Praet et al. *EPJ Data Science* (2021) 10:50
https://doi.org/10.1140/epjds/s13688-021-00305-7

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Predictive modeling to study lifestyle politics with Facebook likes

Stiene Praet[1*], Peter Van Aelst[2], Patrick van Erkel[2], Stephan Van der Veeken[1] and David Martens[1]

*Correspondence:
stiene.praet@uantwerp.be
[1] Department of Engineering Management, University of Antwerp, Prinsstraat 13, 2000, Antwerp, Belgium
Full list of author information is available at the end of the article

Abstract

"Lifestyle politics" suggests that political and ideological opinions are strongly connected to our consumption choices, music and food taste, cultural preferences, and other aspects of our daily lives. With the growing political polarization this idea has become all the more relevant to a wide range of social scientists. Empirical research in this domain, however, is confronted with an impractical challenge; this type of detailed information on people's lifestyle is very difficult to operationalize, and extremely time consuming and costly to query in a survey. A potential valuable alternative data source to capture these values and lifestyle choices is social media data. In this study, we explore the value of Facebook "like" data to complement traditional survey data to study lifestyle politics. We collect a unique dataset of Facebook likes and survey data of more than 6500 participants in Belgium, a fragmented multi-party system. Based on both types of data, we infer the political and ideological preference of our respondents. The results indicate that non-political Facebook likes are indicative of political preference and are useful to describe voters in terms of common interests, cultural preferences, and lifestyle features. This shows that social media data can be a valuable complement to traditional survey data to study lifestyle politics.

Keywords: Data science; Predictive modeling; Political preference; Facebook likes

1 Introduction

Voting is central to the democratic process and the well-functioning of our political system. Therefore, voting behavior and party preference are well-studied by a broad range of social scientists who try to explain how and why decisions are made by the electorate. Traditionally, to understand individuals' votes, scholars relied on socio-structural factors such as group identification, religious affiliation, and socio-economic status. Electoral pioneers from the Michigan and Columbia school both stressed the importance of these long-term factors in explaining voting behavior [1, 2]. In the US context party identity was key, while in many European countries, more religious and class-based cleavages were driving voters [3]. As these stable factors started losing importance, scholars gradually started devoting more attention to short-term factors such as concrete issues and popular candidates [4, 5]. Studies showed that a better educated electorate became more volatile and more affected by issue priorities [6] and candidate evaluations [7].

"Lifestyle politics" represents an alternative way to understand political preference and voting behavior. Authors such as Bennett [8] and Giddens [9] argue that in our post-modern society, personal identity is replacing collective identity and that individuals increasingly let their personal politics depend on lifestyle choices. Politics focuses more and more on broad issues like identity, values, and moral and social orientation [9]. According to DellaPosta et al. [10] (p. 1474) we "are increasingly likely to find our local communities and social networks populated by individuals with similar aesthetic tastes, leisure activities, consumer preferences, moral practices, and ways of life." Our attitudes on climate change, religion, migration, sexual minorities or workers' welfare are strongly connected to our consumption choices and daily lifestyle [11]; in turn, lifestyle is becoming more important in understanding individuals' political opinions and voting behaviors. In recent studies, political preference has been associated with leisure activities and personal tastes [10], brand choice in supermarkets [12], sustainable behaviors [13], food and music taste [11, 14] and movie preferences [15]. These lifestyle choices reflect deep-seated personality traits that not only drive the culture wars over issues like race, or sexuality but also influence how citizens think about classic economic issues [16]

Although the linkage between politics and many aspects of our personal life were documented, what is driving this alignment between ideological preferences and seemingly unrelated lifestyle dimensions is less clear. While some suggest that these lifestyle choices are a way to express one's (political) identity [15], others rather stress the deeper role of social and cultural homogeneity [10]. Yet others point out the political relevance of seemingly apolitical domains like entertainment news, novels, films, television sitcoms, and popular culture more broadly to distribute news and stimulate political debate [17]. Although a clear connection between lifestyle and political preference is not carved in stone, politicians seem well aware of their lifestyle choices when acting in the public sphere. For instance, studies show that politicians prefer certain types of food on the campaign trail, in particular to relate to the common man or to show cultural sensitivity. In this way politicians, subtly present or frame themselves in line with the cultural preferences of their potential voters [18]. Finally, "branded communication" refers to the idea to attach political messages to so-called lifestyle brands to capture the attention of individuals, and by which the alignment between ideological preferences and lifestyle is further strengthened [19].

The growing political polarization renewed the issue of "lifestyle enclaves" as a prime interest for a wide range of social scientists [20, 21]. Furthermore, we know relatively little on whether these "cultural fault lines" that are clearly visible in the two party system in the US [14] are as present and determining in a multi-party system with a much broader variety of ideological players. Yet, detailed information on individuals' lifestyles is very difficult to collect, which complicates empirical and comparative studies in this domain. The options of leisure activities, movies, music, cultural activities, etc. that could be included in survey questions on lifestyle politics are endless, therefore, including these type of questions will not only occupy a lot of survey space and time but will also result in non-exhaustive option lists. Moreover, it implies that researchers are able to deductively make a selection of the lifestyle indicators (i.e. cultural places, movies, products) that should be included to best explain or predict vote choice.

A potential valuable alternative data source to capture these cultural and lifestyle practices is social media data. Today, thanks to the Internet and social media, an unseen amount and granularity of data are available. People visiting webpages or liking Facebook content leave little "bread crumbs" behind in the digital world that are indicative of their interests and personality. From this fine-grained behavioral data, inferring unknown information about a user is possible by applying predictive modeling techniques [22]. We explore the potential of Facebook "like" data for capturing lifestyle and predicting political and ideological preference. With Facebook "likes" we refer to the mechanism used by Facebook users to express their positive association with public Facebook pages of products, sport clubs, musicians, books, restaurants, etc.

Digital trace data (such as Facebook likes) seem to offer valuable insights in political attitudes and behavior [23], and several scholars have combined survey data with digital traces to study and understand political behavior. For example, Bond and Messing [24] estimate the ideology of politicians and their supporters using individual citizens' Facebook likes of political figures to study the relationship between ideology and age, social relationships and ideology, and the degree of polarization among the electorate. Eady et al. [25] apply a method developed by Barberá [26], to quantify the ideological distributions of users' online political and media environments on Twitter and study the extent to which liberals and conservatives live in so-called "echo-chambers". Some other examples include comparing individual exposure to news and politics content [27] and understanding news sharing behavior [28, 29].

Our study adds to this existing work by using non-political Facebook pages to study lifestyle politics. As most previous literature mainly focused on the inference of ideology based on the accounts or pages from political actors or news media, the inclusion of non-political Facebook pages is innovative.

To explore the value of digital trace data to complement classical survey research, we gathered Facebook likes and survey data of more than 6500 participants in Flanders, the Dutch-speaking part of Belgium. The data gathering was of course done with user consent and a clear privacy statement. Based on these Facebook likes, we build models to predict political and ideological preference, and compare this to predictive models based on traditional survey data. The interpretation of these models can be used to gain insights into voter profiles. The main contributions of this study are twofold. First, we explore the use of Facebook like data to complement and improve traditional survey data to study lifestyle politics, with an explicit focus on the contribution of non-political likes. Second, we collect a unique dataset of Facebook likes and survey data to gain insights into voter profiles in Belgium, a fragmented multi-party system.

2 Predictive modeling with Facebook likes

Facebook likes[1] express a certain positive attitude or interest in a public Facebook page, similar to visiting a webpage, purchase behavior, payment or location data. For example, observing a users' Facebook likes related to books provides similar information to watching someone's book closet or a list of purchased books online. Liking a Facebook page signals a user's desire to see more posts from the page's publisher. Moreover users can actively interact with the page creator as well as other people that have liked the page, by commenting, liking, and sharing content [30].

[1] We refer to *Public Page Likes*, i.e. the public Facebook pages that a user likes and that show up as being liked in the About section of that person's profile (see https://www.facebook.com/help/171378103323792?helpref=uf_permalink). We do not include likes or emotional reactions (Love, Haha, Wow, Sad, and Angry) to Facebook posts.

Additionally, a Facebook like has been conceptualized as a form of social endorsement, since these likes are publicly observable by friends in one's network [24]. Since like behavior is observable, it is used as a form of self-expression, and in line with the theory of Goffman [31] it is believed that users are building an idealized version of themselves on social platforms. At the same time, this "ideal" self should remain congruent with how one is perceived in the offline world, as Facebook networks are often grounded in offline relationships [32]. Indeed, in a study on consumer identity, Hollenbeck [33] finds that users like Facebook pages (of brands) to present—an ideal version of—themselves on the platform. Given these characteristics, Facebook likes have been proven powerful to infer interests and psychological traits, and became invaluable for user profiling and personalized advertising applications [34]. More concretely, Piazza et al. [35] investigated the relationship between Facebook likes and individual lifestyle, and they found a correlation between the activity, interests, and opinions of an individual and their like information. Given these characteristics, we believe Facebook likes can be used to operationalize lifestyle preferences.

We argue that Facebook likes could be a valuable addition to survey data for electoral research because of at least four different reasons: (1) through the direct measurement of actual behavior we avoid social desirability—though this might, to some extent, also be present in like behavior, see below—, recall error, and subjectivity[2] that are specific to answering survey questions [37], (2) it requires less effort from the respondents, and thus it is less time-consuming and costly than collecting survey data, (3) they can provide an unfiltered look and unique information on interests and lifestyle that we cannot grasp (fully) with survey-based data and (4) they can inductively help to identify the most important indicators of a phenomenon (such as lifestyle politics in the case of our study), which can then later be included in a survey.[3] Social media data allow to measure real behavior rather than self-indicated behavior or attitudes.

