Journal of Computational Social Science (2023) 6:289–313
https://doi.org/10.1007/s42001-022-00196-2

SURVEY ARTICLE

A scoping review on the use of natural language processing in research on political polarization: trends and research prospects

Renáta Németh[1]

Received: 24 September 2022 / Accepted: 29 November 2022 / Published online: 19 December 2022
© The Author(s) 2022

Abstract

As part of the "text-as-data" movement, Natural Language Processing (NLP) provides a computational way to examine political polarization. We conducted a methodological scoping review of studies published since 2010 ($n = 154$) to clarify how NLP research has conceptualized and measured political polarization, and to characterize the degree of integration of the two different research paradigms that meet in this research area. We identified biases toward US context (59%), Twitter data (43%) and machine learning approach (33%). Research covers different layers of the political public sphere (politicians, experts, media, or the lay public), however, very few studies involved more than one layer. Results indicate that only a few studies made use of domain knowledge and a high proportion of the studies were not interdisciplinary. Those studies that made efforts to interpret the results demonstrated that the characteristics of political texts depend not only on the political position of their authors, but also on other often-overlooked factors. Ignoring these factors may lead to overly optimistic performance measures. Also, spurious results may be obtained when causal relations are inferred from textual data. Our paper provides arguments for the integration of explanatory and predictive modeling paradigms, and for a more interdisciplinary approach to polarization research.

Keywords Language polarization · Political polarization · Partisan language · Natural language processing · Text mining · Computational text analysis

✉ Renáta Németh
nemeth.renata@tatk.elte.hu

[1] Research Center for Computational Social Science, Faculty of Social Sciences, ELTE Eötvös Loránd University, Budapest, Hungary

Introduction

Language polarization – conceptualization and operationalization

As textual data sources grow in number and size, Natural Language Processing (NLP) is gaining ground in many social science subfields, including research on political polarization. In contrast to votes or polls, texts allow their authors to express a more nuanced opinion. Internet textual data reflect observed behavior as opposed to polls, and computational methods provide access to these vast amounts of data.

To review the most diverse approaches of polarization research, we wanted to refer to political polarizations in a broader sense. Indeed, characteristically different definitions of polarization can be found in the literature. Even one of the classic papers [1] treats polarization as multidimensional in character. According to the authors, polarization can be measured as (1) the dispersion of opinions, (2) the bi-modality of opinions, (3) the close association between different social attitudes, or (4) the correlation between social attitudes and salient individual characteristics. Lelkes [2] distinguishes two other forms: perceived polarization and affective polarization. A recent writing on the polarization of the digital sphere, Yarchi et al. [3], in addition to definitions already mentioned, distinguish interactional polarization that focuses on whether homophilic interactions are dominant over heterophilic ones (exploiting the network nature of the digital sphere). We included polarization not only around political ideologies, but also around public policy issues.

When searching for relevant studies, we also included "partisanship" as an alternative to "political polarization" in the search terms. Originally, partisanship was understood as affective or rational party identification [4], hence it could be causally linked to polarization, although it was not considered to be equivalent to it. Recently, however, many authors have been using partisanship and polarization as interchangeable concepts (e.g., [5]).

Turning to language polarization: Fiorina and Adams' [6] comprehensive paper is one of the seminal works on political polarization, with over 1000 citations, listing different kinds of empirical evidence used to study political polarization, but not listing linguistic features. Linguistic manifestations of political polarization (*language polarization* for short from now on) has entered the scientific discourse at a later point, mostly in the last decade. The term "political polarization" has been first used at the largest conference of computational linguists (Annual Meetings of the Association for Computational Linguistics) in 2012.

When measuring language polarization, one either tries to adapt existing measurement practices of political polarization to textual data, or they develop a new approach. The former solution is possible for several traditionally used polarization measures. For example, DiMaggio et al.'s [1] measures, that are based on the distribution of a numerical variable, can be adapted if the textual data can be converted into numerical data in some appropriate way. As we will see in the review, affective polarization [2] and interactional polarization [3] can be also

defined on textual data. In other cases, it is not possible to directly match the approach applied to texts with that applied to non-textual data, see e.g., topic choice, which, as the review shows, is often the focus of NLP analysis.

As the review will show, NLP methods most often introduce new ways for measuring polarization. The approaches differ according to the underlying conceptualization of political position and the data available. A focus of the review is on investigating whether researchers are indeed identifying traces of political polarization when they detect differences in language use between ideological sides—a question that arises in supervised classification.

The aims of our paper

We provide a methodological scoping review on how researchers have used NLP to study language polarization. We identify data sources and computational techniques, and review the different conceptualizations and operationalizations of polarization.

Our review captures several features that describes the difference between the research paradigms of social science and computer science-based NLP. These two approaches can be best identified as explanatory and predictive in nature [7, 8]. Whether the research is embedded in theory, whether qualitative approach is also used, whether the study address causality, or whether the results are interpreted, all can be linked to this duality. Social scientists traditionally prioritize explanations, invoking causal mechanisms derived from theory. However, computer scientists are more concerned with developing accurate predictive models, leaving interpretability aside. The degree of integration of the two approaches is also characterized in the review.

The overarching aim of this scoping review is to provide a starting point for future research by synthesizing approaches, potential flaws, and solutions. We would also like to give social science researchers who are unfamiliar with NLP a picture of the ongoing research.

The review's methodology

We decided to write a scoping review because of the multidisciplinary nature of the papers to be reviewed, and because of the differences in terminology across disciplines. We performed the searches using Google Scholar, and included studies published between January 1, 2010 and June 29, 2021. Our initial search terms were political "polariz(s)ation" AND "natural language processing", then added synonyms to both terms (for the detailed methodology of the review see the Supplement).

As is usually the case with scoping reviews, because of the undefined nature of the search terms, and because of searching anywhere within the article, we got many irrelevant hits, so a range of hard and soft exclusion criteria was defined. We synthesized our findings in a narrative report. Due to space limitations, technical details have been moved to the Supplementary Material.

Results

Summary

After deleting duplicates, our search returned 3078 unique hits, to which we added 6 more papers manually. Of the initial 3084 records, we identified 154 relevant studies (see Table S1 in the Supplement).

According to Fig. 1, the number of publications has risen during the past decade. Table 1 presents top 10 countries of the paper's focus, and countries the authors are affiliated with. (Full information can be found in the Supplement, Table S2.) As Table 1 presents, more than half of the studies were conducted using US data and the US is also the most publishing country affiliated with half of the studies. The US dominance is probably a phenomenon that manifests itself throughout the scientific literature, and most explicitly in the areas with a strong methodological basis. Similar results suggesting US dominance were found by recent bibliometric analysis on carbon emissions from transport sector [9], on effects of COVID-19 pandemic on mental health [10], or in a more technical area, on NLP in medicine [11], finding US dominance of 29%, 21% and 63%, respectively.

Only a tenth of the studies were single-authored. Figure 2 demonstrates the research collaboration among countries (full information can be found in Table S4 of the Supplement). The network shows co-authorship relations. The size of a node indicates the number of studies affiliated to the country, the width of the edges is proportional to the number of publications produced in collaboration, and colors

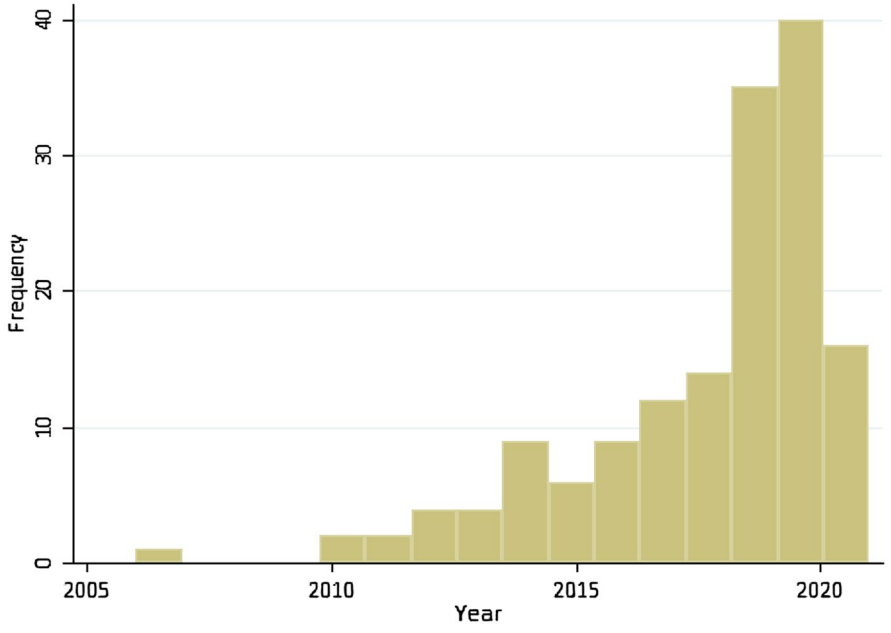

Fig. 1 The included studies by year of publication

Table 1 Top countries of the paper's focus, and top countries the authors are affiliated with (with frequency of occurrence)

