing only these classified hashtags are then selected to create a training set, comprising 253482 tweets, which was then employed to train a classifier (depicted in the left panel of Fig. 1).

To identify the best classifier, Zhou et al. [37] tested five different models: Logistic Regression (LR) with L2 regularization, Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Decision Tree (DT). These models were validated on 10% of the classified tweets. As shown in Table 4 of [37], the Logistic Regression model performed the best, with an average group accuracy, recall, and F1-score all at 83%. The SVM followed with an 81% accuracy, then Naive Bayes with 79.5%, and finally Random Forest and Decision Tree. Logistic Regression assigns a probability p to each tweet, indicating its likelihood of supporting either candidate. A probability closer to one indicates support for Macri, while a probability closer to zero indicates support for Fernandez. Ultimately, users' opinions are inferred based on the latest number of tweets classification, defining



User @username · Aug X, 2019
Ya te vote #YoTeVotoAlberto

User @username · Jul X, 2019
Pablo **Zurro**: "Macri es un subnormal y perverso"
eldestapeweb.com/nota/pablo-zur... vía @eldestapeweb

Pablo Zurro: "Macri es un subnormal y perverso"

El intendente de Pehuajó disparó duramente contra el presidente de la Nación. "Murió el padreo y lo estaba acusando de corrupción", dijo.

Figure 1 Methodology pipeline. *Top:* example of raw data of tweets from social media Twitter (now X). *In the left:* hashtags were utilized to train a logistic regression model for classifying tweets as supportive of one candidate or the other. Users are assigned to the candidate for whom they exhibit the highest number of supportive tweets (see more details in [37]). *In the right:* the news URLs in the tweets are utilized to extract the text by web scrapping to execute then all the necessary steps leading to perform sentiment and topic analysis

loyalty classes. The methodology proves robust even when varying the number of tweets considered to determine user loyalty.

It's important to note that both in the training process and in the classification process, the model employed in [37] only considers the text of the tweet; any external information contained in the tweet, such as references to a news outlet, is not taken into account. Further details can be found in [37].

3.2 Data

This study starts with an existing Twitter dataset [37] containing tweets collected between March 1, 2019, and August 27, 2019. The data was obtained using keywords associated with candidates for the 2019 Argentina primary election, including alferdez, CFK, CFKArgentina, Kirchner, mauriciomacri, Macri, and Pichetto. A bots and fake accounts cleaning process was performed over this dataset in the original work (see details in [37]). However, we have run an additional analysis of the impact of potential bots based on Botometer API [49], whose details can be found in Additional file 1.

We refined the original dataset by a) including only tweets containing an external URL linking to an Argentinian news outlet and b) considering users involved in computing the final vote intention. This process yielded 65,971 tweets from 17,466 users intending to vote for the Center-Left (CL) coalition (Fernández-Fernández) and approximately 40,211 tweets from 15,425 users intending to vote for the Center-Right (CR) coalition (Macri-Pichetto). Intending CL coalition voters shared 19,395 news articles, while intending CR voters shared 10,219. The tweets considered in this work represent approximately 0.1% of the raw data (see Supplementary Information of [37]).

It's noteworthy that the while users' political orientation was computed in [37] by considering all the tweets of a user (with and without a URL), and the model was trained using a set of hashtags, in this paper, we concentrate on a subset of those tweets (those containing a URL) and on the text of the news articles themselves, which was not utilized in [37].

3.3 Data filtering

In order to acquire the primary dataset for our analysis, we implemented the following procedures:

1. *Tweets with shared news selection:* We filter all tweets from the data collected by Zhou et al. (2021) [37] that contained a URL in the *url_expanded* Twitter field. This included tweets, retweets, and quotes.
2. *Urls expansion:* Requests python library [50] is used to expand the urls, applying multiprocessing.Pool.map() [51] to parallelize the process.
3. *Urls filter by media:* We retain only the URLs corresponding to news from Argentine media outlets based on ABYZ News Links Guide [52].
4. *Scraping news articles:* For each media outlet, we develop a dedicated code to scrape the content from their respective web pages based on the python libraries Requests [50], Selenium [53] and Beautiful Soup [54]. We acquire the texts of the news articles shared by users.

3.4 News articles sentiment analysis

After scraping, we perform sentiment analysis on the text of the shared news articles. We decompose each article into multiple sentences and apply Pysentimiento algorithm [55] to each sentence within every article. This allows us to calculate positivity, neutrality, and negativity levels with regard to the two main election candidates. Sentiment is defined only for sentences that mention the candidates. If there is a single mention, it is counted as one. If there are multiple mentions, sentiment is calculated separately for each mention, categorizing them as neutral, positive, or negative. Given potential misleading in sentiment classification and the role of irony, we performed a hand-labeled classification in order to measure the accuracy of the model and the potential presence of ironic mentions, which can be found in Additional file 1.

3.4.1 Sentiment bias

We define the Sentiment Bias (*SB*) [41, 56] of a news article as the balance between positive and negative mentions of the candidates of the *CL* coalition (Fernández-Fernández) versus the candidates of the *CR* coalition (Macri-Pichetto) using the following formula:

$$SB = \frac{(\#CR_{+} - \#CR_{-}) - (\#CL_{+} - \#CL_{-})}{\#CR_{total} + \#CL_{total}} \tag{1}$$

where each mention is defined per sentence and the total number of mentions counts positive, negative and neutral ones.

For example, if an article has six sentences with mentions to candidates: one negative mention of CR candidates ($\#CR_{-} = 1$), two positive mention of CL candidates ($\#CL_{+} = 2$), and three neutral mention to CL candidates, then $\#CR_{+} = 0$, $\#CL_{-} = 0$, $\#CR_{total} = 1$ and $\#CL_{total} = 5$. The Sentiment Bias of the article is calculated as calculate $SB = \frac{(0-1)-(2-0)}{1+5} = \frac{-3}{6} = -0.5$.

