Journal of Information Technology & Politics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/witp20

Predicting political attitudes from web tracking data: a machine learning approach

Nora Kirkizh, Roberto Ulloa, Sebastian Stier & Jürgen Pfeffer

To cite this article: Nora Kirkizh, Roberto Ulloa, Sebastian Stier & Jürgen Pfeffer (2024) Predicting political attitudes from web tracking data: a machine learning approach, Journal of Information Technology & Politics, 21:4, 564-577, DOI: 10.1080/19331681.2024.2316679

To link to this article: https://doi.org/10.1080/19331681.2024.2316679

JOURNAL OF INFORMATION TECHNOLOGY & POLITICS
2024, VOL. 21, NO. 4, 564–577
https://doi.org/10.1080/19331681.2024.2316679

Routledge
Taylor & Francis Group

Predicting political attitudes from web tracking data: a machine learning approach

Nora Kirkizh, Roberto Ulloa, Sebastian Stier, and Jürgen Pfeffer

ABSTRACT

Anecdotal evidence suggests that the surge of populism and subsequent political polarization might make voters' political preferences more detectable from digital trace data. This potential scenario could expose voters to the risk of being targeted and easily influenced by political actors. This study investigates the linkage between over 19,000,000 website visits, tracked from 1,003 users in Germany, and their survey responses to explore whether website choices can accurately predict political attitudes across five dimensions: Immigration, democracy, issues (such as climate and the European Union), populism, and trust. Our findings indicate a limited ability to identify political attitudes from individuals' website visits. Our most effective machine learning algorithm predicted interest in politics and attitudes toward democracy but with dependency on model parameters. Although website categories exhibited suggestive patterns, they only marginally distinguished between individuals with anti- or pro-immigration attitudes, as well as those with populist or mainstream attitudes. This further confirm the reliability of surveys in measuring attitudes compared to digital trace data and, from a normative perspective, suggests that the potential to extract sensitive political information from online behavioral data, which could be utilized for microtargeting, remains limited.

KEYWORDS

Political attitudes; web tracking data; machine learning; surveys; life-style, immigration, climate change, democracy, European union

Introduction

Increasing political polarization makes voters' policy preferences easier to identify from self-reported vote choice and political ideology. However, anecdotal evidence shows that effects of political polarization may expand beyond politics. For example, an online quiz published by New York Times, a newspaper in the United States, demonstrated that some Donald Trump voters could be identified from their food diets.[1] Republican party in the United States targets with political ads Facebook users who are hunting, fishing, or playing golf.[2] As a result of the potential for vote choices to be identified from digital trace data the industry became more cautious. For example, Google, Facebook, X (former Twitter) made significant changes to their political advertising policies to prevent the display of ads containing potentially false information prior to the US presidential election in 2020.[3,4,5]

However, despite major social media platforms and search engines adapting preemptive privacy policies, research offers mixed evidence of political features being identifiable from digital trace data. ML models trained on Facebook likes, including lifestyle-related ones, can predict if a person is Democrat or Republican (Kosinski et al. 2013), and even vote choices themselves (Cerina & Duch, 2020). Visits to untrustworthy news websites are related to people's populist attitudes (Stier et al. 2020) and right-wing political ideology (Guess, Nyhan, & Reifler, 2020). Praet, Guess, Tucker, Bonneau, and Nagler (2021), however, show that lifestyle Facebook likes have limited prediction power when used to identify political ideology. The source of this mixed evidence may be traced back to the data-generating process: Since people may be reluctant to publicly show their true life-style choices, social media might not offer a complete picture. In this article, we use browsing histories, which directly identify peoples' everyday decisions, to explore the link between political orientations and lifestyle beyond the image of users displayed on social media. We also go beyond a political ideology argument, which is often applied to the US samples, by testing the predictive power of website choices to identify political

CONTACT Nora Kirkizh eleonora.kirkiza@tum.de The School of Social Sciences and Technology, Technical University of Munich, Munich, Germany
Supplemental data for this article can be accessed online at https://doi.org/10.1080/19331681.2024.2316679

attitudes in broader policy domains in a European country.