Country of paper's focus		Affiliation country	
USA	91	USA	81
United Kingdom	9	Italy	14
Italy	6	United Kingdom	13
Spain	6	Germany	11
Canada	5	Qatar	10
Germany	5	Canada	8
India	3	Spain	7
Scotland	3	India	6
Turkey	3	Ireland	5
Ukraine	3	France	4

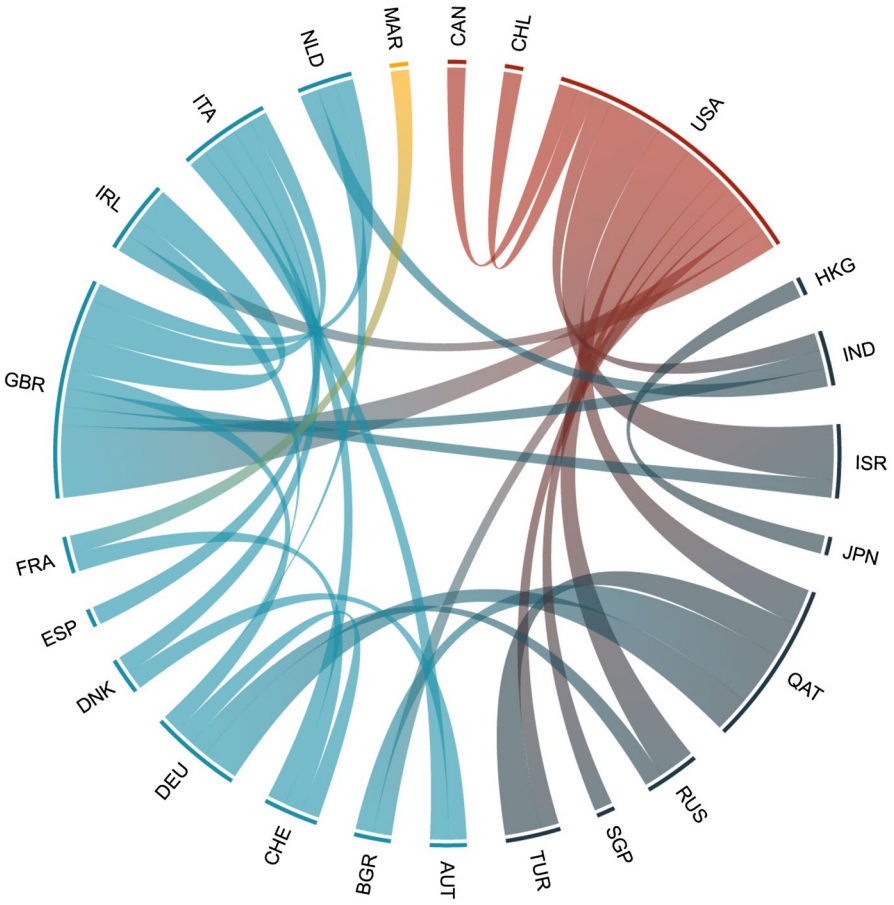

Fig. 2 Co-authorship network of countries

represent continents. Countries have been indicated by their corresponding three-letter abbreviation according to the ISO 3166 standard.

Figure 2 shows that in absolute numbers, the US leads in collaborations, GBR is second, and overall European and Asian countries are also very active. However, in terms of proportions (Table S4), it can be seen that 80% of US studies had no authors from other countries, while almost all of the non-US authors are collaborators, typically with a US co-author. These findings underline the US-centricity of the research topic.

Figure 3 represents co-authorship network of disciplines, its edges and nodes are defined in a similar way to Fig. 3, with disciplines instead of countries. Vertical edges at the bottom of the figure correspond to collaborations between authors from the same disciplines. (Table S4 and S5 in the Supplementary Material present full information and also describe the science classification method).

Two-thirds of the studies were assigned to Social Sciences and the same proportion to Technology. The number of non-interdisciplinary studies is very high: about one-fifth of the articles were written by social scientists only, and a further 40% (!) were written only with collaborators from the field of Technology. By merging Physical Sciences and Life Sciences & Medicine with Technology, we can conclude that 45% of the articles were co-authored exclusively from these areas, without subject matter collaborator.

As Table S1 shows, the range of sources is very diverse, with four fifths of the papers appearing in a source that is listed only once in the database, which indicates the multidisciplinary nature of the topic.

Figure 4 presents a word cloud of the studies' abstracts, the figure clearly shows the research focus.

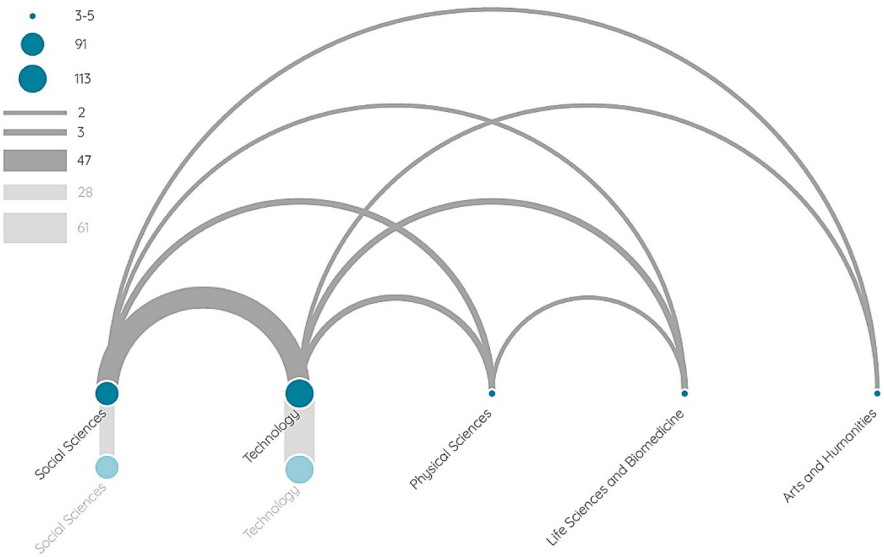

Fig. 3 Co-authorship network of disciplines

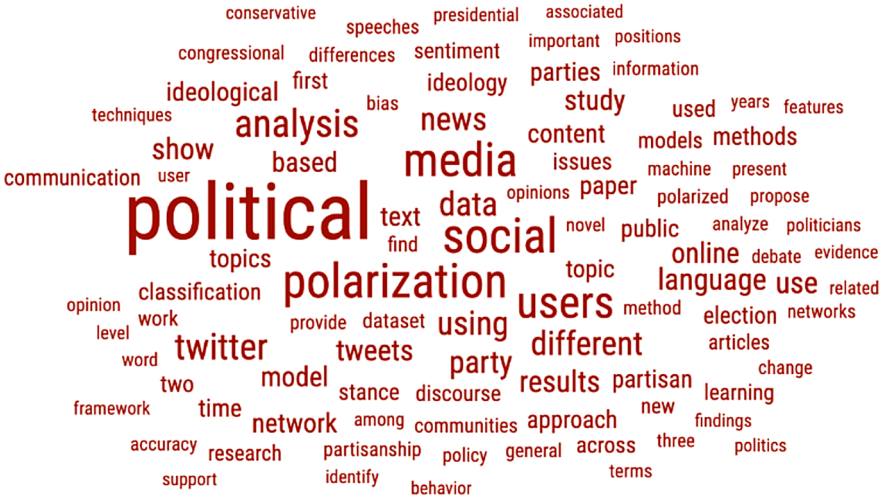

Fig. 4 Word cloud of the studies' abstract

Data

Data sources

This section describes the data sources that the studies used. Choice of data source has an impact on the end result, we therefore registered the type of data sources (see Table 2 for the top 10, the full list is in the Supplement, Table S3.). According to Table 2, more than 40% of the studies used Twitter data. We agree with a reviewer's comment that when interpreting this result, different aspects of data usability should be taken into account. Twitter is the most accessible source of politically relevant texts in most countries. It is also much less time-consuming and requires less technical expertise to use than other sources. However, in addition to accessibility, the fruitfulness of the data should also be taken account—there are research questions for which the use of other sources, such as parliamentary speeches, may be much more justified.