Of course, we must keep in mind that online data is a proxy for the real behavior under study, and thus, drawing causal inference and generalizing results from this type of data must be performed with caution [38, 39]. For example, what you like on Facebook is not necessarily what you like in real-life and thus true behavior can only be studied indirectly [39]. More specifically, two potential hidden biases need to be considered. First, Facebook like behavior may be affected by user induced biases such as social desirability and intentional misrepresentation. Second, Facebook profile data are affected by the mechanics of the platform, such as the personalization by Facebook algorithms [40]. The pages that you like are at least partially influenced by what is recommended to you. These recommendations are based on your previous behavior on the platform, your profile attributes, and who you are connected to [40]. Therefore, connections between lifestyle and political preference are possibly reinforced owing to algorithmic curation and recommendation systems. Furthermore, behavioral data can suffer from exhibiting a low signal-to-noise ratio, since the behavior we capture can be unrelated to the target question. Surveys, if

[2] For example, [36] linked original survey data with respondents' observed social media data to validate self-reports of political activity; and discovered a substantial discrepancy between objective and self-reported posting behavior, which could be due to a subjective interpretation of what is considered to be "political".

[3] Although this was not the focus of this research, the last three arguments could also be applied to the context of experimental—instead of survey—data, and show how Facebook likes could be a valuable addition to experimental data for electoral research. Additionally, behavior observed in designed experiments is difficult to generalize to the real world due to the artificial conditions the subjects are put in, while social media data emerge from user behavior in a more natural setting.

properly designed and implemented, provide better quality controls and allow for more targeted questions and responses [41]. A last issue with behavioral data is privacy. Since these data are often very personal and sensitive, the data must be collected and stored with respect for users' privacy, addressing challenges on user consent, data anonymization, secure storage, etc. [42].[4] Yet, because of the enormous amount of information available in behavioral data, there is much to learn from it when following a rigorous research approach [38, 39].

We will use a predictive modeling approach to study Facebook likes and political preference. Although most empirical political science research relies on explanatory modeling to test causal explanations [43], predictive modeling is more suitable to uncover complex patterns from data that might lead to the generation of new hypotheses [44]. The distinction between explanatory and predictive modeling exhibits some practical implications to each step of the modeling process [45].

At the very essence, the scientific goals that they are aimed at are different: explaining or predicting. Explanatory modeling is theory-driven: given a causal theoretical model, statistical models are applied to observational data to test hypotheses. In contrast, in predictive modeling statistical models are used for generating good predictions of new values, irrespective of the true underlying causal relationship [45]. Since the goal of prediction and explanation is different, they lead to different modeling approaches. Explanatory modeling requires interpretable statistical models; however, this is not a strict requirement for good prediction, although interpretability of the model is desired in our case. Moreover, some methods exist that are very useful for prediction, but not for explanation. When building prediction models, we will make use of regularization: an extra constraint introduced to the optimization function that penalizes the weights of coefficients. This is especially useful when modeling Facebook likes: because they are sparse data, simply optimizing the standard objective function would lead to estimates that are not unique and can overfit the data.

Finally, validating a causal model involves goodness-of-fit tests (e.g., normality tests) and model diagnostics such as residual analysis. Prediction performance is evaluated on a part of the data that was not used for training the model (a holdout- or test set). The evaluation metric we will use is Area Under the ROC Curve (AUC). AUC is a frequently used metric in data science to measure the performance of a classification model, independent of the frequency of the classes. It can be interpreted as the probability that the model ranks a random positive example higher than a random negative example [46]. A perfect model would achieve an AUC of 100%, while an AUC of 50% indicates a random model.

3 Data collection

This study is part of a broader research project that focuses on (social) media and electoral preferences, and was approved by the Ethics Committee for the Social Sciences and Humanities (EA SHW), at the University of Antwerp. The data collection started in March 2018 and focused on Flanders, the Dutch-speaking part of Belgium, representing around 60% of the population. Belgium is a federal country with a fragmented multiparty system and large coalition governments. A detailed survey with questions on sociodemographics, media consumption, political preference and attitudes was sent to 4500 respondents online. Surveying was done by Dynata/SSI, who distributed the survey among

[4] We elaborate on our research design with regard to these issues in Appendix A.1.5.

their own online panel, which consists of a diverse collection of citizens in terms of age, gender, and educational level.[5] Targets were set on these socio-demographic characteristics to increase the representativeness of the sample.[6] Of these respondents, 524 agreed to provide us access to their Facebook like data via Facebook Login (See Appendix A.1.4). In May–June 2018, a second round of data collection was conducted, where we disseminated a shorter survey and Facebook Login through the online webpages of popular Flemish newspapers. We asked people to use our tool, which would predict their ideological position based on their personal Facebook likes. An additional 6209 respondents agreed to provide access to their Facebook data, and they completed 12 survey questions about their media consumption and political preference. The Facebook likes and survey questions were collected with user consent and stored anonymously and securely on a local server. The data are used for scientific purposes only and will under no circumstances be shared with other institutions or companies. Results will be shown on an aggregate level only and participants have the right to stop their collaboration at any time and ask for their data to be removed. More details about the data collection and a full discussion of the privacy and ethical concerns can be found in Appendix A.1.

Per user, we stored the name and timestamp of all public Facebook pages they liked. This resulted in a total of 595,994 unique Facebook pages. For privacy reasons (see Appendix A.1.5), only pages that are liked by a minimum of 50 respondents in our dataset will be reported later in the analysis, which results in 5357 pages. For these pages we searched for the category that was assigned to them on Facebook. In total, Facebook shows over 1300 different page categories.[7] We started from the categorization that Facebook uses to classify public pages, but we refined and adjusted it into 20 categories (see Table 1). Note that our categories are not mutually exclusive; for example, an art festival can be included in the category *Arts & Culture* as well as in the category *Event & Festival*.

The survey data (see Appendix A.1.2) include some of the basic variables that are generally included in models of voting behavior in Belgium, see for instance [47, 48]. Specifically, we added several of the most important structural determinants of ideology and voting behavior (gender, age and education), as well as the use of different media sources (TV, newspaper, social media, etc.), interest in the news (sports, culture, home affairs, foreign affairs, etc.), and general interest in politics.

Two survey questions will be used as target variables: (a) *Ideological leaning*. The participants positioned themselves on a scale of 0 (most left) to 10 (most right). In the analysis we consider the numbers 0 to 3 as "left", 4,5 and 6 as "center", and numbers 7 and higher as "right", and (b) *Party preference*. The participants indicated how likely they are to ever vote for each of the seven main Flemish political parties on a scale of 1 (never vote for party) to 10 (definitely vote for party): the worker's party (PVDA), the green party (Groen), the social democratic party (Sp.a), the Christian democratic party (CD&V), the liberal party (Open VLD), the Flemish nationalist party (N-VA), and the extreme right party (Vlaams Belang, VB). The preferred party per participant is the one with the highest score. In case of a tie between two or more parties, all tie-parties are considered equally important. The number of participants per target variable can be found in Table 4 in the Appendix.

[5] For more information on Dynata and their panel see https://www.dynata.com/.

[6] With regard to age our sample of 4500 respondents is slightly older than the population.

[7] An overview of all possible Facebook categories can be found on https://www.facebook.com/pages/category/.

Table 1 Description and number of pages for the 20 Facebook categories

Category	Description	# pages
Music	Music, bands, producers, record labels, albums etc.	835
Companies & Business	Companies, entrepreneurs, stores, shops etc.	833
Communities	Communities, interests and places	824
Apps, Websites & Blogs	Apps, websites and blogs	511
Products & Services	Products, brands, financial services, marketing etc.	487
Artists & Public Figures	Artists and public figures	414
News & Media	News, media, radio, magazines etc.	396
Tv Shows	TV shows and episodes	349
Games, Humor & Entertainment	Games, humor, amusement, comedy etc.	316
Sports & Health	Sports, athletes, gym, health	250
Civil Society	Nonprofit organizations, labor unions and religious oranizations	249
Politics	Politicians, political parties and government organizations	229
Food, Drinks & Restaurants	Food, cooking, restaurants, breweries etc.	228
Movies	Movies, films, actors and cinema	212
Arts & Culture	Arts, culture, photography, museums etc.	193
Events & Festivals	Events, festivals and concerts	166
Books & Authors	Books, libraries, publishers, writers	109
Bars, Cafes & Night clubs	Bars, cafes, pubs, clubs etc.	108
School, University & Education	Schools, universities, student organizations and education	99
Travel	Travel, tour agencies and tourism	92

Our sample of Facebook users consists of a diverse mix of users in terms of gender, age and education levels. Yet, our sample contains less females, less participants older than 55 and less lower educated participants compared to the general population. Weights were applied to our survey samples (see Appendix A.1.3) but this did not influence the results,[8] therefore we will only report the unweighted results hereafter. Moreover, through the self-selection of participants, Facebook users with a higher political interest are overrepresented. Likewise, some parties are over- or underrepresented in our sample. However, since the goal of our study is not to predict aggregated election results but rather to gain insights into voter groups, this is not a particular stumbling block in this study. Nevertheless, it is likely that we will achieve more accurate results for political parties that are well-represented in our sample.