We test this argument based on several types of ML models which we supply with three-month web browsing histories from 1,003 individuals living in Germany and survey data measuring political attitudes toward (1) *immigration*, (2) *democracy*, (3) *climate change policies*, (4) *trust in public institutions*, and (5) *populist attitudes* — policy dimensions that reflect manifestos of major political parties in Germany, and parties' Facebook pages (Kirkizh et al. 2022). We also measured participants' interest in politics and attitudes toward the European Union (the EU). Overall, we examine if individual political attitudes are identifiable from *general* website choices, not just their news-related behavior or Facebook likes, which, as mentioned above, has been the focus of most previous research.

The contribution of this paper lies in methodological and policy making dimensions. First, from a methodological perspective, we offer an ML application to investigate political attitudes measured with surveys. ML algorithms used in this study allow to capture complex non-linear patterns in the data and obtain more robust predictions to advance theory on relationships between political attitudes and web browsing behavior (Leist et. al., 2022). Second, we show whether web tracking data can be used as a measurement of attitudes and compete with survey-based measures. Third, we offer the investigation of potential and the limits of web tracking data in predicting political attitudes based on the 2019 data setting up a pipeline for future research in different time frames. From normative perspective, our study sheds light on how much third parties can potentially learn about voters from their browsing histories, which, in turn, is connected to whether the urge for recent developments in digital privacy policies is justified. This, in turn, is connected to our initial argument about political polarization expanding beyond consumption of political content.

Theory and literature

Can website choices reveal relevant signals to identify political attitudes, and if yes, what is the underlying theory? We rely on two bodies of literature. One proposes theory and evidence that personality is linked to political attitudes; the other is that personality can define lifestyle preferences.[6] Establishing these two links and following the transitive property, we posit the link "*political attitudes – online behavior*". In Figure 1, we visualize our theoretical model. The right part of the model shows the link between personality traits and political attitudes, and the left part –– personality traits and online behavior. However, empirical evidence for this link is limited. Praet, Guess, Tucker, Bonneau, and Nagler (2021) used lifestyle Facebook likes to predict political ideology based on the US sample. Consistent with the existing literature, the authors found that Facebook pages related to politics are the strongest predictors of political ideology, while other topic domains, such as sports, food, and music among others did not show significant effects on ideology. Other studies show similar results. For example, political Facebook likes can predict individuals' vote choices (Cerina & Duch, 2020), whether a user is a democrat or republican (Kosinski et al. 2013), and visits to untrustworthy news websites are associated with populist attitudes (Stier et al. 2020) and political ideology (Guess et al. 2020). Overall, the predictive power of lifestyle website choices is still understudied and limited to social media data and the United States context. Since people may be reluctant to publicly show their true lifestyle choices, social media might offer an incomplete picture. In this paper, we use more advantageous data source than social media –– web tracking data that can show a closer to a complete picture of respondents' lifestyle behavior than what social media or surveys are able to demonstrate. A primary reason of this advantage of web tracking data is that it measures online behavior directly while social media and surveys are data sources significantly altered by users or respondents.

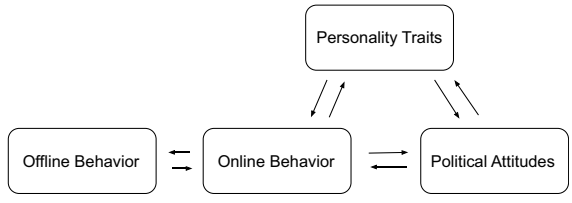

Figure 1. Theoretical model.

We present an identification of a broad set of political attitudes based on voters' website choices that are observable from browsing histories. We further leverage the premise that lifestyle choices or life circumstances, which we observe from web tracking data, are affecting political attitudes. A limited number of studies show that lifestyle can be tied to political views (DellaPosta et al. 2015), book shopping signal political ideology (Shi et al. 2017) and surveys conducted by Pew Research Center show that size and location of the house can predict political ideology.[7] For example, visits to accommodation services (e.g., booking.com), or flight booking websites (e.g., google.com/travel/flights) may signify frequency or interest in traveling and therefore signal potential support for open-borders policies and welcoming immigrants; gambling platforms (e.g., lotto.de) may be linked to financial issues and therefore potentially directed toward support for populist politics, which often exploits economic hardship (Wiedemann, 2023); job search websites (e.g., indeed.com) signal about employment status (Kerna et al. 2019) and therefore, if unemployed, could correlate with populist attitudes; political online media outlets signify interest in politics (Möller et al. 2020), and visits to pirate video streaming websites (e.g., uTorrent.com) could be linked to low trust in institution. And this list can continue: Shopping (amazon.com), sports, dating websites, well-being online services (meditations, yoga, etc.), and websites related to food diets, which may also be linked to political attitudes (Althoff et al. 2022). We leverage browsing behavior data from users to count their visits to this kind of lifestyle-related website and link them to their political attitudes.