3.4.2 Interpretation of the sentiment bias

Since Sentiment Bias (*SB*) is a fundamental metric in this study, this section delves into its analysis and provides a detailed interpretation. To conduct this analysis, we first manually classified a group of articles by selecting a random sample of 120 articles with well-defined SB, that is, articles in which candidates from either the Center-Left (CL) or Center-Right (CR) coalitions are mentioned. We then applied the majority rule to this manual classifications to obtain a unique label for each coalition. For instance, if an article received classifications of two positive, two negative, and two neutral with respect to a given coalition, we labeled the article as neutral for that coalition. Finally, we determined the overall connotation of the article based on the criteria shown in Table 1.

Table 1 Criteria to determine the overall connotation of each article

Connotation over CL	Connotation over CR	Overall connotation
–1	–1	Neutral (0)
–1	0	Favorable CR (1)
–1	1	Favorable CR
0	–1	Favorable CL (–1)
0	0	Neutral
0	1	Favorable CR
1	–1	Favorable CL
1	0	Favorable CL
1	1	Neutral

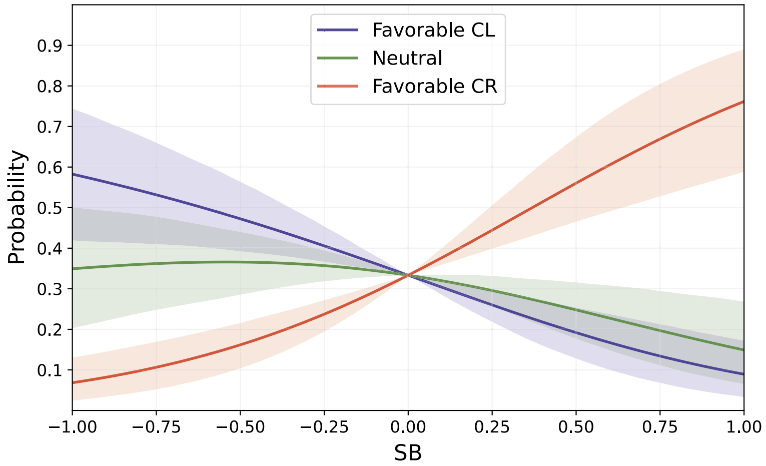


Figure 2 Interpretation of the SB. This figure displays the probability of an article being favorable towards CL, CR, or neutral, given the value of SB measured over the article content. The shaded regions represent the 90% confidence intervals calculated by bootstrapping

Given the overall connotation of each article, we applied logistic regression to correlate the SB value assigned to an article with its label. Specifically, we propose:

$$P(l = i|SB) = \frac{e^{a_i SB}}{\sum_i e^{a_i SB}}$$

where l is the connotation of the article, and $i = -1, 0, 1$ represents being favorable towards CL, neutral, and favorable towards CR, respectively. The coefficients a_i are inferred by fitting the model to the labeled data. In order to keep the model as simple as possible, we chose not to include intercepts b_i in the exponent of the exponential functions (i.e., $a_i SB + b_i$), after finding them to be insignificantly different from zero. The estimated coefficients are as follows: $a_{-1} = -0.89$ $[-1.44, -0.44]$, $a_0 = -0.37$ $[-0.82, 0.07]$, and $a_1 = 1.26$ $[0.82, 1.94]$. The numbers in brackets denote the 90% confidence intervals, which were calculated using bootstrapping.

In Fig. 2, we present the inferred probability of an article's connotation based on the measured value of SB. This figure facilitates the interpretation of the SB value. For instance, a $SB = 0$ indicates an equal probability for an article to be either neutral or positive towards a given coalition. An article with a SB slightly deviating from zero already indicates a clearly favorable trend towards a specific coalition. On the other hand, extreme values ($SB = -1$ or $SB = 1$) do not necessarily represent a probability equal to 1 of being favorable to a certain coalition. Instead, there is a significant fraction of neutral articles with these SB values, and a small fraction of articles that express the opposite opinion, likely due to misclassifications by the sentiment detection algorithm [55]. Additionally, we observed a slight asymmetry for extreme SB values, with a higher probability of an article being neutral when $SB = -1$ compared to when $SB = 1$.

3.5 Topic decomposition

We process the content of the articles by describing the texts within the bag-of-words framework. Specifically, we represent the corpus as a matrix of documents and terms, allowing subsequent topic description. To do this, we proceed with the following steps:

- *Pre-Processing Text.*

  Given that we use a text representation based on word frequency, it is important to delete, on one hand, those words that are redundant and, on the other hand, those words that are non-informative, such prepositions and articles, in order to represent texts on a reduced set of meaningful words. This set of words will constitute our "vocabulary".

  With this in mind, we perform two things: First, we apply lemmatization on the texts using the python library Spacy [57], specifically we use *es_core_news_md* model [58]. Lemmatization transforms all the words to their roots, for instance, all verbs are transformed to their infinitive form and all substantives are transformed to their singular form. Then, we remove stopwords defined in NLTK python library [59] (which, for instance, includes articles and prepositions), as well as rare words (that we defined as those that appears in only one text of the corpus) and very frequent words that were not included in the stopword list but were present in more than 70% of the news articles.
- *TF-IDF*

  After defining the vocabulary, we proceed to describe texts in the bag-of-words framework.

  We start by describing each article by a term-frequency (TF) vector. This description transforms a given text to a vector where each component points out the number of times a given word of the vocabulary appears in the text. We construct this representation through the object *CountVectorizer* from the python scikit-learn library [60].

  Moreover, to reduce word frequency bias and boost the impact of meaningful words, we compute for each word the Inverse Document Frequency (IDF) coeficient, defined as $\text{idf}_j = log(\frac{N}{N_j})$, where N represents the total number of articles within the corpus, while N_j denotes the count of articles containing the j-th term. To do this calculation, we apply the object *TfidfTransformer* from [60].