4 Methods

As discussed in Sect. 2, our methodology consists of predictive modeling to study lifestyle politics. We apply predictive models to infer the political and ideological preference of our respondents, based on their Facebook likes.[9] The goal of these models is to optimize prediction accuracy at the individual level. Yet, the coefficients of these models reveal which likes are predictive for a certain political preference, providing insight in the interconnection between lifestyle and politics. To show Facebook likes are not just capturing socio-demographics in an indirect way, we compare the models built on Facebook likes to models built on survey data and demonstrate the potential added value of combining the predicted Facebook likes models with traditional survey data.

More specifically, we compare the predictive performance of Facebook like data to the survey data in five different set-ups: models based on all Facebook likes (M1), models based solely on non-political Facebook likes (M2), models based on the survey data (M3),

[8] The spearman rank correlation between the coefficients of the models with and without survey weights applied was 0.94

[9] Facebook pages are encoded as dummy variables where the value 1 indicates the user has liked the page and 0 indicates the user has not liked the page.

Table 2 Overview of the modeling set-ups with a description of the different datasets and the number of variables. In the description we refer to the survey questions in Appendix A.1.2

Data	Description	# Variables
Facebook likes (M1)	Public Facebook page likes	595,994
Facebook likes* (M2)	Non-political public Facebook page likes, liked more than 50 times	5128
Survey (M3)	Use of media sources (Q1), interest in news topics (Q2), interest in politics (Q5) Gender (Q6), age (Q7), education level (Q8)	27
Target variables	Political leaning (Q3) and party preference (Q4)	10

models based on a combination of all Facebook data and survey data (M4), and models based on a combination of non-political Facebook likes and survey data (M5) (see Table 2). As explained in Sect. 2, prediction models are evaluated based on out-of-sample predictive accuracy. We report the average AUC over ten folds.[10]

Political leaning is divided into three classes (left, center, and right, see Sect. 3). For each of the three classes, we transfer the variable into a binary classification problem (one-vs-all) and train three binary Logistic Regression (LR) models.[11,12] Next, we calculate the weighted average AUC for the three classes. Similarly, we built seven binary classifiers for the seven political parties and calculate weighted average AUC. Though multinomial LR (one-vs-one) could also be used in this case (and is in fact more common in electoral research), we prefer binary classification to be able to create general distinct profiles for each party electorate in comparison to all other citizens, rather than compare the voter profiles to only one single reference group. However, both methods (binary and multinominal LR) show similar predictive performance in our analysis.

Additionally, we want to know how predictive each category of pages is for political leaning. In other words, how accurate can political leaning be predicted when using only the Facebook likes of the concerning category. This tells us which aspects of our social lives are most related to politics. To report the predictive performance of each category independent of the amount of pages per category, we randomly sample (with replacement) 100 pages per category and use only those pages as features.[13] This procedure (of random sampling) is repeated 10 times, and the average AUC is reported.

Finally, we compare the insights in ideological position and party preference based on Facebook likes versus survey data. Analyzing the coefficients of a logistic regression and their "p-values" is common practice in traditional explanatory and predictive modeling with dense data [50]. However, when using a regularization term, assumptions about the asymptotic distribution of parameters do not apply, and therefore, different methods for significance testing are needed and suggested in literature [51, 52]. We will follow the bootstrap procedure as described in [52]. From the original dataset we take a random sample with replacement and built a model from this dataset to estimate the betas. This step is repeated 1000 times to obtain 1000 values for each beta. For each beta we estimate the

[10] This means that we train the model on 90% of the data and evaluate on an unseen 10% of the data, this procedure is repeated ten times with different parts of the data.

[11] We used the scikit-learn implementation for logistic regression [49].

[12] The classifiers were trained using 5-fold cross validation to optimize the regularization penalty ($L1$ or $L2$) and the optimal regularization value C in $[0.001, 0.01, 0.1, 1, 10]$. The data is split into five folds, for the different parameters a logistic regression is trained on four folds and the parameters that resulted in the highest predictive performance on the test fold are selected.

[13] Because we do not have all categories of the less frequent pages available, we will only include pages with 50 likes or more.

Table 3 AUC and the standard deviation (std) for models build on (non-political) Facebook pages, survey data or a combination thereof.

		Facebook (M1)		Facebook* (M2)		Survey (M3)		Combined (M4)		Combined* (M5)	
		AUC	Std	AUC	Std	AUC	Std	AUC	Std	AUC	Std
Leaning	Left	83%	2%	75%	3%	71%	3%	84%	2%	81%	1%
	Center	64%	2%	62%	3%	62%	2%	66%	2%	64%	2%
	Right	81%	3%	75%	3%	70%	4%	81%	3%	80%	2%
	Weighted avergage	73%	2%	68%	2%	65%	2%	74%	2%	72%	1%
Party preference	PVDA	79%	3%	73%	3%	66%	3%	79%	2%	77%	2%
	Sp.a	70%	4%	70%	2%	61%	2%	69%	3%	65%	4%
	Groen	79%	1%	71%	1%	69%	1%	79%	1%	76%	1%
	CD&V	73%	3%	68%	3%	58%	3%	72%	2%	66%	3%
	Open VLD	77%	4%	73%	3%	67%	3%	77%	4%	74%	4%
	N-VA	84%	3%	80%	2%	70%	2%	84%	3%	80%	2%
	Vlaams Belang	86%	5%	76%	7%	76%	7%	85%	7%	83%	8%
	Weighted average	78%	1%	73%	1%	67%	1%	78%	1%	74%	1%

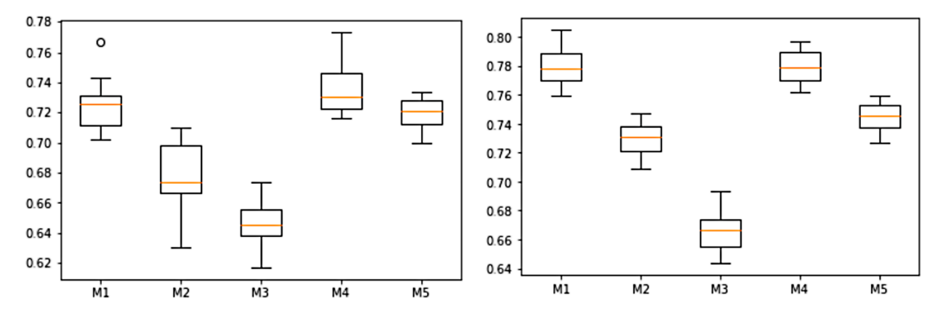


Figure 1 Boxplot of predictive performance for the five models (M1)–(M5) for ideological leaning (left) and party preference (right)

probability density function using a Gaussian kernel[14] to calculate the probability (p-value) that the parameter is less than or equal to zero. Facebook pages will be ranked based on the mean coefficients over 1.000 bootstraps while the p-value indicates significance on a $\alpha = 0.05$ level.

5 Results

Table 3 compares the predictive performance for models based on Facebook likes and survey data for all users that participated in our study (M1)–(M5). Figure 1 shows the boxplots for the predictive performance of these models. In Appendix A.2 we perform a corrected paired differences t-test [55] and a non-parametric Wilcoxon signed-ranks test [56] (based on 10-fold cross-validation results) to compare the AUC of the different models and apply the Bonferonni correction for multiple pairwise comparison [57, 58].

Both for ideological leaning and party preference, nearly all pairwise differences between the models are significant at the $\alpha = 0.05$ level,[15] see Appendix A.2. The Facebook likes (M1) exhibit a higher predictive power than the survey data (M3). Still, the survey ques-

[14] We used gaussian_kde from scipy.stats [53] with the default Scott's Rule [54] for bandwidth selection.

[15] with the Bonferonni correction this is reduced to $\alpha/10$.

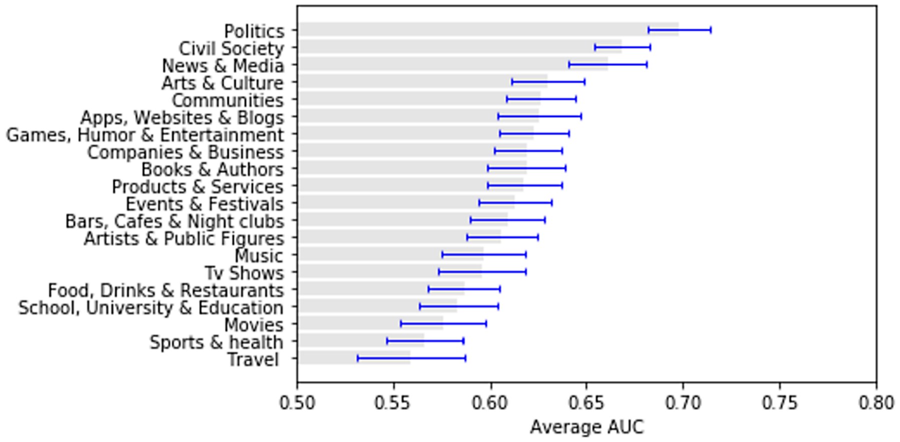
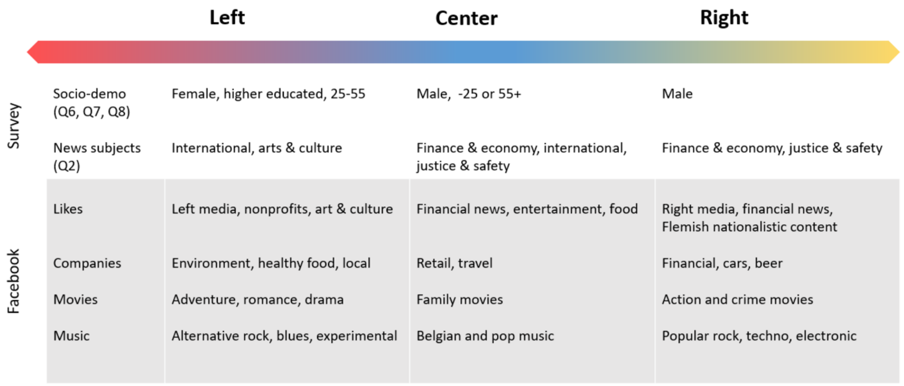


Figure 2 Average AUC (+/- 1 std) for the Facebook categories with target "left"

tions do seem to capture some information that Facebook likes do not, since the combined model (M4) achieves the highest predictive power. One might argue that the high predictive performance of Facebook likes is due to the presence of political Facebook pages. Naturally, if someone likes the Facebook page of a certain party or politician this is very indicative of their political preference. Therefore, to assess the predictive value present in non-political Facebook pages, we built the same prediction models on the Facebook data without political pages (M2). As expected, the predictive performance decreases when excluding explicit political content. However, the AUC is still higher than the survey data (M3). Moreover, when combining non-political likes and survey data (M5), the performance is higher than for the survey data alone, which indicates that purely non-political lifestyle choices can augment surveys in predicting political preferences.