Importantly, we do not test mechanisms that can be behind of the link between *political attitudes and online behavior*. In this paper, we are strictly interested in the predictive power of online behavior concerning political attitudes. One of the reasons for this theoretical strategy is that establishing mechanisms based on online behavioral data is challenging. For example, theoretically, hotel and flight booking platforms can be a proxy of cosmopolitan or, exactly opposite, nationalist orientations because it is important where exactly the respondent travels; visits to gambling websites can be because respondent has extra budget or, the opposite, lack of financial flexibility; visits to job search websites may be a sign of unemployed status or, the opposite, it could be a routine procedure for a professional to stay sharp in the profession; real estate websites might be visited by tenants as well as by owners. Consequently, our article rather focuses on methodological advantages of web tracking data for predicting political attitudes, which may facilitate further studies that are using web tracking data, including the study of mechanisms.

Data and measurement

In this paper, we use two types of data: web browsing logs and online survey responses. The data was collected with approval from the Oxford Internet Institute's Departmental Research Ethics Committee at the University of Oxford (Reference Number SSH IREC 18 004). We chose web tracking data over surveys to measure online behavior to avoid biases and the incomplete picture that survey panelists may have in their responses when asked to disclose or recall websites they visited during a particular week. Existing research shows that direct measure of online behavior with web tracking data is more accurate than self-reported measures and, to some extent, social media (Araujo et al. 2017; Englehardt et al. 2016; Scharkow, 2016; Stier et al. 2020).

Web tracking

We acquired web browsing histories of respondents from an online access panel maintained by Netquest, a market research company (please, see more details on recruiting in the Online Appendix.) Personally identifiable information is algorithmically anonymized by Netquest. We utilize web browsing histories from 1,003 study participants living in Germany. The tracking period is between mid-March and mid-June 2019. The dataset includes anonymized IDs, visited URLs, domains, and time spent on a web page. The dataset comprises 19,026,887 URLs (96,093 unique domains), with an average number of URL visits of 18,000 per respondent (Please, see more details on descriptive statistics of web tracking data in Table 1.) We specifically focus on cumulative

Table 1. Descriptive statistics of web tracking variables. There were 1,003 panelists 19,026,887 unique URLs, and 96,093 unique domains.

Statistic	N	Mean	St. Dev.	Min	Max
N visited URLs	1,003	18,080.07	23,864.05	53	191,526
N unique domains	1,003	362.04	328.97	9	2,279
μ visits per unique domain	1,003	43.36	37.30	3.28	376.76
μ duration per unique domain (sec.)	1,003	1,373.64	1,955.56	56.90	44,116.23
μ duration per URL (sec.)	1,003	33.34	23.58	1.62	276.12

number of visits to the websites, which we further group into topic domains (please, see the Models subsection), since we are striving for automated ML analysis. For more nuanced analysis of repeated visits to individual website domains, further in-depth research is required.

Further, we eliminated respondents who made less than 50 visits and visited less than nine unique domains. We also eliminated visits on which respondents spent less than three seconds, which allows us to avoid unintentional visits. Table 1 illustrates the distribution of means on a respondent level. Most of the respondents in our sample spend between 20 and 50 s on a unique web page (URL). Overall, the mean duration per unique domain and URL reported in Table 1 demonstrates regular browsing behavior, suitable to capture lifestyle preferences and daily life routines rather than incidental behavior.

We also tested to what extent our collected data represents the behavior of the general population. Since our panelists were aware of the tracking, they might have altered their behavior. In addition, we evaluate the extent to which tracking panelists' privacy attitudes diverge from panelists who participate in surveys but do not have tracking tools installed. Both validity tests are available in the Online Appendix.