Data sources also differ in the time period they cover, with longer time periods allowing for changes to be detected (see Table 2). Longer time scales also provide a basis for benchmarking the degree of polarization detected in the present. For example, Jensen et al. [12] found that although the political discourse became more polarized in the late 1990s, polarization remained low relative to the late 19th, and much of the twentieth century. Tracing changes over time is also possible by measuring changes in polarization not over a 'common' historical time span, but over the course of individual lives. For example, Iliev et al. [13] examined the legislative rhetoric of US Senators as a function of their time in office.

Table 2 Top 10 data source, layers of political public sphere examined and length of period covered in the studies (with frequencies)

Type of data source		Which layers of the political public sphere are examined		Length of period covered	
Twitter	66	Lay only	63	Present data or period shorter than 2 years	105
Congressional/Parliament speeches	32	Official (politicians) only	50	2–5 years	9
News sites	22	Media only	19	5–10 years	11
Speeches outside Congress*	8	Expert only	5	10–20 years	8
Written public political documents**	7	Media and lay	4	20–50 years	10
Facebook	6	Official and media	4	50–100 years	2
Reddit	5	Official and lay	3	100 + years	9
Blogs	4	Official and expert	2		
Texts produced by non-political experts***	4	Official and media and expert	2		
Newspapers	3	Official and media and lay	1		

*Presidential/presidential candidates' speeches, campaign speeches, public statements, presidential candidacy announcements, **party manifestos, candidate manifestos, party platforms, press releases, coalition agreements, ***judges' written opinions, business reports, scientific papers)

Layers of political discourse

Language polarization can appear at *different layers of the political public sphere*, including the official channels of political communication (e.g., parliamentary speeches); the different types of media, and also the user-generated contents (e.g., social media,). In the following, we will make a distinction between professional (official sources and the press) and non-professional (lay) communication. A fourth layer emerged during our reviewing process: the expert layer (e.g., texts written by judges). In this domain, ideological expressions are considered inappropriate, so it is a particularly intriguing question whether ideology can be detected by computational tools.

According to Table 2, most of the papers were concerned with the lay public ($n = 71$); they all used only social media data. The majority of them researched tweets or posts; exceptions were KhudaBukhsh et al. [14], who studied the comments sections of YouTube channels, or Wu et al. [15], who analyzed Twitter bios.

Studies that were concerned with the official layer ($n = 60$) were mostly based on legislative speeches. Exceptions were, for example, Gross and Jankowski [16], who used a dataset of local party manifestos in Germany, or Wang and Tucker [17], whose dataset consisted of press releases issued by members of the US Congress.

The expert sphere was studied by a smaller number of studies ($n = 9$). Among them were Jelveh et al. [18], who detected latent ideological bias in academic papers in economics, Diaf et al. [19], whose dataset consisted of business cycle report sections issued by German economic research institutes, and Hausladen et al. [20], who studied the ideological direction of US Circuit Court decisions.

Studies very rarely involved more than one layer of the public sphere (only 9%). Serrano-Contreras et al. [21] studied comments made on YouTube videos uploaded by politicians. This approach could be generally used to examine the reactions of one layer to another. Karamshuk et al. [22] studied both the media and the lay public during the Ukrainian–Russian conflict of 2013–14, but the two layers were researched separately.

Similarities between the polarization of different layers (e.g., whether partisan terms diffuse from one to the other) was rarely investigated. An exception was Hofmann et al. [23], who measured differences between the language of the political parties and their media representation. Acree [24], going even further, compared the structure of the layers he studied and showed that ideology represented in the expert discourse is rich and varied, but professional political debate compresses ideological expression into a single (left–right) dimension. Yan et al. [25, 26] studied three different layers, the expert sphere (conservative and liberal wikis), alongside with the media and politicians. Their main research question was to what extent the polarization of the three layers differed [25] and how well the models can be transferred across domains [26]. In the latter, they showed that models based on the Congressional Record have some success in classifying articles from the media, that is, there is a diffusion process between the Congress and the media.

Key lessons learned from data

Some papers call attention to an often-overlooked problem: the importance of *the nature of the data*. The way the data were collected and filtered, the context in which they were created, and the genre of the texts are all important factors that determine, for example, the performance of the model, or whether a machine learning model can be transferred from one database to another (a problem called transferability, or cross-domain generalizability).

Cohen and Ruths [27] proved that using politically discriminative hashtags to define the corpus (which is a standard research practice) favors including an artificially enriched population of politically polarized users, who can therefore be more easily classified by the model. Their result also suggests that previously reported performances had been systematically overoptimistic. They also showed that classifiers cannot be used to classify users outside the range of political orientation on which they were trained.

In other words, *groups of people with different political activities* also use very different languages. Therefore the common implicit assumption according to which the most extreme cases exhibit the same phenomenon, only in a more detectable way, does not stand. E.g., Diermeier et al. [28], Morini et al. [29], and Grover et al. [30] follow this assumption implicitly when selecting more explicit/extreme cases, as well as Cotelo et al. [31], who only included in their analysis clearly codable tweets that were given the same political label by the coders.

Yan et al. [25] achieved a result similar to that of Cohen and Ruths [27] on three text corpora of *different genres from different public spheres*. They studied different layers of political discourse, and found that although the models perform well on within-dataset, their ability to generalize from one dataset to another is poor.

Another of their important results show that the success of prediction is not only domain-specific but also *time-specific*: effectiveness of prediction decreases as the test data is further removed in time from the training data. This raises questions about Diermeier et al.'s [28] methodology, whose training set consists of speeches in the 101st-107th Senates, and the test set comes from the 108th Senate.

Potthast et al. [32] made the important finding that the language of news sites is in fact not so much divided along right and left ideological lines, but rather along the *mainstream/hyper-partisan* dichotomy, i.e., along *stylistic lines*. Hirst et al. [33, 34] also detected poor transferability when classifying across two Canadian Parliaments (with the Liberal Party governing in the first case, and the Conservative Party in the second). It is partly this result that led them to the realization that *party status (being in opposition vs. being government)* is an often-overlooked factor when classifying texts by ideology.

Methods

Text analytic methods

For the benefit of readers who are not familiar with text analytic methods, in the Supplement, we give a brief intuitive outline of the methods most used in the studies. Some of them are specific to political science (Wordfish, Wordscores, Wordshoal), others are general, widely used NLP methods (topic model, word embedding, supervised machine learning, sentiment analysis).

Operationalization of polarization

When operationalizing polarization, measurement of political positions must first be determined, then, based on it, measurement of polarization must be defined. Typically, one of the following two approaches were used to define the political position of a given text: ideological scaling or classification.

Scaling was mostly done using standard political science approaches (Wordscores, Wordfish, or Wordshoal, $n=8$). Gross and Jankowski [16], using Wordfish on a dataset of local party manifestos in Germany, identified dimensions of party conflict. Medzihorsky et al. [35], using Wordfish, were able to show that the 2012 US Republican candidates moved farther away from the more traditional Republican ideology. Wordshoal was used, for example, by Goet [36].

Another approach to scaling comes from ideal point models (the DW-NOMINATE, [37], also belongs here). The models estimate political positions for legislators from legislative votes. The purpose of generalizing these models is that they can also incorporate texts. Nguyen et al. [38] introduced the hierarchical ideal point topic models, which uses not only votes but also associated bill text and the language of the legislators themselves and incorporates topic of the bills. Gerrish and Blei [39] develop the issue-adjusted ideal point model, which accounts for the contents of the bills. The idea is that the votes on a bill depend on a legislator's general position, adjusted for the bill's content.

When using scaling, degree of polarization can be approximated from the distribution of positions. In this context, an explicit measure is defined by Goet [36], which captures the consistency with which MPs fall within their party label across multiple policy issues. A score of "1" represents perfect polarization, with zero overlap between left and right wing parties.

Scaling has been used more by political science authors, while others used *classification* (51 studies used classification, mostly a supervised version). Classification in this context is a method that tries to identify the position of an author based on the words he or she has used. These studies, explicitly or implicitly, consider polarization as a classification problem. High classification performance suggests that language used on one political side is homogeneous, and different from language used on other sides. A polarization metrics can be defined: the greater the classification model's ability to identify the position of the author, the greater polarization there is. Works written most explicitly in this approach are Goet [36] and Green et al. [40]. Following this approach, Bayram et al. [41] analyzed House of Representatives floor speeches, and detected a clear upward trend in classification performance, indicating that polarization have become more obvious in the language of the speeches. In their highly cited paper, Gentzkow et al. [42], building on methods developed by Taddy [43, 44], recommend a polarization measure that follows a similar logic. They specify a multinomial model of speech with choice probabilities that vary by party, and polarization is measured by how easily an observer who knows the model can guess a speaker's party just from the speaker's choice of a single phrase. Kelly et al. [45] developed this method further.