The high predictive performance of non-political Facebook likes is an indication that non-political interests are interrelated with political preferences. The fact that this performance is higher than for the available survey data indicates that indeed additional information is captured with Facebook data that was not covered in the survey questions. To further explore which lifestyle categories are connected to political preference, we calculate the predictive performance per Facebook category. Figure 2 shows which categories of Facebook pages are most predictive for ideological preference. A prediction model built on only political Facebook pages achieves, not surprisingly, the highest AUC, followed closely by *Civil Society* and *News & Media*, which could be considered as semi-political categories. However, also the non-political categories of *Arts & Culture* and *Communities* are very telling for ideological preference. Conversely, the categories *Movies*, *Sports & Health* and *Travel* are the least predictive for ideological preference. This indicates that some aspects of our social lives are more tied to our ideological views than others.

Finally, we note the non-political likes perform better at discriminating between left versus non-left people and between right versus non-right than between center versus non-center (see Table 6 in Appendix for the t-test results). This might indicate that a less clear pattern is present in the characteristics and behavior of people with a center political leaning and that they exhibit a less distinguishable profile. Similarly, voters for some parties can be classified more accurately than others (e.g. the party N-VA, see Table 7 in the Appendix).



		Left	Center	Right
Survey	Socio-demo (Q6, Q7, Q8)	Female, higher educated, 25-55	Male, -25 or 55+	Male
	News subjects (Q2)	International, arts & culture	Finance & economy, international, justice & safety	Finance & economy, justice & safety
Facebook	Likes	Left media, nonprofits, art & culture	Financial news, entertainment, food	Right media, financial news, Flemish nationalistic content
	Companies	Environment, healthy food, local	Retail, travel	Financial, cars, beer
	Movies	Adventure, romance, drama	Family movies	Action and crime movies
	Music	Alternative rock, blues, experimental	Belgian and pop music	Popular rock, techno, electronic

Figure 3 Some examples to illustrate the insights we can gain into voter profiles based on Facebook data and based on survey data

From the survey data (see Table 11 in the Appendix), we learn that traditional variables, such as gender, age, education, and interest in certain topics help to predict people's vote. For instance, left voters are often highly educated women, between the age of 25 and 55, and interested in news about culture, arts, and international politics. Conversely, right voters are more often male than female and generally demonstrate a strong interest for financial and economic news. What additional insights do Facebook likes provide? To answer this question, we examine the most related pages when excluding political pages (see Table 10 in the Appendix and Fig. 3 for a summary).

With Facebook likes, analyzing cultural taste and differences in lifestyle between left and right voters is possible at a very detailed level. The pages most related to a left political leaning are (alternative) media outlets that are considered to be leaning more toward the left, and nonprofit organizations for climate and human rights, which can be considered as outspoken left-wing topics. For example, all else equal, the odds of demonstrating a left political ideology are almost 40% higher for someone who liked the website of the left-leaning newspaper *De Morgen* than for who did not, and they are almost 20% higher for someone who liked the nonprofit organization *Amnesty International*. Right pages are dominated by Flemish nationalistic content and memes. The odds of demonstrating a right political leaning are almost 20% higher for someone who liked the alternative right-wing news website *SCEPTR* or the popular mainstream paper *HLN.be* than for someone who did not. Similarly to the left pages, most of these pages carry a subtle, or sometimes an outspoken, reference to political ideology. In contrast, center voters like less explicit or implicit political pages on Facebook.

Similarly, to predict political parties, politically loaded pages are ranked high, such as newspapers or organizations with a certain ideology. Bluntly summarized, voters for the workers'—and social democratic party like solidarity content such as refugee or third-world organizations, green voters mainly like environmental pages (e.g. Greenpeace) Christian democratic voters are interested in religious organizations and the royal family, liberal voters in financial news and businesses, and finally the Flemish nationalists and extreme rightists tend to like Flemish nationalistic and identity content on Facebook. At first sight, most of these likes seem to be in line with the broader ideological or issue profile of the party family.

When delving deeper into aspects of lifestyle and cultural preferences our methodology allows us to focus on the most related pages per specific Facebook category. For example, when considering only the pages in the category *Movies*, left voters more often like adventure, romance and drama movies, such as *The Hunger Games* (Odds Ratio (OR) = 1.06) or *500 Days of Summer* (OR = 1.06), and visit arthouse cinemas in larger cities. In contrast, action movies such as *Scarface* (OR = 1.04) and *Fast & Furious* (OR = 1.04) are mainly liked by voters on the right. In the category *Music*, the genres of alternative rock, blues, and experimental music are most related to a left political leaning (e.g. *Bob Dylan* (OR = 1.08) or *Tom Waits* (OR = 1.07)) whereas popular (hard) rock music (e.g. *AC/DC* (OR = 1.05)) is often liked by right voters, next to the genres of techno and electronic dance music (e.g. *Justice* (OR =1.04)). The same analysis can be done for other categories such as books, food, brands, sports, etc. Clearly, the lower odds ratios of a single movie or rock band compared to those of a civil-society organization or alternative news outlet indicate that their connection with ideological preference is present, but much more modest.

6 Conclusion and future research

The starting point of this study is that lifestyle and politics are closely related. Next to classical variables such as socio-demographics and issue preferences, our personal values and lifestyle choices correlate with our political preferences. The fact that (non-political) Facebook likes achieve high predictive accuracy (in addition to survey data) shows that they are indeed capturing additional information, which we argue are related to values and lifestyle.

Consequently, we looked into which aspects of our social life are most predictive for our political preference by analyzing the predictive performance of Facebook categories (e.g. movies, music, food, etc.) separately. Non-surprisingly, politicians, media and civil society are most predictive for political leaning, but also arts, culture, entertainment and books help to predict where people stand politically, whereas, in particular sports and travel are less predictive. This raises the question why some aspects of our social lives are more connected to political preference than others. For instance, our study seems to suggest that Belgian citizens are more likely to meet people with different political convictions when cheering for their favorite sports team than when going to a music festival. However, the extent to which these results can be generalized to other contexts or countries is uncertain. For instance, the study of [14] found that for Twitter users in the US musical preference was less connected to political preference than following sports teams. To reach insights that travel across countries and time periods, future research needs to be comparative or at least more similar in terms of data and methods.

In contrast with most previous research on lifestyle politics in the (polarized) two-party system in US, we examined lifestyle politics in a multi-party system, with much more subtle ideological differences between parties. Our study indicates that Facebook likes are less predictive for center voters and for traditional political parties. A less clear pattern is present in the characteristics and behavior of those voters compared to voters with a more outspoken ideological position. We find, for instance, that our social media data are much better in predicting who votes for an extreme-right party compared to the social democrats. Is this simply due to the more radical or straightforward ideological position of these parties, or do certain politicians link their ideology or party platform more to lifestyle choices?

The same goes for placement on the left-right scale. Our model is better at predicting positions to the left or right on this scale than predicting the center position. Again, this might simply be because center voters have less distinguishable profiles, but it may also be driven by the fact that respondents who in surveys position themselves at the center tend to be less politicized. Studies have demonstrated that at least part of these respondents have no clear political attitudes and tend to have lower political interest and less political knowledge [59–61]. In this sense, these scales cannot fully disentangle respondents who are politically neutral from those who are non-political. Nevertheless, the fact that it is also more difficult to predict actual votes for center parties—although to a lesser extent than the center ideological position—suggests that at least part of the difference can likely be attributed to the fact that the ideological position at the center is less connected to distinct lifestyle profiles. How to explain these differences and why certain ideological and political opinions are more connected to lifestyle than others, are interesting follow-up questions for social scientists, and electoral scholars in multi-party systems in particular. At this stage it remains unclear whether an ideological position leads to a specific lifestyle, or rather whether certain lifestyle choices facilitate the adoption of certain political attitudes. Future research should try to dig deeper into the causal mechanism of this intriguing finding.