  With these ingredients, each text is finally described with the Term Frequency - Inverse Document Frequency (TF-IDF) coeficients [61], where the j-th component of the "article vector" i is given by:

  $$v_{ij} = f_{ij} \cdot \log\left(\frac{N}{N_j}\right)$$

  where f_{ij} is the frequency of term j in article i and N_j is the number of documents where the term j appears, as it was stated before.

  Then the articles corpus is described as a matrix $M \in \mathrm{R}^{n\times m}$, with n the number of articles in the corpus and m the number of terms included in the vocabulary. This matrix is a concise representation of the corpus where the meaningful words (both frequent and specific words) are enhanced for each text.
- *Topic Decomposition.*

  In this step, in order to identify the main topics of the corpus of news articles, we apply the unsupervised topic detection algorithm Non-negative Matrix Factorization (NMF) model from scikit-learn python library [60] on the news-term matrix M constructed in the previous step. NMF decomposes matrix M into the product of two matrices, ensuring that all elements are non-negative:

$$M \approx H \cdot W\text{ , }where\ H \in \mathrm{R}^{n\times t}\ and\ W \in \mathrm{R}^{t\times m}$$

Here, t represents the selected number of topics, and H and W denote the resulting matrices of the decomposition. In particular, H defines how each article is described in terms of topics. The element h_{ij} points out the weight of topic j on article i. In other words, it quantifies how much article i belongs to topic j. In order to interpret these weights in terms of probabilities, each row is normalized such as $\sum_j^t h_{ij} = 1$.

On the other hand, rows of matrix W specify the description of each topic in terms of the vocabulary built above. In this case, the element w_{ij} denotes the weight of term j in topic i, i.e, how well term j describes topic i. In this case, by only identifying the weightiest terms allows to interpret what the topic talks about.

3.5.1 Media agenda

Following the procedure outlined in [8], we define the media agenda as the proportion of articles associated with each topic. Specifically, we define the weight of topic j, T^j, as:

$$T^j = \frac{1}{n}\sum_i^n h_{ij} \tag{2}$$

with h^{ij} being the weight of topic j on article i (as defined earlier) and n the number of unique articles shared by the media outlets. We interpret T^j as the collective interest of media outlets in topic j. This measure indicates the likelihood of finding an article associated with topic j in our dataset. (in this case, we are not considering the number of times each article was sharing in social media. Therefore, this definition holds for unique articles).

3.5.2 Partisans agenda

In order to distinguish the interest of partisans groups over the topics found above, we define the interest of partisan group p over topic j as the average of elements h_{ij} (weight of topic j on article i) weighted by the number of times group p shares article i (s_{pi}):

$$T_p^j = \frac{\sum_i^n s_{pi}h_{ij}}{\sum_i^n s_{pi}} \tag{3}$$

where $\sum_i^n s_{pi}$ is equal to the total number of times users from group p shared an article i and n being the total number of articles. T_p^j tells us the probability that an article associated with topic j is shared by an user identified with group p.

4 Results

As outlined in the Introduction, the aim of this study is to investigate how and which characteristics of shared news, and to what extent, correlate with the political ideologies of the users sharing them. To achieve this, we analyze various characteristics of news articles shared by users with identified political leaning, encompassing their sources, distribution of topics, and the political biases that manifest at several levels.

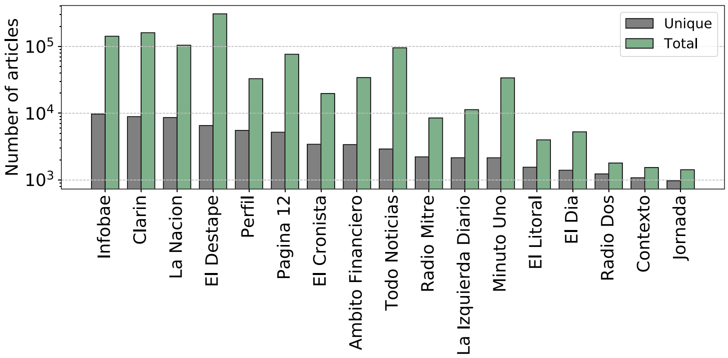


Figure 3 Distribution of News Articles by Media Outlet. Grey bars show the unique articles on Twitter by media outlet (*unique* means counted once), ordered descending. Green bars show the distribution considering each instance of sharing, highlighting user preferences

4.1 Data description

An essential characteristic of news, potentially informative for analyzing the relationship between users' ideologies and the news they share, is the source from which they originate—the media outlet. We begin by analyzing the distribution of news articles on Twitter categorized by their originating media outlets, as illustrated in Fig. 3. The grey bars represent the descending order of the number of unique articles shared from each media outlet, where *unique* indicates that each article is counted only once, regardless of how many times it was posted. Meanwhile, the green bars depict the distribution of news articles shared by Twitter users, taking into account the frequency of each article's sharing.

From the grey bars, it can be deduced that approximately 60% of the articles originate from a specific set of outlets: Infobae, Clarín, La Nación, El Destape, Perfil, and Página 12, listed in descending order by count. This distribution mirrors the activity level of these outlets, with Infobae being the most active in terms of articles published.

On the other hand, the green bars highlight the impact of user preferences on the distribution. For instance, articles from El Destape constitute about 30% of the shared content, underscoring its significance despite not being the highest in publication volume. The same six outlets (with Perfil replaced by Todo Noticias) account for approximately 80% of the shared articles.

4.2 Sentiment bias

We then analyze the political bias of these media outlets using the Sentiment Bias (SB) metric introduced in Sect. 3. This metric measures the tendency of an article to lean positively or negatively towards one of two political coalitions, CL and CR. The SB metric provides a score between –1 and 1 for each article that mentions a candidate from either coalition. A score closer to –1 indicates a favorable stance towards CL, while a score closer to 1 indicates a favorable stance towards CR. This metric helps us define the bias of each article and, consequently, of each media outlet.