In addition to scaling and supervised classification, other less common methods were also used. Samantray and Pin [46] used the ideological divergence indicator developed by Lelkes [2] that characterizes the level of polarization based on bimodality of the distribution of a numerical variable. The latter variable traditionally comes from opinion polls, here it is generated as a feature of texts. Darwish [47] used a polarization measure on Twitter data following Garimella et al. [48], measuring the amount of controversy from characteristics of the conversation graph. Budhiraja and Pal [49] represented each politician as a word embedding vector based on the content of their tweets, then identified polarization as the partitioning of the resulting point cloud by party. Villa-Cox et al. [50], generalizing KhudaBukhsh et al. [14], interpreted polarization through machine translation. Their framework assumes that two sub-communities are speaking in two different languages and obtains single-word translations. Number of disagreed word-pairs present a quantifiable measure of polarization. Other proposals have been also made to operationalize polarization, see Gross et al. [51], and Acree et al. [52].

Methods used for examining changes over time

Most studies investigating changes over time simply divide the time interval into several sections, and carry out the same analysis on each section separately. Other studies combined NLP with traditional time modeling statistical methods. For example, Tsur et al. [53] applied time series regression on topic affiliations resulting from the topic model. Gross and Jankowski [16] used linear mixed-effects models with

the dependent variable being the positions estimated by Wordfish. Hofmann et al. [23] employed time series modeling using generalized additive models to compare the lexical differences between parties.

There were studies using the structural topic model (STM), an NLP model that directly incorporates time. Farrell et al. [54], for example, analyzed texts written by US organizations about climate change over a 20-year period, and applied STM to examine how corporate funding ties influenced the change in topics over time.

Other authors, for example Gross et al. [51], Acree et al. [52], and Iliev et al. [13] developed their own statistical models to identify temporal changes in ideological positions.

Classification models

Unsupervised classification The main advantage of unsupervised classification is that it does not require any prior labeling of users, therefore there is no need for domain knowledge. Stefanov et al. [55] and Darwish et al. [56] used unsupervised user stance detection on Twitter. After projecting users onto a low-dimensional space, they applied clustering, which allowed them to find core users who were representative of the different stances. Another example for unsupervised classification is cluster analysis, applied for example by Giglietto et al. [57].

Supervised classification We have reviewed the relevant studies according to several criteria: nature of the target variable, way of annotation, use of structural information, and nature of classification features.

According to different definitions of political position, different target variables were used: stance [58], party [15], or ideology [59]. Specific target variable was used, for example, by Gerrish and Blei [39], who predicted the vote cast based on a legal text. The paper is unique in that it predicts the reaction to a text, not the author's ideological position. Cotelo et al. [31] used stance on two parties instead of using it on only one (with positive, negative or neutral stances on each) as target variables, and defined the task as classifying tweets into any of the nine combinatorial categories.

The creation of labels for supervised learning (or "annotation") is also an important issue, as obtaining labeled texts can itself be challenging. A *manually labeled* corpus to predict the ideological direction of US circuit court decisions was used by Hausladen et al. [20], with annotation being obviously a challenge in this case. If the annotation task is easy to teach, *crowdsourcing* can be used, like in the case of Wang and Tucker [17], who used Amazon's Mechanical Turk.

In many other cases, manual annotation is not necessary, because the labels can be obtained from external data. This is obviously the case with politicians' texts. A less trivial example was given by Jelveh et al. [18], who investigated academic articles written by American economists, and determined the author's political leaning based on their political campaign contributions and petition signing activities. Zubiaga et al. [60] classified the stance of Twitter users on the independence movement in Catalonia, and their labels relied on users' self-reported

territory that they claim to be citizens of, which is directly indicative of their stance toward the independence movement.

In other cases, the label is inherited from larger units of analysis by smaller ones. Kulkarni et al. [61] used the classification of news sources provided by AllSides.com (an American company that assesses the political leaning of prominent media outlets), and applied labels to articles according to their sources. Karamshuk et al. [22] applied manual labeling on news sources, and labeled Twitter users based on the news sources they shared. Rao et al. [62] continued the chain of inheritance even further, carrying over the labels from web domains to users, and then from users to other users based on retweet patterns. However, Kobayashi et al. [63] pointed out that the above inheritance-based solutions assume that Twitter users prefer to follow media and politicians whose ideological positions are similar to their own, and these assumptions are not necessarily true.

In addition to the texts, some papers also exploited the *structural information* on the database for classification. This is especially feasible for social media data, where a network of relationships between users is available. It had been commonly found that involving structural information increases classification accuracy [31, 64]. Wang et al. [65] classified tweets and found that the best model is one that integrates texts and *pictures*. However, it is worth noting that if the primary goal is to study language polarization (instead of defining the "best" classifier), the importance of language itself should be examined. What may be a relevant question, however, is the fraction of linguistic information that constitutes total polarization.

Another relevant question is *what features were used*, if text-based classification was applied. Most often a bag-of-words approach or n-grams were used, ignoring syntax [41, 52]. Newer models model the compositional aspect of language, e.g., by applying a neural network framework [61, 66].

Several papers used pre-defined text properties as features in addition to the raw text. Potthast et al. [32] conducted a study of hyper-partisan news, and employed different stylometric features, such as readability scores, ratios of quoted words, number of paragraphs, etc. Hashtags proved to improve classifiers' performance in several studies [64]. As hashtags are brief and information rich, they may reduce noise.

Word embedding was used for creating features in some studies [22, 60, 62, 65]. Zubiaga et al. [60] used word embedding for dimension reduction: word embedding representation of the content of a user's timeline was used as a feature. Karamshuk et al. [22] illustrate another important application of word embedding: their starting dictionary consisted of terms considered to be indicators of partisan rhetoric, and these words were then matched to the most similar ones according to their word embedding representation to get the final feature set.

Other studies used the output of a topic model [58, 62], author-level text features [27], or media-level features [67]. Baly et al. [67] illustrate how to utilize texts from a variety of different sources: the authors' aim was to classify the political ideology of news articles, and they determined media-level features based on the word embedding representation of (1) the bios of the medium's Twitter followers, and (2) the content of the Wikipedia page describing the medium, and (3) Web-traffic information about the medium's website.

Topic modeling

One in six studies applied topic modeling. The main topics and the words typical of the topics were often used for frame analysis [68]. According to Tsur et al. [53], topic co-occurrence can also approximate the way topics are framed. In other cases, see e.g., Sinno et al. [69], topic modeling was used only for technical reason to create a topically coherent corpus by omitting articles that were not related to the relevant topics.

Structural topic models are advantageous because they allow the incorporation of document metadata that affects how the topics vary by document. Including financial metadata allowed Farrell et al. [54] to test the effects of corporate funding on how organizations discuss climate change.

Others proposed the introduction of new types of topic models, e.g., Thonet et al. [70], Trabelsi and Zaiane [71], and Koylu et al. [72].

Sentiment analysis

One in six studies applied sentiment analysis. Most often *dictionary-based sentiment analysis* was used [20, 73, 74]. Grover et al. [30] for example examined moral, affective, and cognitive differences in language use between the two opposing sides of the debate over immigration in the US, using the LIWC dictionary. Among those using *non-dictionary methods* are Wang and Tucker [17], who used supervised machine learning models to assign sentiment scores to press releases.

Word embedding

One in eight studies applied embedding, mostly word embedding methods. An often-used application of word embedding was mentioned above in "Classification models", where it was used to extend the initial feature set with similar terms in meaning. This method was used in an unsupervised context as well [49] to get a polarization dictionary.

Another inspirational use of the model is exemplified by Brigadir et al. [75], or by Bonikowski et al. [76]: in their case word embedding vector space is used to detect changes in the meaning of certain characteristic terms. Brigadir et al. [75] considered changes in word semantics, both over time and between ideological positions. Bonikowski et al. [76] examined campaign speeches of US presidential candidates Classification models in 2016, by focusing on, for example, the word embedding neighborhood of the term "dangerous," which illustrates what the candidate views as the most pressing concerns (for example, whereas in case of Trump "refugees" is close to "dangerous" in meaning, in Clinton's case it is close to "prejudice").

KhudaBukhsh et al. [14] showed a powerful and interpretable application of word embedding, using word embedding-based machine translation on discussion section of YouTube channels of four prominent US news networks. They showed, that the two sub-communities of CNN and Fox News speak two different languages: what the former label for example, as "biden" and "kkk," the latter label as "creep" and "blm," respectively.

Finally, word embedding was also used for user clustering, e.g., Rashed et al. [77] represented users in an embedding space based on their texts, and the representations were then projected onto a lower dimensional space to which cluster analysis was applied.

Role of domain knowledge in the analysis

Recent studies, both academic and business, highlighted the importance of building domain knowledge into data science [78, 79]. Domain knowledge is important at each step of a data science project, including research question formulation, data collection, preprocessing, modeling, result interpretation and validation. Therefore, we reviewed whether the authors used domain knowledge at any stage of their research, for example whether they tried to interpret the classification results beyond the assessment of model performance.