Survey

We measured political attitudes with surveys, which we conducted in Germany parallel to the web tracking. We measured political attitudes based on survey questions from established annual survey panels such as *Eurobarometer*, *European Social Survey*, and *World Values Survey*. We also relied on systematic research of agendas of the major political parties and voters in Germany provided in Kirkizh, Froio, and Stier (2022). After the content analysis of party programs, political Facebook pages, and text analysis of open-ended questions related to the most critical issues in the country, Kirkizh, Froio, and Stier (2022) identified the four most prevalent policy domains: *immigration*, *democracy*, *climate change*, *the European Union* (the EU), and *populism*. Using these policy domains, we asked the respondents a set of attitudinal questions listed, along with the summary statistics, in Table 2. In addition to questions about attitudes toward democracy, we also measure trust in democratic institutions. Following a common political science practice, we also included a question measuring *political interest*. We placed responses to each survey question on Likert (from strongly disagree to strongly agree) or 1–11 scales (Please find the entire question wordings in the note of the Table 2). Distributions of a selected set of survey items are provided in the Online Appendix, Figure B1.

In addition to attitudinal questions, we asked demographic questions such as *age*, *gender*, *education* based on the German education system, and *income*. Overall, the sample composition consists of 1,003 respondents living in Germany, of which, 51% identified as female, and 49% as male. 24% of participants held at least elementary-level education, 54% had a mid-level education, and 22% reported a high education level (high school or above). Respondents were also distributed in the following age groups: 0.07% in 18–24, 21% in 25–54, 21% in 55–64, and 10% in 65+ age group. Median income of the respondents is 34,000 EUR. (See more details on sampling in the Online Appendix.) The following demographics distributions are deviating from nationally representative samples. Our respondents on average younger, more educated and have higher incomes than average population in Germany, which is common for online survey panels.

Table 2. Descriptive statistics of survey-based political attitudes.

Statistic	N	Mean	St. Dev.	Min	Max
Interest in politics	1,019	2.86	0.86	1.00	4.00
Trust in parliament (D)	1,019	3.24	1.17	1.00	5.00
Trust in the police (D)	1,020	2.54	1.10	1.00	5.00
EU integration (EU)	871	6.79	2.92	1.00	11.00
Income redistribution (P)	871	3.30	1.14	1.00	5.00
Big business and the people (P)	869	3.72	1.03	1.00	5.00
Social benefits and laziness (P)	870	2.79	1.15	1.00	5.00
Islam (I)	940	3.46	1.31	1.00	5.00
Immigrants and jobs (I)	1,020	2.83	0.91	1.00	4.00
Immigrants and crime (I)	1,020	2.04	0.89	1.00	4.00
Climate change and humans (C)	869	3.49	0.89	1.00	5.00
Free elections (D)	866	9.60	2.19	1.00	11.00
People obey their rulers (D)	866	3.96	2.92	1.00	11.00
Democratic political system (D)	868	3.39	0.69	1.00	4.00
Satisfaction with democracy (D)	1,019	2.63	0.80	1.00	4.00

Political attitudes question wordings and scales: interest in politics (1 - not at all, 4 very interested); trust in parliament and trust in the police (1 - not at all, 5 - a great deal); EU integration (1 - gone too far, 11 - should be pushed further); government should redistribute income from the better off to those who are less well off, big business takes advantage of ordinary people, social benefits make people lazy, Islam promotes violence more than other religions (1 - strongly disagree, 5 - strongly agree); immigrants take jobs away from German people, immigrants make crime problems worse (1 - strongly agree, 4 - strongly disagree); climate change is caused by natural processes, human activity, or both (1- natural processes, 5 - human activity); the following things are essential characteristics of democracy: free elections, and obeying the rulers (1 - not essential for democracy, 11 - essential for democracy); having a democratic political system (1 - very good way of governing this country, 4 - very bad way of governing this country); satisfaction with democracy (1 - not at all, 4 very satisfied).

Methods

We measure respondents' political attitudes with surveys and match them with their lifestyle choices, which we learn from web browsing histories. We combine these data types to find meaningful associations between political attitudes and daily life choices. We have over a thousand survey participants and their corresponding browsing histories, which generated millions of URLs over three months, which we further grouped into categories. Each website category can potentially be associated with a specific political attitude. Hence, each website category is an independent variable in a regression model, while political attitudes are dependent variables. The number of models equals to the number of attitudinal questions in Table 1.

However, in our data, the number of websites exceeds the number of respondents: Each regression model will have one dependent variable, thousands of independent variables, and only one thousand respondents. Because many websites in our data will have no visits since outside most popular websites like google.com or amazon.com, very few users visit the same web pages, it contributes to the increase of data sparsity, meaning that many cells in the data frame do not carry data points, which is in other words, missing data. There are several methods to deal with data sparsity (Dixit et al. 2020). In this paper, we use a multidimensionality reduction method (Engel, Hüttenberger, & Hamann, 2012), which helps to compress a data frame with thousands of websites. Following this approach, we offer a multidimensionality reduction method: Grouping websites by categories. Categories (specifically, the sum of visits for each category) are features that we used to train the algorithms.