Figure 4 shows the average sentiment bias ($\bar{SB}$) for each media outlet, calculated as the mean of all SB scores from their respective articles. For instance, Página 12 and El Destape exhibit $\bar{SB}$ values favoring CL, whereas La Nación and Clarín show $\bar{SB}$ values favoring CR. Notably, Infobae, shared by both supporter groups, falls between these two groups of outlets. Regarding the absolute value of $\bar{SB}$, we interpret $\bar{SB} = 0$ as a neutral position

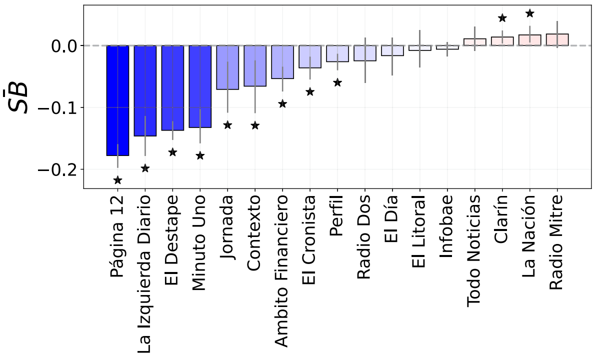


Figure 4 Mean Sentiment Bias ($\bar{SB}$) by media outlet. $\bar{SB}$ represents the average SB across all articles from a specific media outlet. Media outlets positioned on the left side are interpreted as having a bias towards the Center-Left (CL), while those on the right side are considered to have a bias towards the Center-Right (CR). Gray bars indicate the centered 99% quantile of the estimator determined through bootstrapping, and stars denote those estimates significantly different from zero

(see Sect. 3.4.2), meaning most outlets slightly favor CL during the analyzed period. El Destape and Página 12 are more extreme in their positions and can be certainly considered as Center-Left outlets, while Clarín, and La Nación, closer to the center, can be also considered centrist media but slightly lean towards the Center-Right position. For all mentioned media outlets, $\bar{SB}$ significantly deviates from zero marked with stars in Fig. 4), unlike Infobae, which underscores its apparent centrist position.

4.2.1 Selective sharing

After identifying the bias of the articles, we further explore the relationship between users' political ideologies and the news they share on social media by incorporating their political leaning at the time of sharing. This leaning, as computed by Zhou et al. [37], identifies users as belonging to either the Center-Left (CL) or Center-Right (CR) factions during the 2019 Argentine presidential elections.

In order to incorporate this information and motivated by studying confirmation bias, in Fig. 5 we examine the behavior of CL and CR user groups in relation to sharing media outlets, previously identified with specific political biases in Fig. 4, and the bias of the news each group shares. We select media outlets with at least 100 articles shared by each user group, ordered by increasing SB, as shown in left panel of Fig. 5. We then examine the partisans' media preferences by calculating the percentage of each media outlet's articles shared by CL and CR users, as shown in the middle panel of Fig. 5. This panel depicts how the 100% of news for each media outlet is distributed, with articles shared by CL users in red and by CR users in blue. The percentage of the majority group is specified in white within each bar. The right panel displays the average *SB* of articles shared by each group and categorized by outlet. Detailed observations from each panel are discussed below.

A pattern that is, to some extent, anticipated emerges upon examining users' media preferences, in Fig. 5 middle panel. We can see that media outlets with strong biases towards CL, such as Página 12, El Destape and La Izquierda Diario, are primarily shared by CL supporters (for instance, in the first two outlets, CL users share 95% of the articles). Conversely, outlets with biases on the other end of the spectrum, such as La Nación and Clarín, are mostly shared by CR users. We also observe this tendency when examining

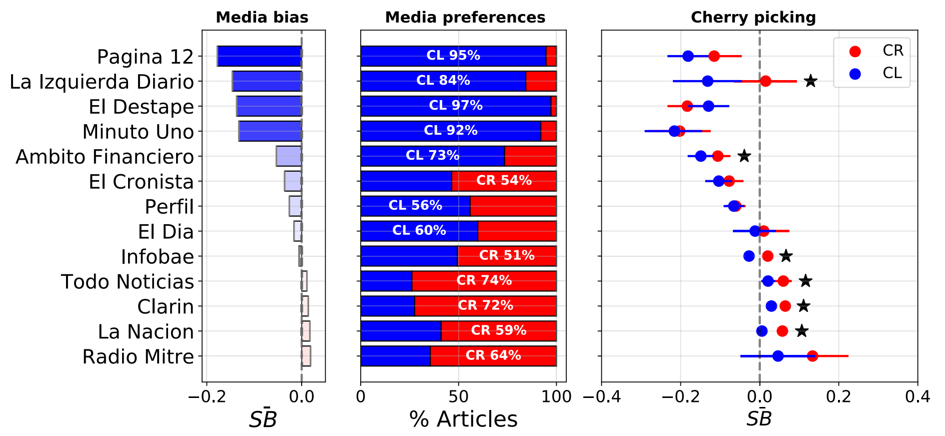


Figure 5 Confirmation bias analysis. *Left:* Mean Sentiment Bias ($\bar{SB}$) of media outlets with at least 100 articles shared by users labeled as CL and CR. *Middle:* Percentage of news articles shared by CL (blue) and CR (red) supporters from each media outlet. *Right:* Average Sentiment Bias of news articles shared by CL and CR partisans across media outlets. Each point represents the average sentiment bias across all articles from a given media outlet shared by each user group, with CL in blue and CR in red. Only media outlets with at least 100 articles shared by each group are included (see Additional file 1). The horizontal bars indicate the centered 99% quantile of the estimate obtained via bootstrapping. Black stars highlight instances where the difference in $\bar{SB}$ between CL and CR supporters is statistically significant, with $p < 0.01$

the *SB* of all articles shared by Center-Left (CL) and Center-Right (CR) supporters (see Supplementary Information for more details). These observations reinforce the selective exposure theory [29], suggesting that users tend to select news from media that favor or align with their pre-existing ideologies. However, the proportion of users sharing news from media outlets with biases similar to their own differs between the two groups. It's noteworthy that while CR users predominantly share CR-favored media, a considerable number of CR-biased news articles are also shared by users with opposing biases (CL).