Our analyses showed that with Facebook data, different interest categories can easily be analyzed and compared to improve our understanding of public opinion and voter behavior. As such, digital trace data can complement survey data. Although we believe that it is theoretically possible to achieve the same predictive power with targeted survey questions, we think it is not practically feasible in the context of exploratory research for two reasons. First, capturing this amount of detailed information using traditional surveys is very difficult, as it would require very long question batteries on different categories related to lifestyle and consumer choices. The survey of DellaPosta et al. [10] is a rare example of such an extensive survey, but requires a great effort from respondents. In addition, it would request scholars to define a priori the most important political lifestyle indicators. Therefore, we believe digital trace data are invaluable for exploratory research, as they can capture many different lifestyle options with relatively little effort. The patterns extracted from behavioral data can then help scientist to refine research questions and formulate targeted survey questions.

An important aspect of future research is then how to summarize the insights gained from Facebook likes in such a way that they can support theory building. Our classification using Facebook categories is one way. Other possibilities include dimensionality reduction or clustering Facebook likes to learn a certain structure from the data. For instance, movies or music of the same genre could be grouped together, or food of the same cuisine. These dimensions or clusters can help us to identify overarching cultural taste or lifestyle patterns which can thereafter be transformed into a series of survey questions and used for explanatory analysis.

At the same time, we do not argue that social media data can replace survey research. Probably both advanced survey research and social media data are needed to understand how lifestyle and political preferences influence each other. For instance, DellaPosta and colleagues (2015) explain the puzzling association between lifestyle preferences and political affiliation by the self-reinforcing effect of homophily and social influence. In short, they argue that people make lifestyle choices and opt for places where they meet likeminded people (homophily), in those (online) places people's attitudes are affected by exposure to each other (social influence). However, to better understand how this self-reinforcing effect takes place, further research is needed. By using a survey panel design [62] or by simultaneously analyzing friend networks on Facebook [63], the relationships between homophily and social influence could be further explored. Interviews or surveys could seek for explanations for this association and its direction by asking questions such as "Do you like to go to events where people have the same opinions as you?" or "Do you talk about politics or societal issues at music concerts?".

We argue that combing social media and survey data is not only useful to understand where "the cultural fault lines" are most persistent, but it can simultaneously provide insights in which (type of) organizations, brands or events "unite" people with different political views. For example, our data indicate that people from all sides of the ideological spectrum "like" a political comedy show from the public broadcaster (*De Ideale Wereld*). From the perspective of the recently growing polarization of the audience for late night comedy in the US [64], this is not a trivial finding. Therefore, this type of knowledge may be increasingly relevant and useful, as ever more countries are facing increasing levels of (affective) polarization [20, 21]. Furthermore, the growing affective polarization in society might make people more eager to express their political identity online by liking pages or following actors that are seen in line with their political views and cultural tastes. This would imply that similar factors drive both online and offline lifestyle choices. Further research could tackle this, among others by studying the linkage between social media data and political preferences over time.

In conclusion, we argue that fine-grained behavioral data are useful to discover unknown patterns or better understand existing relationships and are therefore equally valuable for political and social science research. However, the mining of behavioral and online data has raised new and unexpected ethical and regulatory questions related to online collection, storage, and use of human subjects' data [65]. We therefore underline the importance of an ethical discussion of the research design in journal manuscripts, which we included in Appendix A.1.5. Unfortunately, Facebook restricted data collection through the APIs of Facebook, Instagram, and other platforms it owns. While this intervention certainly is positive for the privacy protection of their users, it is also locking out third parties and diminishing transparency of the platform. Social scientists depend on the data collected through the API to investigate the impact of such platforms on our media and society. A potential promising avenue of Facebook is to work in a coordinated way with researchers.[16] However, some scholars question whether this initiative will provide sufficient support for free and independent scientific research [66]. Nonetheless, the potential of other social media data (e.g. Twitter) or other types of behavioral data (e.g. location data, payment data or browsing data) needs to be further explored as they complement the insights of survey research into different aspects of citizens' social and political life.

Appendix: additional files

A.1 Data collection

The data collection took place in two waves in March and June 2018. In the first wave (March 2018), a detailed survey with questions on socio-demographics, media consump-

[16] see https://socialscience.one/.

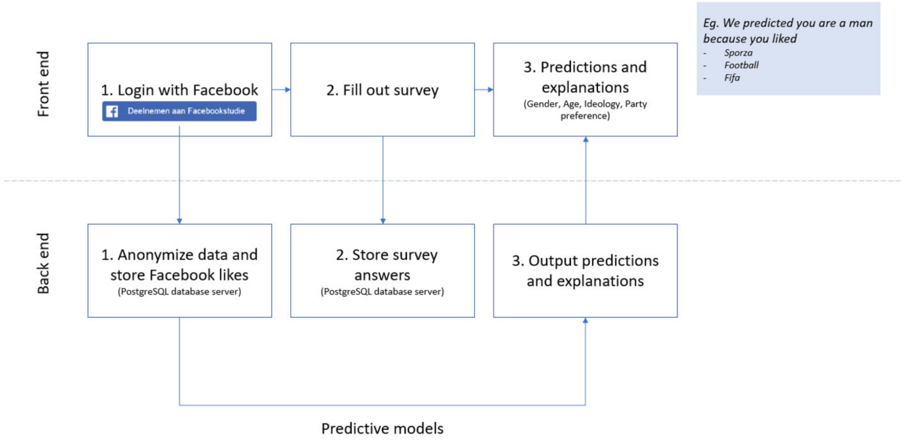

Front end
1. Login with Facebook
Deelnemen aan Facebookstudie
2. Fill out survey
3. Predictions and explanations
(Gender, Age, Ideology, Party preference)
Eg. We predicted you are a man because you liked
- Sporza
- Football
- Fifa
Back end
1. Anonymize data and store Facebook likes
(PostgreSQL database server)
2. Store survey answers
(PostgreSQL database server)
3. Output predictions and explanations
Predictive models

Figure 4 Overview of the front- and back end of the webpage that was created to collect Facebook like data for our study

tion, political preference and attitudes was sent to a representative panel of around 4500 respondents. From these respondents, 524 agreed to give us access to their Facebook like data, via Facebook Login (for more details about Facebook Login see Appendix A.1.4). Based on the data of this small set of respondents, models were built to predict gender, age, political leaning and party preference based on Facebook likes. These initial prediction models were used to develop a tool that shows participants which characteristics can be inferred about them based on their Facebook likes. The goal of this tool was two-fold: (1) to convince people to participate in our study and (2) to create awareness about the personal information that you might disclose about yourself on Facebook. In the second wave (May–June 2018), this tool was disseminated through the online webpages of popular Flemish newspapers to reach a broad audience. Via an online webpage, users could give consent to collect their Facebook likes and they were also asked to complete 12 survey questions about their media consumption and political preference. In return for completing the survey, people could see a prediction of their gender, age, ideology and party preference based on the Facebook likes they provided. The Facebook likes and survey questions were anonymously and securely stored on a PostgreSQL database server. An overview of the front- and back end of our webpage can be found in Fig. 4.

A.1.1 Survey invitation

In wave 1, this survey invitation was sent to a representative panel (translated from Dutch to English):

"People increasingly keep themselves up to date on the news via social media such as Twitter and Facebook. Therefore, as researchers, we are very interested in the public pages you follow (like) on Facebook and the posts you share or post on your timeline. In addition, we want to examine what companies (such as Cambridge Analytica) can deduct from the like behavior of Facebook users and provide users insight into this.

For this we need your help. By participating in our study, you allow us access to your Facebook data and will need to answer ten questions. We will use those questions to determine what we can predict of people on the basis of their page likes on Facebook.

Of course, all data will be processed and stored fully anonymously. This research pursues scientific aims only, and data will not be shared with others. Your privacy is of high importance to is. Read our privacy policy here.

If you have any further questions about this research you can read our frequently asked questions section. If you still have any questions left you can contact us via nws-data@uantwerpen.be."

In wave 2, this text was added to the previous invitation:[17]

"Based on your Facebook likes we use data mining models to predict your gender, age, ideological leaning and party preference. We will also show you which specific likes (i.e. the Facebook pages that you have liked yourself) are the most important contributors to this prediction. Do you want to find out what your Facebook likes reveal about you? You can test it here!"

A.1.2 Survey questions

1. *How often did you use the following channels in the past month to follow news about current events? Never, less than once a week, once or twice a week, 3 or 4 times a week, daily, several times a day* (a) Radio (b) TV (c) Online (d) Newspapers (e) Facebook (f) Twitter
2. *To what extend are you interested in the following news subjects on a scale of 1 (not interested at all) to 5 (very strongly interested)?* (a) International (b) Politics (c) Local (d) Finance and economics (e) Entertainment (f) Lifestyle (cooking, fashion, travel) (g) Arts and culture (h) Sports (i) Science and technology (j) Justice and safety (k) Bizarre/funny
3. *In politics we often use the terms "left" and "right". Where would you situate your own opinion on a scale from 0 to 10, where 0 means "left", 5 "center", and 10 "right"?*
4. *Please indicate on a scale from 1 to 10 how likely it is that you will ever vote for each of the parties listed below. 1 means that you will never vote for this party and 10 means that you will definitely vote for this party.* (a) CD&V (b) Groen (c) N-VA (d) Open VLD (e) PVDA (f) Sp.a (g) Vlaams Belang
5. *To what extend are you interested in politics in general on a scale of 0 (not interested at all) to 10 (very much interested)?*
6. *What is your gender?* Male, female
7. *What is your age?* Younger than 25, between 25 and 55, older than 55
8. *What is your highest level of education?* No education, lower education, general education, technical education, vocational education, higher education (non-university), University education

A.1.3 Survey weights

We used Iterative proportional fitting (IPF)[18] to adjust survey weights to reflect the overall population distribution in terms of gender, age and education levels (see Table 5). However, the weighting process did not influence the outcomes all that much (e.g. the spearman rank correlation between the coefficients of the models with and without survey weights

[17] The original webpage can be found here (only in Dutch): https://www.uantwerpen.be/nl/projecten/nws-data/facebookstudie/dataverzameling/.