If the analysis is specified along substantive questions domain, knowledge is needed to formulate these questions. Stecula and Merkley [80] for example displayed relevant subtleties in the analysis of framing climate change when they defined three types of frames based on previous research. Decadri and Boussalis [81] studied the link between populism/party membership and speech complexity in Italy, with substantive hypotheses, a dictionary of Italian populist rhetoric, and well-chosen metadata.

It is worth noting that in the case of the latter two, the studies were co-authored not only by computer scientist but by domain experts as well. Some examples show that if there are only computer scientists among the authors, the interpretation is often missing, and the research questions are rather technical [82]. And vice versa: if there are only social scientist co-authors, the methodology is fairly simple, although the paper is rich in content: the research questions are explained, the meta-variables are well-chosen, and the results are embedded in an existing scientific discourse [51, 52].

The role of domain knowledge is the most important one at the interpretation step, and especially in supervised classification. Without interpretation, predictive models are black boxes [83]. Understanding *why the model made a certain decision*, and finding the terms that are the most indicative of e.g., conservative versus liberal positions, bring us closer to the understanding of polarization, and help us position the results in the scientific discourse. Three out of five studies that used classification focused only on optimizing classification performance, and did not discuss which linguistic features played a role in the classification [65, 67, 84]. However, there were also articles that went through the interpretation in detail. Diermeier et al. [28] investigated the most distinctive linguistic features in the US Congress, with conclusions such as the one that cultural references are more important than economic references in distinguishing conservative from liberal speeches. Another example is Gentzkow et al. [42], who also very thoroughly went through the interpretation of partisan phrases.

In case of Hirst et al. [33, 34], the main lesson of the paper came from model interpretation: they observed that some of the most distinctive terms of the first analysis (with the Liberal Party as governor) “swapped sides” when turning to the

second analysis (with the Conservative Party as governor), which provided evidence that the classifier really picked up features that are related to the government/opposition dimension, instead of political ideology. In other words, the model's real functioning was revealed by the interpretation.

On the interpretability of classification models, it is worth mentioning Praet et al. [86] and Goet [36]. Praet et al. systematically explored the efficiency and interpretability of classification models. Their results showed a clear trade-off between interpretability and discriminative power, e.g., an expert-driven model showed the worst prediction and the best interpretability. According to Goet [36], classifiers ignore dimensionality (contrary to scaling methods), and when we use these models, we sacrifice our ability to make substantive claims about the drivers of polarization. We have seen, however, that the interpretation of the most distinctive words provides some answer to this question, if not as explicitly as in the case of scaling.

It is not only in the case of classification that the question of which words had an impact arises. Medzihorsky et al. [35] used Wordfish to follow the ideological shift of the Republican Party, and identified the most widely used terms and the extent to which they discriminate on the shift dimension. Rashed et al. [77], using cluster analysis, interpreted semantic differences between clusters based on their most prominent terms. Rumshisky et al. [85] gave a detailed interpretation of important "drifter" words that (according to word embedding) changed their meaning the most during the period under examination.

Domain knowledge is also important in the *validation phase* of studies, when the researcher judges the conceptual validity of the results. Goet [36] is an important reference as it provides criteria for evaluating text-based measures of polarization that can be easily followed in practice. According to his criteria, a valid text-based measure, for example, should correspond well to our a-priori expectations, e.g., outliers in our estimates should reflect what we know historically about polarization in the given context. However, validation is completely missing from many of the reviewed papers. Some studies using topic modeling included a validation phase, where the researchers qualitatively assessed the effectiveness of the models [54, 87]. Another way of validation was used in classification studies for example by Taddy [43], Diermeier et al. [27], Bayram et al. [41], Rashed et al. [77], and Gerrish and Blei [39]. Their approach was *to detect the outlier/mispredicted cases*, and to examine them qualitatively. For example, Gerrish and Blei [39], when studying the text of US bills to predict the lawmakers' vote, detected Rep. Ron Paul as an outlier, and found that he voted more conservatively than expected on healthcare issues. Validation also involves the most critical question that arises when approaching polarization with supervised classification, namely whether ideological differences really underlie the perceived linguistic differences. We will return to this when we discuss causality.

The use of qualitative methods

In examining use of qualitative methods, it should be noted that studies using only qualitative text analysis or simple quantitative approach were excluded from the review, so our question was precisely whether a qualitative approach was used in

addition to NLP. We have included this issue in our analysis because social research using mixed methods is generally considered to be more valid: it allows for a deeper examination of the phenomenon, supports generalization, increases confidence that findings are not driven by a particular method, and aids interpretation.

Most of the studies reviewed did not use qualitative methods. Some of those that did, used it in the *validation* phase. For example, the validation of topic models requires considerable qualitative work, as the researchers can only assess the interpretability and the effectiveness of the models by actually reading the most relevant texts representing each topic [54, 87]. Yarchi et al. [3] presented a new method that combines NLP with manual content analysis to understand the topics. Taddy [88], on the other hand, is a counterexample, where the model is completely detached from the text, with the analysis not going back to the text during the interpretation of the topics.

Other studies applied a qualitative approach to support *interpretation:* Rho et al. [89] for example used discourse analysis to analyze all comments that contain the top relevant terms computationally detected beforehand. Similarly, Grover et al. [30] carried out a qualitative analysis of tweets containing terms that were found to be important by LIWC analysis.

"*Close reading*," i.e., a thoughtful interpretation of texts, was explicitly mentioned as the method used by Bonikowski et al. [76] and Sinno et al. [69], for example. In the latter work, close reading was used to understand the motivations behind annotators' decisions.

Dornschneider and Todd [90] was among the few studies in which a *qualitative approach dominated*. They conducted interviews and used sentiment analysis combined with qualitative discourse analysis. In the work of Budak et al. [91], machine learning was used only for technical reasons, to identify the relevant documents, and the essence of the research was given by a close reading of the articles.

Addressing causality during the NLP analysis

Though the problem of causation has been extensively studied in empirical social science, it is often neglected in text classification, even though erroneous causal inference does not only concern interpretation, but also compromises the robustness of models.

The most basic causal approach is to refer to timely prevention. E.g., Jensen et al. [12] found that polarized phrases increase in frequency in Google Books before their use increases in congressional speeches. Although they emphasized that causal inference is beyond the scope of their paper, they suggest that their finding is consistent with an autonomous effect of elite discourse on congressional speech.

The most common (even if not explicitly stated) causal problem that arises in the study of ideological polarization is the question of potential confounders, i.e., whether ideological differences are indeed the cause of the detected differences in language use. Very few studies have explicitly mentioned or tried to address this problem. Lin et al. [92] investigated the bitterlemons.org website, which published articles on issues related to the Israeli–Palestinian conflict on a weekly basis with an Israeli and Palestinian editor and guest. The authors applied

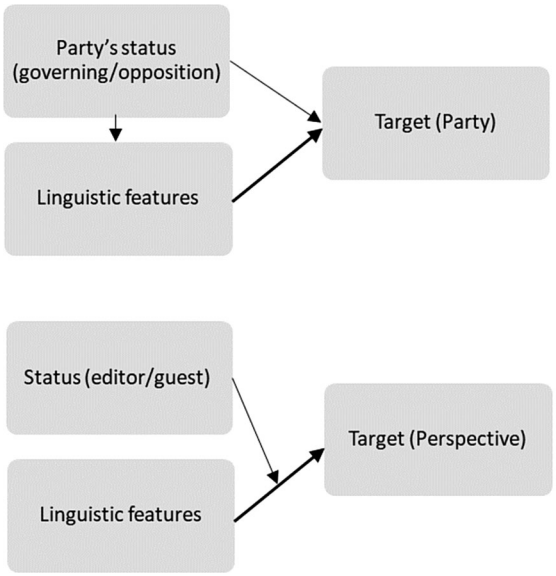

Fig. 5 Classification with the presence of a confounder/a moderator. Moderators lie on the causal pathway (bold line), while confounders do not

supervised classification to predict perspective (Israeli or Palestinian). As perspective was better predicted on editors, the authors suggested that there might exist differences between the writing styles of Israeli and Palestinian editors, and it is this that the models had found, not political perspectives. To test this hypothesis, they conducted experiments in which they trained their algorithm on editors and tested it on guests, and vice versa.

Hirst et al. [33, 34] discussed a similar analytic issue. They classified party affiliation across two Canadian Parliaments and revealed that what their models were sensitive to was not expressions of ideology but rather expressions of attack and defense used in a position of opposition and in a position of governing. That is, according to their result, party status (opposition/government) is a confounder when classifying parties.