What is most interesting is when we break down the sentiment bias of each media outlet into two groups: news shared by CL users and the ones shared by CR users. As discussed above, although certain groups of media outlets tend to be more shared by each political coalition, there is a subset (such as Clarín, La Nación, and Infobae) which is significantly shared by both coalitions. However, the content extracted by each coalition from these outlets differs. The right panel of Fig. 5 displays the average *SB* of articles shared by each group for each outlet. This illustrates the phenomenon known as "cherry picking", where users share news that align with their political beliefs, even from opposing outlets. For example, left-leaning supporters share news from La Nación (identified as a right-biased news outlet in Fig. 4) with an average *SB* close to zero, while right-leaning supporters share news from the same outlet with a higher average *SB*. Similar trends are observed in Clarín (Center-Right), Infobae (Centrist), and La Izquierda Diario and Ámbito Financiero (Center-Left). We have statistically validated significant differences between the groups for each outlet, with a p-value below 0.01. Statistically significant differences in news sharing biases are marked with a star.

4.3 Topics interest

This section delves into whether users' political inclinations also affect the topics of the news they share. The findings in previous section establish a link between the users' ideologies and the political bias in the news they share. Here, we aim to determine whether

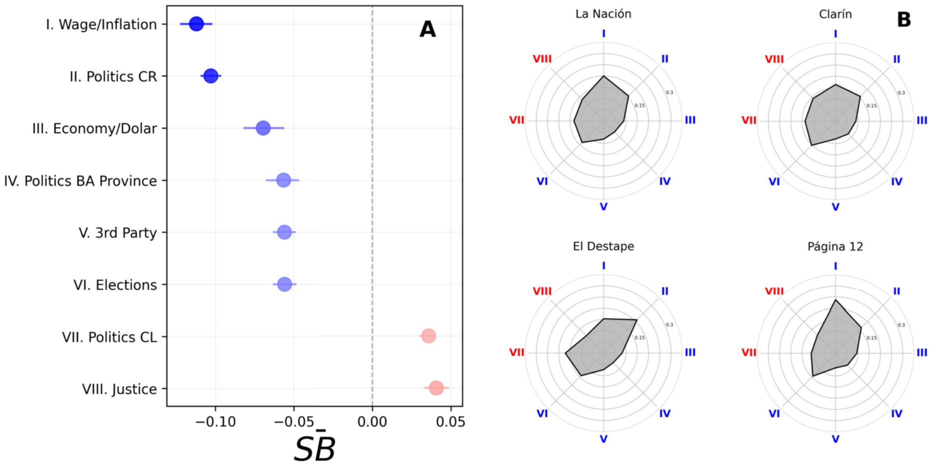


Figure 6 (A) Average Sentiment Bias of emergent topics. Topics are enumerated for subsequent reference. Colors indicate the sign of the *SB* to highlight its orientation: blue signifies a positive bias towards CL, and red indicates a positive bias towards CR. (B) Media Outlet Agendas. Numbers correspond to the topics identified in panel A, with colors reflecting their respective biases. The outlets displayed are those with a clear leaning towards CR (La Nación and Clarín) and CL (El Destape and Página 12), as demonstrated in Fig. 4

specific themes are more supportive of particular candidates and if supporters of each coalition show a preference for these topics.

We initially conduct a topic decomposition of the news articles to identify the principal themes within the dataset, as detailed in Sect. 3. We identified two main families of topics: the first related to economic issues, such as *Wage/Inflation* and *Economy/Dollar*; the second pertains to topics associated with the presidential elections occurring during the analyzed period, including *Politics CR*, *Politics BA Province*, *3rd Party*, *Elections*, *Politics CL*, and *Justice*. Descriptions of these topics, including word clouds and examples of related news articles, are available in the Additional file 1.

Regardless of the interpretation of these topics, which depends heavily on context, panel A of Fig. 6 provides insight into which topics are supportive or against each coalition by displaying the estimated $\bar{SB}$ for each topic. Given that each article is associated with each topic to a varying degree (refer to Sect. 3), $\bar{SB}$ reflects the weighted average *SB* of each article according to this association. For example, this panel indicates that the topic *Wage/Inflation* supports the CL stance, while *Justice* leans towards CR. Notably, topics labeled as *Politics CR* and *Politics CL* appear to favor the coalition contrary to what their labels suggest, likely because they group articles critical of those coalitions. The remaining topics exhibit a slight preference towards CL, aligning with the overall tendency observed during the analyzed period (refer to, for instance, Fig. 4).