[18] We used the Python package ipfn https://pypi.org/project/ipfn/.

Table 4 The number of participants in our dataset per target variable

		All participants (6.733)
Leaning	Left	2196 (33%)
	Center	2322 (34%)
	Right	1958 (29%)
	No answer	257 (4%)
Party preference	PVDA	750 (11%)
	Sp.a	781 (12%)
	Groen	1988 (30%)
	CD&V	620 (9%)
	Open VLD	1150 (17%)
	N-VA	1270 (19%)
	VLaams Belang	150 (2%)
	No answer	24 (0%)

Table 5 Gender, age and education of the Belgian population [67]

	Variable	Survey total	Population percentage	Weighted survey total
Gender	male	3747	50%	2868
	female	2078	50%	2957
Age	younger than 25	1598	28%	1645
	between 25 and 55	3558	40%	2311
	older than 55	669	32%	1868
Education	no education	75	3%	147
	lower education	41	5%	307
	vocational education	190	9%	550
	technical education	578	12%	679
	general education	937	16%	932
	higher education	1661	18%	1022
	university education	2343	38%	2188

applied was 0.94) and therefore the unweighted results will be reported in the manuscript.

A.1.4 Facebook login

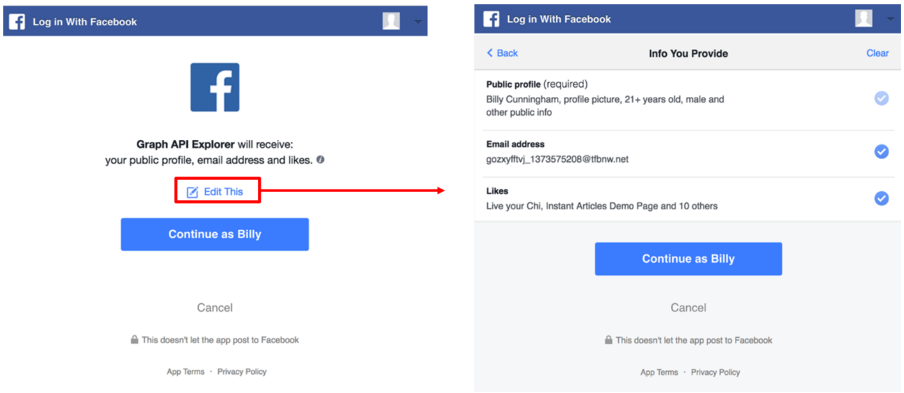
Facebook Login[19] is a developers' tool that allows users to authenticate using their Facebook credentials and log into an app. Facebook Login also enables to ask for permissions to access items of user data. These permissions, if granted by the user, provide access to a subset of that person's data stored on Facebook. To be able to ask users for such permissions, passing a review process (called App Review[20]) is necessary, so that Facebook can make sure that the data are not misused.[21] Researchers that wish to access a user's personal Facebook data thus need to obtain permission from Facebook as well as from the user himself.

After we passed the review process, we integrated the Facebook Login button on our webpage. Respondents could click on this button, authenticate with their Facebook credentials, and grant us permission to their Facebook likes (see a mockup example in Fig. 5).

[19] https://developers.facebook.com/docs/facebook-login/.

[20] https://developers.facebook.com/docs/facebook-login/review.

[21] It should be noted that after the Cambridge Analytica scandal, the Facebook Platform has further restricted data access in March 2018, so that it is no longer allowed to access Facebook profile data for academic research—unless in collaboration with Facebook, e.g. through the Social Science One project (https://socialscience.one/).



Graph API Explorer will receive:
your public profile, email address and likes.

This doesn't let the app post to Facebook

Public profile (required)
Billy Cunningham, profile picture, 21+ years old, male and other public info

Email address
gozxyfftvj_1373575208@tfbnw.net

Likes
Live your Chi, Instant Articles Demo Page and 10 others

This doesn't let the app post to Facebook

Figure 5 Mockup example of the Facebook Login button and the edit permissions window

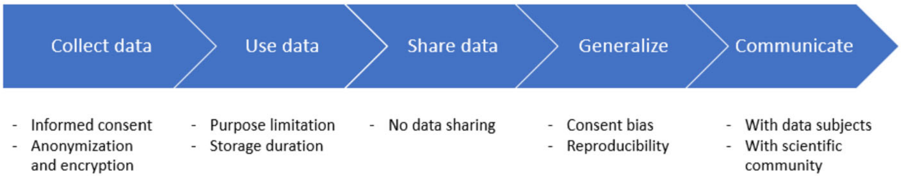


Figure 6 Framework to highlight the ethical concerns of our research, based on [65]

A.1.5 Privacy and ethical concerns

Facebook and other online environments provided tremendous opportunities for social science research to study human behavior. At the same time, the rapid increase in available online data and technological progress raised new and unexpected ethical questions related to online collection, storage, and use of human subjects' data [68]. These questions become urgent as the data and research possibilities move well beyond those typical in the social sciences, to more directly address sensitive aspects of human behavior and our daily lives. Yet, the variety in data sources, research topics, and methodological approaches complicates the draft of universally applicable ethical guidelines. Rather, all researchers engaging with (data from) human subjects, have the ethical responsibility to minimize potential harm. Even seemingly benign data can contain sensitive or private information and has the potential to impact people's lives [68]. From a legal standpoint, the introduction of the European Union's General Data Protection Regulation (GDPR) in May 2018, offers a regulatory framework for researchers collecting (online) personal data. Next to that, various instances are updating ethical guidelines for Internet research and human subjects [69, 70].

We base an ethical discussion of our research and data collection on the work of [65] that unites GDPR regulation and moral concerns into one concise data science framework. We present a version of this framework, tailored to our study in Fig. 6 and will briefly discuss its elements below.

The data needed for this study consist of survey questions and Facebook like data. All data were collected with informed consent. We clearly stated what data would be gathered and for what purpose. Respondents were required to agree with our privacy policy to participate in our study, and have the right to stop their collaboration at any time and ask for their data to be removed. Furthermore, no other data than the data needed for our study were collected. The data are anonymized immediately, meaning that all personal identifiers (i.e. name and Facebook id) are permanently removed. Yet, we are aware of the risk of reidentification [71], especially since like behavior is quite unique per individual. It is increasingly hard to transform the data in such a way that it protects sensitive information [72]. Therefore the -anonymized- data will under no circumstances be made public or shared with other instances and are stored locally and encrypted for optimal security. The data are stored and used for scientific purposes only and will be removed permanently after ten years. Results will be shown on an aggregate level only and will never contain individual examples, nor will we explicitly mention pages that were liked by less than 50 participants. Albeit absolutely necessary to protect privacy of our respondents, our privacy policy demonstrates two important implications for scientific generalizability. First, informed consent automatically induces self-selection or consent bias, and generalizing our results to the total population is difficult. Secondly, since we can only share aggregated results, this complicates the reproducibility of our study. However, the complete computer code and aggregated data can be made publicly available to assess scientific quality. Finally, a last important aspect of ethical research conduct is clear communication with data subjects (participants) before and after the study. Before the start of our data collection we compiled a clear privacy statement, in collaboration with the university's data protection officer and with approval by the Ethics Committee for the Social Sciences and Humanities (EA SHW).[22] The full privacy statement was communicated with participants and can be found on our website,[23] where we have also communicated a concise summary of our results. To conclude, we believe that clear communication about ethical research design in scientific journals benefits the ethical debate and framework for online human subject research. We therefore encourage all researchers and journals to include such discussion.

A.2 Statistical comparison of models

We use a corrected paired differences t-test (based on 10-fold cross-validation results) to compare the AUC of the different models [55] and apply the Bonferonni correction for multiple pairwise comparison [57, 58]. The results can be found in Table 6 (ideological leaning) and Table 7 (party choice). For leaning, the number of models that is being compared is 3 (the number of ideologies) times 10 (the number of folds per binary model), for party choice this is 7 (number of parties) times 10 (number of folds per binary model). Also the predictive performance of non-political Facebook likes for the ideologies and parties are compared, here the number of models is 10 (number of folds). Secondly, in Table 8 and 9 we show the results for a Wilcoxon signed-rank test, a non-parametric alternative to the paired t-test [56]. The results for this non-parametric test largely correspond to the parametric t-test.

[22] https://www.uantwerpen.be/en/research/management/quality-assurance/ethics-screening/eashw/.

[23] https://www.uantwerpen.be/nl/projecten/nws-data/privacybeleid/. (Dutch only)

Table 6 Results for the corrected paired t-tests for ideological leaning

		T-score	P-value
M5	M1	–0.816	1.586
M5	M2	4.745	0.000*
M5	M3	9.358	0.000*
M4	M5	2.251	0.024
M4	M1	5.54	0.000*
M4	M2	13.64	0.000*
M4	M3	31.64	0.000*
M1	M2	9.02	0.000*
M1	M3	21.42	0.000*
M2	M3	8.27	0.000*
Left	Right	0.446	0.656
Left	Center	16.990	0.000**
Right	Center	18.392	0.000**

* Significant with $\alpha/10 = 0.005$.
** Significant with $\alpha/3 = 0.017$.