Data pre-processing

We made two data pre-processing decisions based on our theory. In the analysis, we use website domains ([*domain*].com) to count visits and threshold for a visit duration. If we record 10 URLs with common domain *amazon* we count it as 10 visits to Amazon.com, ignoring URLs. Unlike URLs, the exact website domains appear more often across individuals' browsing histories in the dataset. For instance, users visit amazon.com several times a week, but URLs –– amazon.com/art-supplies/sale/TDFG54jjdiO320 –– they visit only once. The same web page can often have different URLs. Using website domains, we have more data points for each website of interest, e.g., Amazon, Netflix, LinkedIn, and others, than for a single URL. This approach is also a dimensionality

reduction method in addition to the main method we offer in this paper.

Additionally, to the processing of URLs, we use a specific time spent on a web page (TPS) threshold to capture deliberate visits. Since we aim for online behavior that signify individuals' lifestyle and routine behavior, domains larger TSP would more likely represent deliberate and meaningful visit of a web page. Extensive body of literature in the field of human-computer interaction established that TSP is one of the user interests in a web page (see a literature review in (Al Halabi, Kubat et al. 2007)) Empirical evidence offers several different thresholds for TSP to count a visit as a session and thus a deliberate web page visit. The suggested thresholds are between 48 s and 1.5 min (Hofgesang, 2006). We decide to use mean TSP based on this literature, which is 1-min threshold. After we removed "short" domain visits, where individuals spent less than one minute, the data generated 1,632,769 URLs (35,380 unique domains) for 1,003 respondents.

Models

We grouped website domains into categories provided by an online service Webshrinker (webshrinker.com) as a dimensionality reduction method. Webshrinker catalogs and scans websites and uses ML algorithms to categorize website domains in Europe and the United States. Since our web tracking data was collected from German participants, we needed a service that could work with German domains. Being able to match as many websites as possible impacts how to complete the picture of the respondents' web browsing, we will have in our data.

Webshrinker managed to match 49,918 unique domains in our web tracking dataset to categories. After applying a one-minute duration and at least five visit thresholds 13,824 unique domains are left in our dataset. Table 3 shows the domain categorization structure with nested data. The domains fall into the 12 groups of categories listed in the first column of the table, and there are several categories (or sometimes only one category) within each group, for instance, sports, blogs, dating, gambling, social media, travel, news, games, and health. Furthermore, each category is represented by domains, which we matched with domain categories available from Webshrinker. Table 3 also shows the number of visits per domain group. As expected, consumption, general, and communication are the most visited domain groups, followed by education, media, and tech services. Domains from more specific lifestyle groups like adult, life, gambling, sports, and social status are among the least visited categories.

We use three different algorithms to test the predictability of website choices, which we measure by summing the visits to each website category and for each respondent: a baseline model, where we estimate the average predictability from a training dataset, linear model, elastic net regression, which is sensitive to multicollinearity (Zou & Hastie, 2005), and random forest, which identifies variables with the most significant explanatory power (Breiman, 2001). For the modeling, we use the

Table 3. Domain categories, groups, examples, and number of visits per group.

Group	Domain category	Top domains	N of visits
Consumption	shopping, business, vehicles, finance, real estate, weapons, alcohol/tobacco	amazon.de, otto.de, bonprix.de, eclipso.de, deutschebank.de, mobile.de, immonet.de, kotte-zeller.de, flaschenpost.de	8,779,614
General	search engines	google.com, web.de, gmx.net	5,654,703
General	information tech, blacklist, filter avoidance, content server, parked	chip.de, microsoft.com, office.com	648,507
Communication	social media, forums, messaging	facebook.com, twitter.com, instagram.com, live.com, msn.com, spin.de	1,750,701
Media	news and media, streaming media, blogs, illegal content, media sharing	bild.de, welt.de, focus.de bs.to, 9gag.com, serienjunkies.org, share-online.biz	1,242,623
Entertainment	games, virtual reality, humor	gameduell.de, youtube.com, netflix.com, twitch.tv	525,399
Entertainment	adult	xhamster.com, planetromeo.com, pornhub.com	488,999
Entertainment	gambling	jackpot.de, tipico.de, bet3000.com	209,355
Life	education, translators	wikipedia.org, uni-mannheim.de, sfgame.de reverso.net	1,719,116
Life	travel, food/recipes, health, drugs	booking.com, bahn.de, chefkoch.de lieferando.de, docmorris.de, zamnesia.com	308,651
Life	sports	flashscore.de, livetv.sx, sport1.de	154,690
Life	job search, religion, dating	indeed.com, stepstone.de, jw.org, finya.de	83,550
TOTAL			21,591,904