Furthermore, we explore the topics covered in news articles to discern the "agenda" of each media outlet (referred to as the "media agenda" in Sect. 3). This agenda essentially represents how each outlet distributes its coverage across the detected topics. Panel B of Fig. 6 showcases the agendas of four media outlets, two with a right-leaning bias (Clarín and La Nación) and two with a left-leaning bias (El Destape and Página 12). Upon inspecting this panel, clear similarities and differences emerge. Much of this coverage behavior can be understood by considering the overall bias of each topic as shown in panel A of Fig. 6 and the bias of each media as depicted in Fig. 4. For instance, Clarín and La Nación

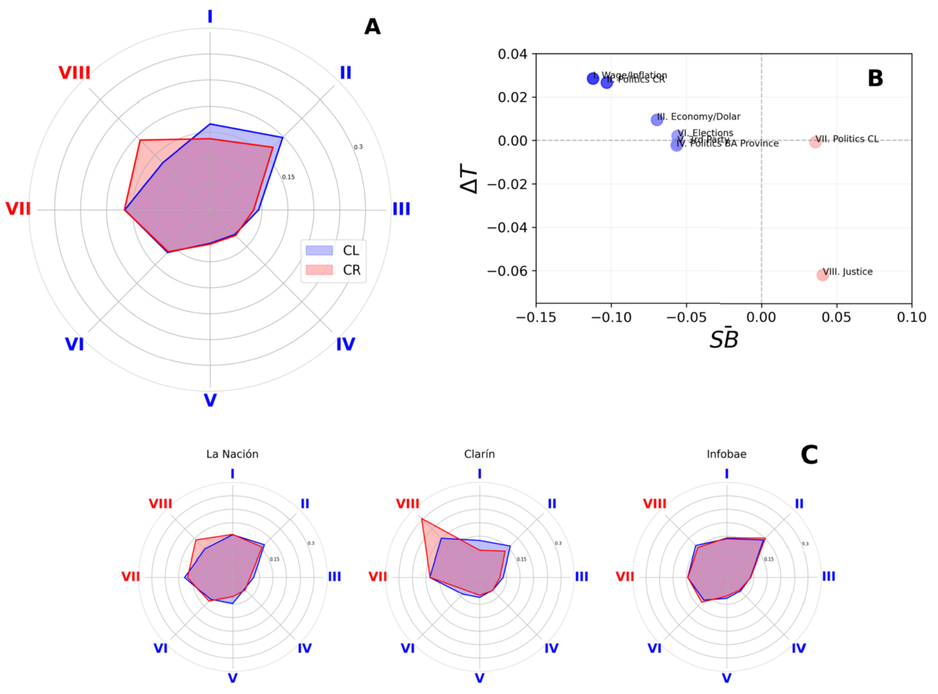


Figure 7 (A) Partisan agendas. Distribution of topic interests for each partisan group. Roman numerals refer to topics outlined in panel A of Fig. 6 and also in panel B of this figure. *(B) Difference in topic interests.* This difference is calculated using equation (4). *(C) Partisan agendas by media outlet.* The outlets depicted are those significantly shared by both coalition groups (see Fig. 5 midle panel)

show a priority for covering *Justice* compared to the other two outlets, whereas Página 12 exhibits a stronger focus on *Wage/Inflation*, and El Destape on *Politics CR*, relative to other topics.

4.3.1 Partisans agenda

The topics described above influence social media users according to their political leanings. These leanings may constrain users to prefer sharing certain topics over others. Panel A of Fig. 7 provides insights into the preferred topics for each partisan group, delineating what we term the "partisan agendas". In this figure, it is evident that CL (Center-Left) users demonstrate a greater interest in topics like *Politics CR* and *Wage/Inflation*, which exhibits a positive inclination towards the CL coalition, whereas CR (Center-Right) users show a preference for the topic *Justice*, with a positive bias towards the CR coalition. Panel B further clarifies the disparity in these interests. We define this difference as

$$\Delta T^j = T^j_{CL} - T^j_{CR} \tag{4}$$

where T^j_{CL} denotes the interest of CL partisans in topic *j*. As illustrated in panel B of Fig. 6, the specified topics significantly align with the coalition of users who share them, indicated by a Spearman correlation coefficient of –0.78 (with a 90% confidence interval of [–1, –0.24]). The sole noticeable exception is the topic *Politics CL*, in which both coalitions appear to have an equal interest, yet it demonstrates an overall inclination towards CR.

Finally, Fig. 5 middle panel showcases the distribution of news shared by each partisan group, this time segmented by media outlet. This panel unveils another dimension of the cherry-picking behavior outlined in Fig. 5 right panel. For example, while users from both coalitions distribute news from Clarín and La Nación, CR users predominantly share content related to the topic *Justice*, which exhibits a CR-favorable bias, whereas CL users are more inclined to share information on *Wage/Inflation*, which aligns with a CL-favorable bias as indicated in panel A of Fig. 6. Nevertheless, this cherry-picking behavior seems to be absent in the topic dissemination from the centrist outlet Infobae, in contrast to the observations made in the right panel of Fig. 5, where each group distinctly shared articles from this media outlet that were biased towards their respective preferences.

5 Discussion and conclusions

In this work, we investigated the relation between shared news on social media and the ideologies of the users who share them. We analyzed both the source of the news articles, their intrinsic bias, and the topics covered, and we related each of these characteristics to the users political ideologies from [37]. To accomplish this, we analyzed the content of news articles shared by politically aligned users on X (ex-Twitter), scraping their content and quantifying both the bias and the topics covered.

Our initial analysis focused on sources (i.e. media outlets). This analysis revealed that the sharing behavior of news by users did not exhibit a distinctly polarized distribution. While certain media outlets may be associated with particular political ideologies (CL and CR), we observed a significant percentage of news from Center-Right (CR) outlets being shared by users identified as Center-Left (CL). See Fig. 5 middle panel. This suggests that the sources of news shared by users on social media may not necessarily indicate their ideology. Our data indicates that Center-Right (CR) media outlets are the most widely consumed in the country, aligning with findings from [40]. Additionally, our results highlight that Center-Right (CR) media outlets reach a more diverse audience in terms of ideological spectra.

We delve deeper into the analysis of the relationship between users' ideologies and the news they share, by examining the bias of news content using the previously introduced Sentiment Bias index [41, 56]. This index effectively categorizes biases of news outlets without making any assumptions, as depicted in Fig. 4. Our findings are consistent with external classifications, where available, validating the accuracy of our approach [38].