Table 7 Results for the corrected paired t-tests for party preference

		T-score	P-value
M5	M1	–10.134	2.000
M5	M2	4.375	0.000*
M5	M3	29.620	0.000*
M4	M5	30.168	0.000*
M4	M1	–0.103	1.082
M4	M2	16.338	0.000*
M4	M3	40.747	0.000*
M1	M2	17.018	0.000*
M1	M3	22.860	0.000*
M2	M3	13.889	0.000*
N-VA	Vlaams Belang	1.754	0.079
N-VA	Open VLD	7.101	0.000**
N-VA	PVDA	8.488	0.000**
N-VA	Groen	11.771	0.000**
N-VA	Sp.a	8.401	0.000**
N-VA	CD&V	13.311	0.000**
Vlaams Belang	Open VLD	1.193	0.233
Vlaams Belang	PVDA	1.104	0.270
Vlaams Belang	Groen	1.890	0.059
Vlaams Belang	Sp.a	3.099	0.002
Vlaams Belang	CD&V	2.946	0.003
Open VLD	PVDA	0.051	0.960
Open VLD	Groen	1.953	0.051
Open VLD	Sp.a	2.639	0.008
Open VLD	CD&V	3.816	0.000**
PVDA	Groen	1.784	0.075
PVDA	Sp.a	2.481	0.013
PVDA	CD&V	3.156	0.002
Groen	Sp.a	1.969	0.049
Groen	CD&V	2.861	0.004
Sp.a	CD&V	0.778	0.437

* Significant with $\alpha/10 = 0.005$.
** Significant with $\alpha/21 = 0.002$.

Table 8 Results for the corrected paired t-tests for ideological leaning

		T-score	P-value
M4	M5	9.000	0.064
M4	M1	0.000	0.002*
M4	M2	0.000	0.002*
M4	M3	0.000	0.002*
M5	M1	23.000	0.695
M5	M2	0.000	0.002*
M5	M3	0.000	0.002*
M1	M2	0.000	0.002*
M1	M3	0.000	0.002*
M2	M3	0.000	0.002*
Left	Right	16.000	0.275
Left	Center	0.000	0.002**
Right	Center	0.000	0.002**

* Significant with $\alpha/10 = 0.005$.
** Significant with $\alpha/3 = 0.017$.

Table 9 Results for the Wilcoxon signed-rank test for party preference

		W-score	P-value
M5	M1	0.000	0.002*
M5	M2	3.000	0.010
M5	M3	0.000	0.002*
M4	M5	0.000	0.002*
M4	M1	26.000	0.922
M4	M2	0.000	0.002*
M4	M3	0.000	0.002*
M1	M2	0.000	0.002*
M1	M3	0.000	0.002*
M2	M3	0.000	0.002*
NVA	Vlaams Belang	15.000	0.232
NVA	Open VLD	0.000	0.002**
NVA	PVDA	0.000	0.002**
NVA	Groen	0.000	0.002**
NVA	Sp.a	0.000	0.002**
NVA	CD&V	0.000	0.002**
Vlaams Belang	Open VLD	16.000	0.275
Vlaams Belang	PVDA	19.000	0.432
Vlaams Belang	Groen	12.000	0.131
Vlaams Belang	Sp.a	5.000	0.020
Vlaams Belang	CD&V	6.000	0.027
Open VLD	PVDA	27.000	1.000
Open VLD	Groen	11.000	0.105
Open VLD	Sp.a	7.000	0.037
Open VLD	CD&V	5.000	0.020
PVDA	Groen	11.000	0.105
PVDA	Sp.a	8.000	0.049
PVDA	CD&V	6.000	0.027
Groen	Sp.a	9.000	0.064
Groen	CD&V	5.000	0.020
Sp.a	CD&V	20.000	0.492

* Significant with $\alpha/10 = 0.005$.
** Significant with $\alpha/21 = 0.002$.

A.3 Voter profiles

Table 10 The ten most related Facebook pages (with exclusion of political pages) to political leaning offer interesting insights in voters' interests

	Feature	Description	Mean coefficient	P-value
Left	De Morgen	News & media website	0.32	0.000
	De Wereld Morgen	News & media website	0.31	0.000
	HART BOVEN HARD	Citizens' initiative for more solidarity	0.26	0.000
	Apache	News & media website	0.20	0.000
	Amnesty International	Nonprofit organization—human rights	0.17	0.000
	Ringland	Citizens' initiative for green mobility	0.15	0.000
	Dagen Zonder Vlees	Citizens' initiative to consume less meat	0.13	0.000
	MO*	News & media website	0.12	0.000
	Vrolijk Relativerende Liga ter Bestrijding van Azijnpis & Verzuring	Playful page with subtle criticism toward society	0.12	0.000
	Ish Ait Hamou	Belgian dancer, choreographer, television presenter and author of Maroccan descent	0.11	0.000
Center	KU Leuven	University	0.09	0.003
	NoodweerBenelux	Weather forecasts	0.08	0.002
	Tasty	Food videos and recipes	0.08	0.011
	She.be	News & media website for women	0.08	0.005
	Ben & Jerry's	Ice cream	0.08	0.001
	Milow	Belgian musician	0.07	0.000
	The Economist	News & media website	0.07	0.004
	Avatar	Fantasy/Sciencefictionfilm	0.07	0.000
	Politie Leuven	Police station	0.07	0.000
	Radio 2	Radio station	0.07	0.002
Right	SCEPTR	News & media website	0.17	0.000
	HLN.be	News & media website	0.17	0.000
	De Tijd	News & media website	0.16	0.000
	Onafhankelijk Verbond Der Vlaemsche Meme	Flemish nationalistic memes	0.13	0.000
	De Fiere Vlaamse Meme	Flemish nationalistic memes	0.12	0.000
	Schild & Vrienden	Flemish nationalistic youth movement	0.12	0.000
	Vlaamse Volksbeweging	Nonprofit organization—pro-Flemish	0.12	0.000
	Duvel	Beer	0.10	0.001
	Club Brugge K.V.	Sports team	0.10	0.002
	Dan Bilzerian	Internet celebrity, poker player and actor known for his lavish lifestyle	0.10	0.000

Table 11 The significant questions to political leaning for the survey (M3)

	Feature	Mean coefficient	P-value
Left	Q2: news subjects—arts & culture	0.40	0.000
	Q6: gender—female	0.33	0.000
	Q2: news subjects—politics	0.26	0.000
	Q7: age—between 25 and 55	0.21	0.000
	Q2: news subjects—international	0.19	0.000
	Q8: education—university	0.08	0.045
	Q5: interest in politics	0.03	0.031
Center	Q2: news subjects—finance & economy	0.11	0.001
	Q2: news subjects—international	0.09	0.006
	Q8: education—technical	0.09	0.046
	Q1: news channels—TV	0.09	0.000
	Q8: education—vocational	0.08	0.045
	Q2: news subjects—arts & culture	0.07	0.009
	Q2: news subjects—arts & culture	0.07	0.009
	Q2: news subjects—arts & culture	0.07	0.016
	Q1: news channels—Twitter	0.06	0.001
	Q1: news channels—radio	0.04	0.012
Right	Q2: news subjects—finance & economy	0.33	0.000
	Q2: news subjects—justice & safety	0.29	0.000
	Q5: interest in politics	0.07	0.000
	Q2: news subjects—bizarre/funny	0.06	0.036
	Q1: news channels—TV	0.04	0.035

Acknowledgements

The authors would like to thank Enago (www.enago.com) for the English language review and Dynata (www.dynata.com) for the survey panel data collection.

Funding

This work was funded by University of Antwerp [FFB160362 BOF GOA].

Availability of data and materials

The datasets analysed during the current study are not publicly available due to privacy restrictions. A limited set of the data can be made available upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SP, PVA and DM conceived the research and designed the analyses. PVE and PVA designed the survey questions. SP, PVE, PVA and DM collected the data. SP performed the analyses and wrote the paper. SVDV and DM provided advice on the analysis. PVE and DM provided general advice and supervised the research. All authors read and approved the final manuscript.

Author details

[1]Department of Engineering Management, University of Antwerp, Prinsstraat 13, 2000, Antwerp, Belgium. [2]Department of Political Science, University of Antwerp, Sint Jacobsmarkt 2-4, 2000, Antwerp, Belgium.