functionality of a scikit-learn library in Python, which provides the tools to build predictive models. The library uses random forest and elastic net specification from (Pedregosa et al., 2011). To demonstrate if the chosen algorithms are working, we compare our estimates with benchmark demographics such as gender, income, education, and age (Kosinski et al., 13). Overall, we have 15 questions measuring political attitudes in Table 3, meaning we run 15 regression models.

Cross-validation

To measure its ability to predict political attitudes for each model, we use 10-fold cross-validation (CV) and repeated 3 times, a method for model validation and out-of-sample prediction accuracy (please, see more details on why CV are important in ML in the Online Appendix). The 3×10-fold CV process includes splitting the initial dataset into 10 parts and using nine parts to predict the 10th part. We then run three repetitions of the CV process while randomly splitting the data into 10 folds each time. Repeating the CV three times ensures that the prediction was not an artifact of the selection of the 10 fixed parts. We considered a dependent variable as "predicted" if p-values are less than 0.05 in all the cases in which the CV was repeated. In addition, we calculate R-squared coefficient to measure the model performance in each CV fold. For further validation of the results, we added MSE as well in the Online Appendix. We measure the prediction accuracy of a political attitude with Pearson correlation between the predicted and actual values of dependent variables on the test splits. We conduct 3×10-fold repeated CV for the 15 political attitudes listed in Table 2.

Variable importance

After running all regression models, we calculate Variable Importance (VI) for each feature. VI is a method to rank each covariate by their prediction power in a single model. For VI, we use an R package caret (Classification and Regression Training) and a function varImp, which provides the following VI measure for random forest: "The measure is computed from permuting out-of-bag (OOB) data."[8] Behind the VI measure lies an algorithm that tracks the model's prediction accuracy change and records it after each predictor is included in the model (Kuhn, 2008). Because VI can differ depending on model specification (Fisher, Rudin, & Dominici, 2019), we will focus on VI for the best performance model. In this paper, VI helps us understand which websites of which category has the highest power in predicting each political attitude of interest. VI can also show behavioral patterns based on visit domain categories. In Table 3, we group domain categories by topic, 12 groups in total (see the first column of the table). VI will show if there is a pattern where a specific group of domain categories has the higher predicting power. Since each model includes more than 30 features (each of which assigned to a domain group), we will primarily focus on the features with the highest performing coefficients. However, VI measures are model dependent. We therefore calculate and interpret the VI for our best performing model.

Results

As we described in Method section, we build predictive models where the predicted outcome is a political attitude of interest, and predicting features are visits to website categories. We focus on five dimensions of political attitudes from Table 2: immigration, democracy, climate change, populism, and the EU. The covariates in the models, website categories, are listed in Table 3.[9]

Predictive models

Our focus is on determining the extent to which website categories, when included into a singular model, can account for the variance measured by R-squared or R^2. Following (Stachl et al., 2020) we compare the performance of three regression models as described in Method section: Linear model, Elastic Net, and Random Forest against average prediction on a test data. We also measure performance with Pearson correlation (r) and with MSE (Mean Squared Error) between actual and predicted values for each political attitude. MSE, unlike Pearson's correlation, is better suited for assessing the distance between predicting models and the actual values r (Waldmann, 2019). This

method also demonstrates the average discrepancy, measured in scale points of attitudes that the models display. We report MSE in the Online Appendix.