When analyzing the average Sentiment Bias ($\bar{SB}$) alongside social media data, a significant trend emerges: users on social platforms tend to share news that aligns with their political beliefs This tendency can be interpreted as indicative of the selective exposure theory [29]. The findings are supported by Fig. 5 right panel, confirming a "cherry-picking" trend: users engage with various journals regardless of their political alignment, yet selectively choose news that resonates with their ideologies. This underscores their preference for content reinforcing their existing beliefs. Furthermore, our analysis extends these patterns to specific topics, as demonstrated in Fig. 7. Users distinctly favor sharing articles related to subjects aligning with their preferred candidates.

While it's expected for users with defined ideological leanings to share news that aligns with their biases, the analysis presented here highlights that this tendency is only apparent when assessing the bias of the content itself, rather than solely relying on media bias. While this phenomenon is predictable, the aim of this study is to introduce a method for quantifying such behavior.

Finally, we'd like to address some remarks and potential limitations of our study. The dataset, while four years old and specific to Argentina, provides unique insights into users' political leanings not found in other datasets. User classification was achieved through a machine learning model, enhancing the dataset's value and enabling us to explore how it correlates with the political bias of shared news content. We believe this analytical framework could be valuable in other countries, especially those with pronounced political polarization like Argentina, and could be adapted to multipolarized scenarios [62].

Abbreviations
CL, Center-Left; CR, Center-Right; SB, Sentiment Bias; $\bar{SB}$, Mean Sentiment Bias.

Supplementary information
Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00493-y.

Additional file 1. (PDF 2.4 MB)

Acknowledgements
We would like to express our sincere gratitude to Alejandro Pardo Pintos, Favio Di Ciocco, and Federico Moss for their valuable contributions during the preparation of this paper.

Author contributions
SMdP, SP, LG, and MS were responsible for collecting the raw data. SMdP developed the computational code utilized consistently throughout the paper. SMdP and SP made contributions to the statistical analysis. PB and HAM conceptualized the research. All authors engaged in discussions about the results and collaborated on the development of the manuscript. All authors read and approved the final manuscript.

Funding
HAM and MS were supported by NSF Grant No. 2214217. PB, SMdP, SP and LG were supported by PICT-2020-SERIEA-00966.

Data availability
The datasets generated and analysed during the current study are available in the OSF repository in https://osf.io/sxwmj/.

Code availability
The corresponding codes are available in https://github.com/sofiadelpozo/SocialMediaBiasAndPolarization.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales, Departamento de Física, Buenos Aires, Argentina. [2]CONICET - Universidad de Buenos Aires, Instituto de Física Interdisciplinaria y Aplicada (INFINA), Buenos Aires, Argentina. [3]Levich Institute and Physics Department, City College of New York, 10031, New York, USA.

Received: 25 April 2024 Accepted: 30 July 2024 Published online: 08 August 2024

References
1. Feynman RP (1998) Cargo cult science. In: Williams J (ed) The art and science of analog circuit design. EDN series for design engineers. Newnes, Amsterdam, pp 55–61
2. Barbier G, Liu H (2011) Data mining in social media. Social network data analytics, 327–352
3. Newman N, Fletcher R, Schulz A, Andi S, Robertson CT, Nielsen RK (2021) Reuters institute digital news report 2021. Reuters Institute for the study. Journalism
4. Newman N, Fletcher R, Eddy K, Robertson CT, Nielsen RK (2023) Digital news report. 2023
5. Chandrasekaran R, Mehta V, Valkunde T, Moustakas E (2020) Topics, trends, and sentiments of tweets about the covid-19 pandemic: temporal infoveillance study. J Med Internet Res 22(10):22624
6. Lee K, Palsetia D, Narayanan R, Patwary MMA, Agrawal A, Choudhary A (2011) Twitter trending topic classification. In: 2011 IEEE 11th international conference on data mining workshops. IEEE, pp 251–258
7. Falkenberg M, Galeazzi A, Torricelli M, Di Marco N, Larosa F, Sas M, Mekacher A, Pearce W, Zollo F, Quattrociocchi W, et al (2022) Growing polarization around climate change on social media. Nat Clim Change 12(12):1114–1121
8. Pinto S, Albanese F, Dorso CO, Balenzuela P (2019) Quantifying time-dependent media agenda and public opinion by topic modeling. Phys A, Stat Mech Appl 524:614–624. https://doi.org/10.1016/j.physa.2019.04.108
9. Anstead N, O'Loughlin B (2015) Social media analysis and public opinion: the 2010 uk general election. J Comput-Mediat Commun 20(2):204–220
10. Klašnja M, Barberá P, Beauchamp N, Nagler J, Tucker JA (2015) Measuring public opinion with social media data