Both Lin et al. [92] and Hirst et al. [33, 34] suspected a third variable that influenced both the dependent variable and the independent variable, causing a spurious association, but their suspicion was only confirmed in the latter case. In the former case, an effect modification was revealed: only the models' performance level differed depending on the editor/guest role (the moderator variable), but the same linguistic features distinguished perspectives within both roles. The logic of confounding and effect modification is illustrated in Fig. 5 below.

Very few articles have applied recent approaches to causality. The exceptions include Landeiro and Culotta [93], who use the statistical framework of Judea Pearl, and Widmer et al. [94], who apply the instrumental-variables framework.

Conclusions

Our review has its limitations. "Political polarization" includes a wide range of definitions, and how "natural language processing" is referred to also varies by disciplines. Although we have attempted to capture these concepts in several different ways, we may have still missed some relevant papers.

Of the initial 3084 hits, we identified 154 relevant studies. The number of papers has risen during the past decade. Most studies focused on the US ($n = 91$), and the cross-national validity of their results has rarely been tested. About 40% ($n = 66$) utilized Twitter data, and one in three studies employed supervised machine learning for predicting ideology/stance.

Some studies demonstrated that the characteristics of political texts depend not only on the political position of their authors, but also on other often-overlooked factors that are not independent of the former (such as the authors' political engagement, whether their party is in a governing position or in opposition, the texts' style, genre, or date). Ignoring these factors during data collection, or distinguishing texts based solely on political position, can lead to serious errors, such as overly optimistic association measures, and confounded associations. Some studies suggested that the same issue lies also behind transferability problems, i.e., the corpora that are assumed to be homogeneous are in fact different.

Although the number of studies has grown rapidly in the last years, only a minority of them used domain knowledge to gain insights. Those that did, showed the need for expert interpretation at different points in the analysis (interpreting the most important features, detecting outliers, comparing different models, etc.).

Most studies did not employ the method of close reading, and did not discuss potential problems arising if causal inference is made on textual data. High proportions of studies were non-interdisciplinary in the sense that they were authored without subject matter involvement (45%), or, conversely, they were authored only by social scientists (20%). These observations may be a consequence of the institutionalization of computational research methods outside the social sciences.

However, we have found several inspiring examples for methodological approaches without these shortcomings. We have seen combined use of several types of databases: ones with extra-textual information like user structure, those with metadata of texts, and those that combine polling data with social media data. We have seen that research cover different layers of the political public sphere (politicians, experts, media, or lay public). However, very few studies involved more than one layer, although it can be used to study diffusion processes between the layers.

Many of the points raised by the review are likely to apply to the use of NLP in the social sciences in general, and not just to the study of political polarization. These may include the infrequent use of domain knowledge and mixed methods, the frequent lack of interpretation, the low number of interdisciplinary papers, or the two often conflicting aspects of the data, ease of access and fruitfulness.

Arguments have arisen in recent years that fields adapting artificial intelligence are facing a reproducibility crisis [95]. Among the papers pointing to the causes

of the crisis, there were quite a few [5, 96] that argued that the meeting of computer science and applied sciences is more than just adapting large data repositories and tools to analyze them. This meeting also represents a convergence of different fields with different methodological paradigms, and the quality of the research depends on how these paradigms are productively integrated.

We can conclude that the potential of NLP in political polarization research is very high indeed. However, our paper also provides arguments for the integration of explanatory and predictive modeling paradigms, and for a more interdisciplinary approach to polarization research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-022-00196-2.

Acknowledgements This work was supported by NKFIH (National Research, Development and Innovation Office, Hungary) grant K-134428. ELTE Eötvös Loránd University, Budapest, Hungary supported the proofreading of the paper. Open access funding provided by Eötvös Loránd University. The author would like to thank Ildikó Barna and Judit Szabó for their contribution to data collection and Eszter Katona for visualizing the data.

Funding Open access funding provided by Eötvös Loránd University.

Data availability List of studies included in the review can be found in the supplementary information file.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. DiMaggio, P., Evans, J., & Bryson, B. (1996). Have American's social attitudes become more polarized? *American Journal of Sociology, 102*(3), 690–755. https://doi.org/10.1086/230995
2. Lelkes, Y. (2016). Mass Polarization: Manifestations and Measurements. *Public Opinion Quartely, 80*(S1), 392–410. https://doi.org/10.1093/poq/nfw005
3. Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2021). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication, 38*(1–2), 98–139. https://doi.org/10.1080/10584609.2020.1785067
4. Carius-Munz, L. M. (2020). Partisanship: Conceptualizations and consequences. In H. Oscarsson & S. Holmberg (Eds.), *Research Handbook on Political Partisanship* (pp. 47–59). Edward Elgar Publishing.
5. Demszky, D., Garg, N., Voigt, R., Zou, J., Gentzkow, M., Shapiro, J., Jurafsky, D (2019) Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, MN, USA.

6. Fiorina, M. P., & Abrams, S. J. (2008). Political polarization in the American public. *Annual Review of Political Science (Palo Alto), 11*, 563–588. https://doi.org/10.1146/annurev.polisci.11.053106.153836
7. Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature, 595*(7866), 181–188.
8. Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–215.
9. Tian, X., Geng, Y., Zhong, S., Wilson, J., Gao, C., Chen, W., Yu, Z., & Hao, H. (2018). A bibliometric analysis on trends and characters of carbon emissions from transport sector. *Transportation Research Part D, 59*, 1–10. https://doi.org/10.1016/j.trd.2017.12.009
10. Akintunde, T. Y., Musa, T. H., Musa, H. H., Musa, I. H., Chen, S., Ibrahim, E., Tassang, A. E., & Helmy, M. S. E. D. M. (2021). Bibliometric analysis of global scientific literature on effects of COVID-19 pandemic on mental health. *Asian Journal of Psychiatry*. https://doi.org/10.1016/j.ajp.2021.102753
11. Wang, J., Deng, H., Liu, B., Hu, A., Liang, J., Fan, L., Zheng, X., Wang, T., & Lei, J. (2020). Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on pubMed. *Journal of Medical Internet Research, 22*(1), e168161. https://doi.org/10.2196/16816
12. Jensen, J., Kaplan, E., Naidu, S., & Wilse-Samson, L. (2012). Political polarization and the dynamics of political language: evidence from 130 years of partisan speech. *Brookings Papers on Economic Activity, 2012*, 1–81. https://doi.org/10.1353/eca.2012.0017
13. Iliev, I. R., Huang, X., & Gel, Y. R. (2019). Political rhetoric through the lens of non-parametric statistics: are our legislators that different? *Journal of the Royal Statistical Society: Series A (Statistics in Society), 182*(2), 583–604. https://doi.org/10.1111/rssa.12421
14. KhudaBukhsh, A. R., Sarkar, R., Kamlet, M. S., & Mitchell, T. M. (2020). We don't speak the same language: interpreting polarization through machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*. https://doi.org/10.48550/ARXIV.2010.02339
15. Wu, P.Y., Mebane, W.R., Woods, L., Klaver, J. & Duek, P. (2019). Partisan Associations of Twitter Users Based on Their Self-descriptions and Word Embeddings. In: Paper prepared for presentation at the 2019 Annual Meeting of the American Political Science Association, Washington, DC
16. Gross, M., & Jankowski, M. (2020). Dimensions of political conflict and party positions in multi-level democracies: evidence from the local manifesto project. *West European Politics, 43*(1), 74–101. https://doi.org/10.1080/01402382.2019.1602816
17. Wang, R. T., & Tucker, P. D. (2021). How partisanship influences what congress says online and how they say it. *American Political Research, 49*(1), 76–90. https://doi.org/10.1177/1532673x20939498
18. Jelveh, Z., Kogut, B., Naidu, S. (2014). Detecting Latent Ideology in Expert Text: Evidence From Academic Papers in Economics. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar. pp. 1804–1809.
19. Diaf, S., Döpke, J., Fritsche, U., & Rockenbach, I. (2022). Sharks and minnows in a shoal of words: Measuring latent ideological positions based on text mining techniques. *European Journal of Political Economy*. https://doi.org/10.1016/j.ejpoleco.2022.102179
20. Hausladen, C. I., Schubert, M. H., & Ash, E. (2020). Text classification of ideological direction in judicial opinions. *International Review of Law and Economics, 62*, 105903. https://doi.org/10.1016/j.irle.2020.105903
21. Serrano-Contreras, I.-J., García-Marín, J., & Luengo, Ó. G. (2020). Measuring online political dialogue: does polarization trigger more deliberation? *Media and Communication., 8*(4), 63–72.
22. Karamshuk, D., Lokot, T., Pryymak, O., & Sastry, N. (2016). *Identifying partisan slant in news articles and twitter during political crises*. Lecture Notes in Computer ScienceCham: Springer International Publishing.
23. Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J., Wissik, T. (2020). Comparing Lexical Usage in Political Discourse across Diachronic Corpora. In: Proceedings of ParlaCLARIN II Workshop of the Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020 European Language Resources Association (ELRA). pp: 58–65.