Publisher's Note

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

Received: 16 March 2021 Accepted: 15 September 2021 Published online: 02 October 2021

References

1. Berelson BR, Lazarsfeld PF, McPhee WN, McPhee WN (1954) Voting: a study of opinion formation in a presidential campaign. University of Chicago Press, Chicago
2. Campbell A, Converse PE, Miller WE, Stokes DE (1960) The American voter. University of Chicago Press, Chicago
3. Lipset SM, Rokkan S (1967) Party systems and voter alignments: cross-national perspectives. Free Press, New York
4. Dalton RJ, McAllister I, Wattenberg MP (2000) The consequences of partisan dealignment. In: Parties without partisans: political change in advanced industrial democracies. Oxford University Press, Oxford, pp 37–63
5. Dalton RJ, Wattenberg MP (2002) Parties without partisans: political change in advanced industrial democracies. Oxford University Press, Oxford
6. Petrocik J (1989) The theory of issue ownership: issues, agendas, and electoral coalitions in the 1988 elections. In: Annual meeting of the American political science association

7. Funk CL (1999) Bringing the candidate into models of candidate evaluation. J Polit 61(3):700–720
8. Bennett WL (1998) The uncivic culture: communication, identity, and the rise of lifestyle politics. PS Polit Sci Polit 31(4):741–761
9. Giddens A (2013) The consequences of modernity. Wiley, New York
10. DellaPosta D, Shi Y, Macy M (2015) Why do liberals drink lattes? Am J Sociol 120(5):1473–1511
11. Purhonen S, Heikkilä R (2017) Food, music and politics: the interweaving of culinary taste patterns, highbrow musical taste and conservative attitudes in Finland. Soc Sci Inf 56(1):74–97
12. Khan R, Misra K, Singh V (2013) Ideology and brand consumption. Psychol Sci 24(3):326–333. https://doi.org/10.1177/0956797612457379
13. Kidwell B, Farmer A, Hardesty DM (2013) Getting liberals and conservatives to go green: political ideology and congruent appeals. J Consum Res 40(2):350–367
14. Shi Y, Mast K, Weber I, Kellum A, Macy M (2017) Cultural fault lines and political polarization. In: Proceedings of the 2017 ACM on web science conference, pp 213–217
15. Roos JM, Shachar R (2014) When kerry met sally: politics and perceptions in the demand for movies. Manag Sci 60(7):1617–1631
16. Johnston CD, Lavine HG, Federico CM (2017) Open versus closed: personality, identity, and the politics of redistribution. Cambridge University Press, Cambridge
17. Carpini MXD, Williams BA (2001) Let us infotain you: politics in the new media environment. In: Mediated politics: communication in the future of democracy, pp 160–181
18. Jones MO (2020) Politics on a plate: uses and abuses of foodways on the campaign trail. J Folk Res 57(2):41–79
19. Bennett WL (2017) Branded political communication: lifestyle politics, logo campaigns, and the rise of global citizenship. In: Politics, products and markets. Routledge, London, pp 101–125
20. Iyengar S, Sood G, Lelkes Y (2012) Affect, not ideologya social identity perspective on polarization. Public Opin Q 76(3):405–431
21. Lelkes Y (2016) Mass polarization: manifestations and measurements. Public Opin Q 80(S1):392–410
22. Martens D, Provost F, Clark J, de Fortuny EJ (2016) Mining massive fine-grained behavior data to improve predictive analytics. MIS Q 40(4):869–888
23. Stier S, Breuer J, Siegers P, Thorson K (2019) Integrating survey data and digital trace data: key issues in developing an emerging field. Soc Sci Comput Rev 38(5):503–516. https://doi.org/10.1177/0894439319843669
24. Bond R, Messing S (2015) Quantifying social media's political space: estimating ideology from publicly revealed preferences on Facebook. Am Polit Sci Rev 109(1):62–78
25. Eady G, Nagler J, Guess A, Zilinsky J, Tucker JA (2019) How many people live in political bubbles on social media? Evidence from linked survey and Twitter data. Sage Open 9(1):1–22
26. Barberá P (2015) Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. Polit Anal 23(1):76–91
27. Wells C, Thorson K (2017) Combining big data and survey techniques to model effects of political content flows in Facebook. Soc Sci Comput Rev 35(1):33–52. https://doi.org/10.1177/0894439315609528
28. Joseph K, Wihbey J (2019) Breaking news and younger Twitter users: comparing self-reported motivations to online behavior. In: Proceedings of the 10th international conference on social media and society, pp 83–91
29. Mosleh M, Pennycook G, Rand DG (2020) Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. PLoS ONE 15(2):1–9
30. Mochon D, Johnson K, Schwartz J, Ariely D (2017) What are likes worth? A Facebook page field experiment. J Mark Res 54(2):306–317
31. Goffman E (1967) Interaction ritual: essays on face-to-face interaction
32. Eranti V, Lonkila M (2015) The social significance of the Facebook like button. First Monday 20:6
33. Hollenbeck CR, Kaikati AM (2012) Consumers' use of brands to reflect their actual and ideal selves on Facebook. Int J Res Mark 29(4):395–405
34. Matz SC, Netzer O (2017) Using big data as a window into consumers' psychology. Curr Opin Behav Sci 18:7–12
35. Piazza A, Zagel C, Haeske J, Bodendorf F (2017) Do you like according to your lifestyle? A quantitative analysis of the relation between individual Facebook likes and the users lifestyle. In: International conference on applied human factors and ergonomics. Springer, Berlin, pp 128–139
36. Guess A, Munger K, Nagler J, Tucker J (2019) How accurate are survey responses on social media and politics? Polit Commun 36(2):241–258
37. Furnham A (1986) Response bias, social desirability and dissimulation. Pers Individ Differ 7(3):385–400. https://doi.org/10.1016/0191-8869(86)90014-0
38. Nagler J, Tucker JA (2015) Drawing inferences and testing theories with big data. PS Polit Sci Polit 48(1):84–88
39. Dalton RJ (2016) The potential of big data for the cross-national study of political behavior. Int J Sociol 46(1):8–20
40. Kosinski M, Matz SC, Gosling SD, Popov V, Stillwell D (2015) Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. Am Psychol 70(6):543–556
41. Buntain C, McGrath E, Golbeck J, LaFree G (2016) Comparing social media and traditional surveys around the Boston marathon bombing. In: Proceedings of the 6th workshop on 'making sense of microposts', WWW2016, pp 34–41
42. Zimmer M (2010) But the data is already public: on the ethics of research in Facebook. Ethics Inf Technol 12(4):313–325
43. Druckman JN, Green DP, Kuklinski JH, Lupia A (2006) The growth and development of experimental research in political science. Am Polit Sci Rev 100(4):627–635. https://doi.org/10.1017/S0003055406062514
44. Cranmer SJ, Desmarais BA (2017) What can we learn from predictive modeling? Polit Anal 25(2):145–166
45. Shmueli G et al (2010) To explain or to predict?. Stat Sci 25(3):289–310
46. Provost F, Fawcett T (2013) Data science for business: what you need to know about data mining and data-analytic thinking. O'Reilly Media, UK
47. Delwit P, Deschouwer K, Hooghe M, Baudewijns P, Walgrave S (2015) De Kiezer Ontcijferd. Over Stemgedrag en Stemmotivaties. LannooCampus
48. Hooghe M, Dassonneville R (2018) A spiral of distrust: a panel study on the relation between political distrust and protest voting in Belgium. Gov Oppos 53(1):104–130

49. 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
50. Francis JG, Payne C (1977) The use of the logistic model in political science: British elections, 1964–1970. Polit Methodol 4(3):233–270
51. Lockhart R, Taylor J, Tibshirani RJ, Tibshirani R (2014) A significance test for the lasso. Ann Stat 42(2):413
52. Tibshirani R, Wainwright M, Hastie T (2015) Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC, London
53. Jones E, Oliphant T, Peterson P et al (2001) SciPy: open source scientific tools for Python. http://www.scipy.org/
54. Scott DW (2015) Multivariate density estimation: theory, practice, and visualization. Wiley, New York
55. Nadeau C, Bengio Y (2000) Inference for the generalization error. In: Advances in neural information processing systems, pp 307–313
56. Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
57. Vázquez EG, Escolano AY, Riaño PG, Junquera JP (2001) Repeated measures multiple comparison procedures applied to model selection in neural networks. In: International work-conference on artificial neural networks. Springer, Berlin, pp 88–95
58. Pizarro J, Guerrero E, Galindo PL (2002) Multiple comparison procedures applied to model selection. Neurocomputing 48(1–4):155–173
59. Inglehart R, Klingemann H-D (1976) Party identification, ideological preference and the left-right dimension among western mass publics. In: Party identification and beyond, pp 243–273
60. Van der Eijk C, Niemöller B (1984) Theoretical and methodological considerations in the use of left-right scales. In: ECPR joint sessions of workshops, Salzburg, Austria
61. Kroh M (2007) Measuring left–right political orientation: the choice of response format. Public Opin Q 71(2):204–220
62. Dvir-Gvirsman S (2017) Media audience homophily: partisan websites, audience identity and polarization processes. New Media Soc 19(7):1072–1091
63. Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489(7415):295–298
64. Young DG (2019) Irony and outrage: the polarized landscape of rage, fear, and laughter in the United States. Oxford University Press, Oxford
65. Greene T, Shmueli G, Ray S, Fell J (2019) Adjusting to the gdpr: the impact on data scientists and behavioral researchers. Big Data 7(3):140–162
66. Bruns A, Bechmann A, Burgess J, Chadwick A, Clark LS, Dutton WH, Ess CM, Gruzd A, Halford S, Hofmann J et al (2018) Facebook shuts the gate after the horse has bolted, and hurts real research in the process. Int Pol Rev
67. StatBel: Onderwijsniveau van de Belgische bevolking van 15-64 jaar (2018) https://bestat.statbel.fgov.be/bestat/crosstable.xhtml?view=631b4535-7a63-4695-967f-fe42238ee9af
68. Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, Goodman A, Hollander R, Koenig BA, Metcalf J et al (2017) Ten simple rules for responsible big data research. PLoS Comput Biol 13(3):1–10
69. Franzke A, Bechmann A, Zimmer M, Ess C (2020) The association of internet researchers: Internet research: ethical guidelines 3.0
70. Office for Human Research Protections U.S. Department of Health and Human Services (2020) International compilation of human research standards. Office for Human Research Protections https://www.hhs.gov/ohrp/sites/default/files/2020-international-compilation-of-human-research-standards.pdf
71. Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: 2008 IEEE symposium on security and privacy. IEEE Press, New York, pp 111–125
72. Monreale A, Rinzivillo S, Pratesi F, Giannotti F, Pedreschi D (2014) Privacy-by-design in big data analytics and social mining. EPJ Data Sci 3:1