11. Tadesse MM, Lin H, Xu B, Yang L (2018) Personality predictions based on user behavior on the Facebook social media platform. IEEE Access 6:61959–61969
12. An J, Quercia D, Cha M, Gummadi K, Crowcroft J (2014) Sharing political news: the balancing act of intimacy and socialization in selective exposure. EPJ Data Sci 3:12
13. Kalsnes B, Larsson AO (2018) Understanding news sharing across social media: detailing distribution on Facebook and Twitter. Journalism Studies 19(11):1669–1688
14. Kümpel AS, Karnowski V, Keyling T (2015) News sharing in social media: a review of current research on news sharing users, content, and networks. Soc Media Soc 1(2):2056305115610141
15. Lee CS, Ma L (2012) News sharing in social media: the effect of gratifications and prior experience. Comput Hum Behav 28(2):331–339
16. Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. J Econ Perspect 31(2):211–236
17. Bovet A, Makse HA (2019) Influence of fake news in Twitter during the 2016 us presidential election. Nat Commun 10(1):7
18. Rocha YM, Moura GA, Desidério GA, Oliveira CH, Lourenço FD, Figueiredo Nicolete LD (2021) the impact of fake news on social media and its influence on health during the covid-19 pandemic: a systematic review. Journal of Public Health, 1–10
19. Kim DH, Jones-Jang SM, Kenski K (2021) Why do people share political information on social media? Dig Journal 9(8):1123–1140
20. Karnowski V, Leonhard L, Kümpel AS (2018) Why users share the news: a theory of reasoned action-based study on the antecedents of news-sharing behavior. Commun Res Rep 35(2):91–100
21. Osmundsen M, Bor A, Vahlstrup PB, Bechmann A, Petersen MB (2021) Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. Am Polit Sci Rev 115(3):999–1015
22. Westerwick A, Johnson BK, Knobloch-Westerwick S (2017) Confirmation biases in selective exposure to political online information: source bias vs. content bias. Commun Monogr 84(3):343–364
23. Oxford English Dictionary. https://www.oed.com
24. Smith J, Noble H (2014) Bias in research. Evid-Based Nurs 17(4):100–101
25. Delgado-Rodriguez M, Llorca J (2004) Bias. J Epidemiol Community Health 58(8):635–641
26. Kunda Z (1990) The case for motivated reasoning. Psychol Bull 108(3):480
27. Williams A (1975) Unbiased study of television news bias. J Commun 25(4):190–199
28. Nickerson RS (1998) Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol 2(2):175–220
29. Stroud NJ (2010) Polarization and partisan selective exposure. J Commun 60(3):556–576
30. Spinde T, Rudnitckaia L, Mitrović J, Hamborg F, Granitzer M, Gipp B, Donnay K (2021) Automated identification of bias inducing words in news articles using linguistic and context-oriented features. Inf Process Manag 58(3):102505
31. McCombs ME, Shaw DL (1972) Public opinion quarterly. Public Opin Q 36(2):176–187. https://doi.org/10.1086/267990
32. Guo L, McCombs M (2015) The power of information networks: new directions for agenda setting. Routledge, London
33. Diaz-Diaz F, San Miguel M, Meloni S (2022) Echo chambers and information transmission biases in homophilic and heterophilic networks. Sci Rep 12(1):9350
34. Wei D, Zhou T, Cimini G, Wu P, Liu W, Zhang Y-C (2011) Effective mechanism for social recommendation of news. Phys A, Stat Mech Appl 390(11):2117–2126
35. Druckman JN, Parkin M (2005) The impact of media bias: how editorial slant affects voters. J Polit 67(4):1030–1049
36. Tucker JA, Guess A, Barberá P, Vaccari C, Siegel A, Sanovich S, Stukal D, Nyhan B Social media, political polarization, and political disinformation: a review of the scientific literature. Political polarization. And political disinformation: a review of the scientific literature (March 19, 2018) (2018)
37. Zhou Z, Serafino M, Cohan L, Caldarelli G, Makse HA (2021) Why polls fail to predict elections. J Big Data 8(1):1–28
38. Media B (2024) Check. https://mediabiasfactcheck.com/. Accessed: 19th March 2024
39. Cantamutto F (2016) Kirchnerism in Argentina: a populist dispute for hegemony. Int Crit Thought 6(2):227–244. https://doi.org/10.1080/21598282.2016.1172325
40. Todo Medios (Comscore data). https://rb.gy/a56hq9
41. Cicchini T, Del Pozo SM, Tagliazucchi E, Balenzuela P (2022) News sharing on Twitter reveals emergent fragmentation of media agenda and persistent polarization. EPJ Data Sci 11(1):48
42. Bonner MD (2018) Media and punitive populism in Argentina and Chile. Bull Lat Am Res 37(3):275–290
43. Mitchelstein E, Boczkowski PJ (2017) Information, interest, and ideology: explaining the divergent effects of government-media relationships in Argentina. Int J Commun 11:20
44. Becerra M, Marino S, Mastrini G, Dragomir M, Thompson M, Bermejo F, Chan Y-Y, Nissen CS, Reljic D, Southwood R, et al (2012) Mapping digital media. Argentina Observatorio Latinoamericano de Regulación, Medios y Convergencia (OBSERVACOM)
45. Yeager RL (2014) Government control of and influence on the press in Latin America: the case of Argentina during the presidency of Cristina Fernández de Kirchner (2007-2014). Inquiry 17(1):5
46. Clarín Bias. https://mediabiasfactcheck.com/clarin-bias/. Accessed: 19th March 2024
47. La Nación Argentina Bias. https://mediabiasfactcheck.com/la-nacion-argentina-bias/. Accessed: 19th March 2024
48. Infobae Bias. https://mediabiasfactcheck.com/infobae/. Accessed: 19th March 2024
49. Yang K-C, Ferrara E, Menczer F (2022) Botometer 101: social bot practicum for computational social scientists. J Comput Soc Sci 5(2):1511–1528
50. Requests Python Library. https://pypi.org/project/requests/
51. Multiprocesing Python Package. https://docs.python.org/3/library/multiprocessing.html
52. ABYZ Web Links Inc. http://www.abyznewslinks.com/
53. Selenium Python Library. https://pypi.org/project/selenium/
54. BeautifulSoup Python Library. https://pypi.org/project/beautifulsoup4/
55. Pérez JM, Giudici JC, Luque F (2021) pysentimiento: a Python toolkit for sentiment analysis and SocialNLP tasks
56. Albanese F, Pinto S, Semeshenko V, Balenzuela P (2020) Analyzing mass media influence using natural language processing and time series analysis. J Phys Complex 1(2):025005

57. Honnibal M, Montani I (2017) spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear
58. Lemmatization model. https://github.com/explosion/spacy-models/releases/tag/es_core_news_md-3.6.0
59. Bird ELS, Klein E (2019) Natural language processing with Python
60. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
61. Nguyen E (2014) Chapter 4 - Text mining and network analysis of digital libraries in R. In: Data mining applications with R, pp 95–115
62. Martin-Gutierrez S, Losada JC, Benito RM (2023) Multipolar social systems: measuring polarization beyond dichotomous contexts. Chaos Solitons Fractals 169:113244

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.