24. Acree, B. (2016). Deep learning and ideological rhetoric PhD dissertation, College of Arts and Sciences Department of Political Science The University of North Carolina at Chapel Hill University Libraries. https://doi.org/10.17615/mm0p-jk38
25. Yan, H., Lavoi, A., Das, S. (2017). The Perils of Classifying Political Orientation From Text. LINKDEM@ IJCAI. Retrieved June 29, 2021 from http://ceur-ws.org/Vol-1897/paper3.pdf
26. Yan, H., Das, S., Lavoie, A., Li, S. & Sinclair, B. 2019. The congressional classification challenge: domain specificity and partisan intensity. In: Proceedings of the 2019 ACM Conference on Economics and Computation, 71–89. https://doi.org/10.1145/3328526.3329582
27. Cohen, R., Ruths, D. (2013). Classifying Political Orientation on Twitter: It's Not Easy!, in: Proceedings of the the 7th International AAAI Conference on Weblogs and Social Media (ICWSM-13). Cambridge, Massachusetts USA. pp. 91–99. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/14434
28. Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2012). Language and ideology in Congress. *British Journal of Political Science, 42*(1), 31–55. https://doi.org/10.1017/s0007123411000160
29. Morini, V., Pollacci, L., Rossetti, G. (2020). Capturing Political Polarization of Reddit Submissions in the Trump Era. In: Paper presented at SEBD 2020, June 21–24, 2020, Villasimius, Italy.
30. Grover, T., Bayraktaroglu, E., Mark, G., & Rho, E. H. R. (2019). Moral and affective differences in US. immigration policy debate on twitter. *Computer Supportive Cooperative Work, 28*, 317–355. https://doi.org/10.1007/s10606-019-09357-w
31. Cotelo, J. M., Cruz, F. L., Enríquez, F., & Troyano, J. A. (2016). Tweet categorization by combining content and structural knowledge. *Information Fusion, 31*, 54–64. https://doi.org/10.1016/j.inffus.2016.01.002
32. Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, pp. 231–240.
33. Hirst, G., Riabinin Y., Graham, J. (2010). Party status as a confound in the automatic classification of political speech by ideology, in: Bolasco, S., Chiari, I., Giuliano, L. (Eds.) Proceedings of 10th International Conference, Journées d'Analyse statistique des Données Textuelles, 9–11 June 2010 - Sapienza University of Rome. Retrieved online from https://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-0731-0742_137-Hirst.pdf
34. Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., & Morris, C. (2014). Text to ideology or text to party status? In B. Kaal, I. Maks, & A. van Elfrinkhof (Eds.), *From text to political positions: Text analysis across disciplines.* Amsterdam: John Benjamins Publishing Company.
35. Medzihorsky, J., Littvay, L., & Jenne, E. K. (2014). Has the tea party era radicalized the republican party? Evidence from text analysis of the 2008 and 2012 republican primary debates. *PS Political Science & Politics, 47*(4), 806–812. https://doi.org/10.1017/s1049096514001085
36. Goet, N.D., (2017). Measuring polarization with text analysis: Evidence from the UK House of Commons, 1811–2015. In: Paper prepared for the Polarization, Institutional Design and the Future of Representative Democracy workshop, Berlin, Harnack Haus.
37. McCarty, N. M., Poole, K. T., & Rosenthal, H. (2006). *Polarized America: the dance of ideology and unequal riches*. MIT Press.
38. Nguyen, V.A., Boyd-Graber, J., Resnik & P. Miler, K. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1438–14
39. Gerrish, S.M., Blei, D.M. (2012). How They Vote: Issue-Adjusted Models of Legislative Behavior. In: Advances in Neural Information Processing Systems 25 (NIPS 2012).
40. Green, J., Edgerton, J., Naftel, D., Shoub, K., Cranmer, S.J. (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. Science Advances 6 (28). https://doi.org/10.1126/sciadv.abc2717
41. Bayram, U., Pestian, J., Santel, D., Minai, A.A. (2019). What's in a word? Detecting partisan affiliation from word use in congressional speeches. In: Paper presented at the International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19th July 2019.

42. Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica, 87*(4), 1307–1340. https://doi.org/10.3982/ecta16566
43. Taddy, M. (2013). Multinomial inverse regression for text analysis. *Journal of American Statistical Association, 108*(503), 755–770. https://doi.org/10.1080/01621459.2012.734168
44. Taddy, M. (2015). Distributed multinomial regression. *Annals of Applied Statistics, 9*(3), 1394–1414. https://doi.org/10.1214/15-aoas831
45. Kelly, B., Manela, A., & Moreira, A. (2021). Text selection. *Journal of Business and Economic Statistics, 39*(4), 859–879. https://doi.org/10.1080/07350015.2021.1947843
46. Samantray, A., & Pin, P. (2019). Credibility of climate change denial in social media. *Palgrave Commun*. https://doi.org/10.1057/s41599-019-0344-4
47. Darwish, K. (2019). Quantifying Polarization on Twitter: The Kavanaugh Nomination. *SocInfo 2019 Social Informatics Lecture Notes in Computer Science.* Cham: Springer.
48. Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying controversy on social media. *ACM Transactions on Social Computing, 1*(1), 3.
49. Budhiraja, A., Pal, J. (2020). Twitter and political culture: Short text embeddings as a window into political fragmentation. In: Proceedings of the 3rd ACM SIGCAS Conference on Computing and Sustainable Societies, 335–336. https://doi.org/10.1145/3378393.3402276.
50. Villa-Cox, R., KhudaBukhsh, A.R. & Carley, K.M. 2021. Exploring Polarization of Users Behavior on Twitter During the 2019 South American Protests. arXiv preprint on arXiv:2104.05611
51. Gross, J., Acree, B., Sim, Y., Smith, N.A. (2013). Testing the Etch-a-Sketch Hypothesis: A Computational Analysis of Mitt Romney's Ideological Makeover During the 2012 Primary vs. General Elections. APSA 2013 Annual Meeting Paper, American Political Science Association 2013 Annual Meeting, Available at SSRN: https://ssrn.com/abstract=2299991
52. Acree, B. D. L., Gross, J. H., Smith, N. A., Sim, Y., & Boydstun, A. E. (2020). Etch-a-sketching: Evaluating the post-primary rhetorical moderation hypothesis. *American Politics Research, 8*(1), 99–131. https://doi.org/10.1177/1532673x18800017
53. Tsur, O., Calacci & D., Lazer, D. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, Beijing, China, pp. 1629–1638.
54. Farrell, J. (2016). Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences of the United States of America, 113*(1), 92–97. https://doi.org/10.1073/pnas.1509433112
55. Stefanov, P., Darwish, K., Atanasov, A., Nakov, P. (2019). Predicting the topical stance of media and popular Twitter users. arXiv preprint on arXiv: 1907. 01260.https://doi.org/10.48550/ARXIV.1907.01260
56. Darwish, K., Stefanov, P., Aupetit, M., & Nakov, P. (2020). Unsupervised User Stance Detection on Twitter. In: Proceedings of the 14th International AAAI Conference on Web and Social Media. pp. 141–152. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/7286
57. Giglietto, F., Iannelli, L., Rossi, L., Valeriani, A., Righetti, N., Carabini, F., Marino, G., Usai, S., & Zurovac, E. (2018). Mapping Italian news media political coverage in the lead-up of 2018 general election. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3179930
58. Fang, A., Ounis, I., Habel, P., Macdonald, C., Limsopatham, N. (2015). Topic-centric classification of twitter user's political orientation, in: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, New York, USA. pp. 791–794. .https://doi.org/10.1145/2766462.2767833
59. Ademmer, E., Stöhr, T. (2019). The making of a new cleavage? Evidence from social media debates about migration. Kiel Working Papers No. 2140, Kiel University
60. Zubiaga, A., Wang, B., Liakata, M., & Procter, R. (2018). Political Homophily in IndependenceMovements: Analysing and Classifying SocialMedia Users by National Identity. arXiv.1702.08388
61. Kulkarni, V., Ye, J., Skiena, S., & Wang, W. Y. (2018). Multi-view models for political ideology detection of news articles. *arXiv Preprint on arXiv 1809 03485*. https://doi.org/10.48550/ARXIV.1809.03485

62. Rao, A., Morstatter, F., Hu, M., Chen, E., Burghardt, K., Ferrara, E., & Lerman, K. (2020). Political partisanship and anti-science attitudes in online discussions about covid-19. *arXiv Preprint on arXiv 2011 08498*. https://doi.org/10.48550/ARXIV.2011.08498
63. Kobayashi, T., Ogawa, Y., Suzuki, T., & Yamamoto, H. (2019). News audience fragmentation in the Japanese Twittersphere. *Asian Journal of Communication, 29*(3), 274–290. https://doi.org/10.1080/01292986.2018.1458326
64. Conover, M.D., Goncalves, B., Ratkiewicz, J., Flammini, A. & Menczer, F. (2011). Predicting the political alignment of twitter users. In: Proceedings of the 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing. 9–11 October, 2011, Boston, MA, USA.
65. Wang, Y., Feng, Y., Hong, Z., Berger, R., Luo, J. (2017). How polarized have we become? A multimodal classification of Trump followers and Clinton followers. arXiv Preprint on arXiv: 1711.00617. https://doi.org/10.48550/ARXIV.1711.00617
66. Belcastro, L., Cantini, R., Marozzo, F., Talia, D., & Trunfio, P. (2020). Learning political polarization on social media using neural networks. *IEEE Access, 8*, 47177–47187. https://doi.org/10.1109/access.2020.2978950
67. Baly, R., Martino, G. D. S., Glass, J., & Nakov, P. (2020). We can detect your bias: Predicting the political ideology of news articles. *arXiv Preprint on arXiv:2010.05338*. https://doi.org/10.4550/ARXIV.2010.05338
68. Shen, Q., Rosé, C.P. (2019). The Discourse of Online Content Moderation: Investigating Polarized User Responses to Changes in Reddit's Quarantine Policy. In: Proceedings of the Third Workshop on Abusive Language Online, pp. 58–69. Association for Computational Linguistics, Florence, Italy, August 1, 2019.
69. Sinno, B., Oviedo, B., Atwell, K., Alikhani, M., & Li, J. J. (2021). Political ideology and polarization of policy positions: A multi-dimensional approach. *arXiv Preprint on arXiv:2106.14387*. https://doi.org/10.48550/ARXIV.2106.14387
70. Thonet, T., Cabanac, G., Boughanem, M., Pinel-Sauvagnat, K. (2017). Users are known by the company they keep: Topic models for viewpoint discovery in social networks. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM, New York, NY, USA. pp. 87–96.https://doi.org/10.1145/3132847.3132897
71. Trabelsi, A. & Zaiane, O. (2018). Unsupervised Model for Topic Viewpoint Discovery in Online Debates Leveraging Author Interactions. In: Proceedings of the International AAAI Conference on Web and Social Media, 12. Retrieved from https://ojs.aaai.org/index.php/ICWSM/article/view/15021
72. Koylu, C., Larson, R., Dietrich, B. J., & Lee, K.-P. (2019). CarSenToGram: Geovisual text analytics for exploring spatiotemporal variation in public discourse on Twitter. *Cartography and Geographic Information Science, 46*(1), 57–71. https://doi.org/10.1080/15230406.2018.1570343
73. Coutto, T. (2020). Half-full or half-empty? Framing of UK–EU relations during the Brexit referendum campaign. *Journal of European Integration, 42*(5), 695–713. https://doi.org/10.1080/07036337.2020.1792465
74. Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2019). Polarization and acculturation in US election 2016 outcomes–Can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change, 145*, 438–460. https://doi.org/10.1016/j.techfore.2018.09.009
75. Brigadir, I., Greene, D., Cunningham, P. (2015). Analyzing discourse communities with distributional semantic models. In: Proceedings of the ACM Web Science Conference. ACM, New York, NY, USA.
76. Bonikowski, B., Feinstein, Y., Bock, S. (2019). The Polarization of Nationalist Cleavages and the 2016 U.S. Presidential Election. In: Paper presented at The UCR Political Economy Seminar, April 12, 2019. https://ucrpoliticaleconomy.ucr.edu/wp-content/uploads/2019/04/Bonikowski-Feinstein-and-Bock-Polarization-of-Nationalist-Cleavages-UC-Riverside.pdf.
77. Rashed, A., Kutlu, M., Darwish, K., Elsayed, T., & Bayrak, C. (2020). Embeddings-based clustering for target specific stances: the case of a polarized Turkey. *arXiv preprint on arXiv:2005.09649*. https://doi.org/10.4550/ARXIV.2005.09649
78. Liu, J., & Zhang, X. (2019). The role of domain knowledge in document selection from search results: the role of domain knowledge in document selection from search results. *Journal of the Association for Information Science and Technology, 70*(11), 1236–1247. https://doi.org/10.1002/asi.24199

79. Deng, C., Ji, X., Rainey, C., Zhang, J., & Lu, W. (2020). Integrating machine learning with human knowledge. *iScience, 23*(11), 101656. https://doi.org/10.1016/j.isci.2020.101656
80. Stecula, D. A., & Merkley, E. (2019). Framing climate change: economics, ideology, and uncertainty in American news media content from 1988 to 2014. *Frontiers in Communication*. https://doi.org/10.3389/fcomm.2019.00006
81. Decadri, S., & Boussalis, C. (2020). Populism, party membership, and language complexity in the Italian chamber of deputies. *Jornal of Elections, Public Opinion and Parties, 30*(4), 484–503. https://doi.org/10.1080/17457289.2019.1593182
82. Tucker, E. C., Capps, C. J., & Shamir, L. (2020). A data science approach to 138 years of congressional speeches. *Heliyon, 6*, e04417. https://doi.org/10.1016/j.heliyon.2020.e04417
83. Molnar, C. (2019). *Interpretable Machine Learning*. Morrisville, NC: Lulu.com.
84. Hemphill, L., Culotta, A., & Heston, M. (2016). #Polar scores: measuring partisanship using social media content. *Journal of Information Technology and Politics, 13*(4), 365–377. https://doi.org/10.1080/19331681.2016.1214093
85. Rumshisky, A., Gronas, M., Potash, P., Dubov, M., Romanov, A., Kulshreshtha, S. & Gribov, A. 2017. Combining network and language indicators for tracking conflict intensity. In: Ciampaglia, G., Mashhadi, A., Yasseri, T. (Eds.), Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13–15, 2017, Proceedings, Part II. as part of Lecture Notes in Computer Science 10540, Springer International Publishing, Cham, pp. 391–404. https://doi.org/10.1007/978-3-319-67256-4_31
86. Praet, S., Van Aelst, P., Daelemans, W., Kreutz, T., Peeters, J., Walgrave, S., & Martens, D. (2021). Comparing automated content analysis methods to distinguish issue communication by political parties on twitter. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3782027
87. Guntuku, S. C., Purtle, J., Meisel, Z. F., Merchant, R. M., & Agarwal, A. (2021). Partisan differences in twitter language among US legislators during the COVID-19 pandemic: Cross-sectional study. *Journal of Medical Internet Research, 23*(6), e27300. https://doi.org/10.2196/27300
88. Taddy, M. (2012). On Estimation and Selection for Topic Models, in: proceedings of the Fifteenth International conference on artificial intelligence and statistics. *PMLR, 22*, 1184–1193.
89. Rho, E. H. R., Mark, G., & Mazmanian, M. (2018). Fostering civil discourse online: Linguistic behavior in comments of #MeToo articles across political perspectives. *Proceedings of the ACM of Human-Computer Interaction, 2*, 1–28. https://doi.org/10.1145/3274416
90. Dornschneider, S., & Todd, J. (2020). Everyday sentiment among unionists and nationalists in a Northern Irish town. *Irish Political Studies, 36*(2), 185–213. https://doi.org/10.1080/07907184.2020.1743023
91. Budak, C., Goel, S., & Rao, J. M. (2016). Fair and balanced? Quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly, 80*(S1), 250–271. https://doi.org/10.1093/poq/nfw007
92. Lin, W.-H., Wilson, T., Wiebe, J., Hauptmann, A. (2006). Which side are you on? Identifying perspectives at the document and sentence levels. In: Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL-X), June 2006. Association for Computational Linguistics, New York City, NY, USA. pp. 109–116.
93. Landeiro, V., & Culotta, A. (2018). Robust text classification under confounding shift. *Journal of Artificial Intelligent Research, 63*, 391–419. https://doi.org/10.1613/jair.1.11248
94. Widmer, P., Galletta, S., & Ash, E. (2020). Media slant is contagious. *Center for Law and Economics Working Paper Series 14/2020*. https://doi.org/10.3929/ethz-b-000454192
95. Driggs, D., Selby, I., Roberts, M., Gkrania-Klotsas, E., Rudd, J. H., Yang, G., et al. (2021). Machine learning for COVID-19 diagnosis and prognostication: lessons for amplifying the signal while reducing the noise. *Radiology. Artificial intelligence*, *3*(4), https://doi.org/10.1148/ryai.2021210011.
96. Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). *The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning* (pp. 335–348). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3514094.3534196.

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