Figure 2 reports the prediction performance of baseline (linear models), Elastic Net, and Random Forest from repeated cross-validation for each political attitude of interest. We also included a model performance for socio-demographic variables: Gender, income, education, and age –– a common practice in ML literature that deals with social science concepts (Kosinski, Stillwell, & Graepel, 2013; Stachl et al., 2020). Comparing the model performance for political attitudes with socio-demographic variables helps assess ML methods' validity. On average, all three regression models (baseline, Elastic Net, and Random Forest) perform moderately compared to gender or age. Across most political attitudes, the random forest method is the best-performing algorithm compared to Linear and Elastic Net algorithms. However, even with one of the most sophisticated algorithms, such as Random Forest, the Pearson correlation coefficients (r) are significant within 2.5% and 97.5% quantiles only for two features out of 15: interest in politics and support for a democratic political system. The correlation coefficients are modest, with a median $r = 0.15$ for interest in politics, $r = 0.13$ for support in the democratic political system. The coefficients are comparable to those that are reported in the existing literature that deals with social science concepts measured with surveys (Kosinski, Stillwell, & Graepel, 2013; Stachl et al., 2020). Random Forest and Elastic Net models were also able (within 25% and 75% quantiles) to signal populist attitudes, attitudes toward Islam, support for free elections, satisfaction with democracy, and trust in a national parliament.

However, the models' performance is not stable. We increased the number of repeats of 10-folds CV from 3 to 10, which is a stricter robustness test. The effect for interest in politics persisted while for the attitude "support for a democratic political system" did not survive. The results from 10×10-fold cross-validated models together with hyperparameters configurations are reported in the Online Appendix. We also added gradient boosting model

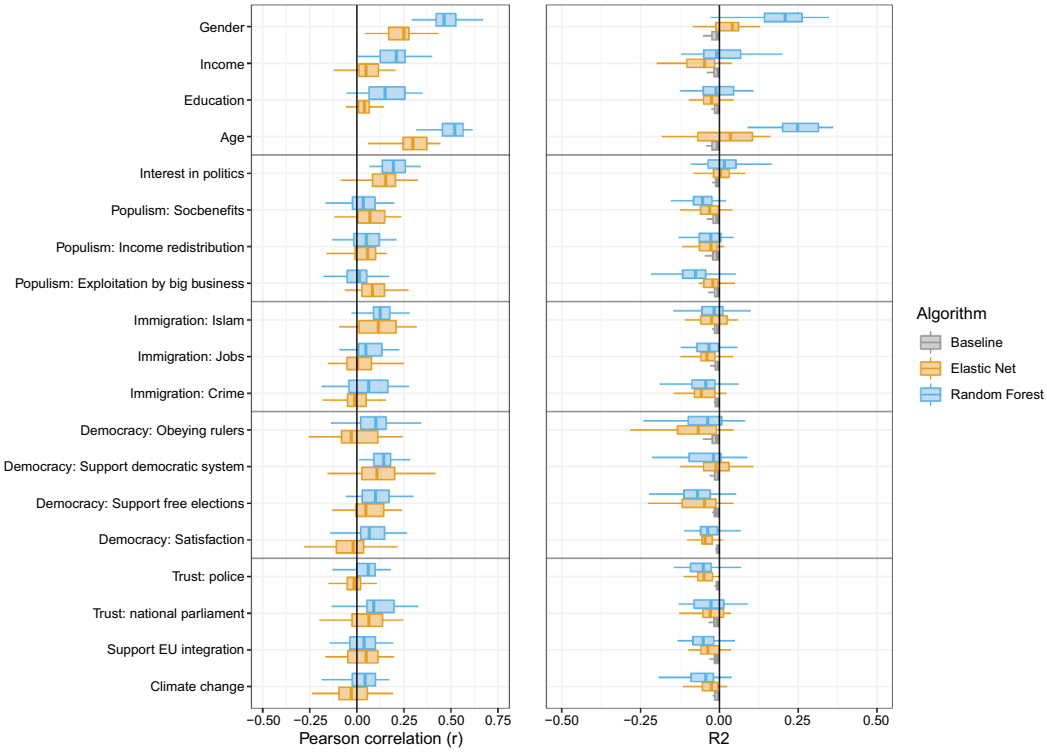

Figure 2. Box and whisker plot of prediction performance measures from repeated cross-validation for each political attitude and demographic category. The middle symbol represents the median, boxes include values between the 25% and 75% quantiles, and whiskers extend to the 2.5% and 97.5% quantiles.

to show if more advanced algorithm would be capable to improve the predictions. The results improved only slightly further demonstrating the challenge of predicting political attitudes based on web tracking data and showing that this kind of data can offer only suggestive evidence.

Although some statistically significant predictions were reached, R^2 is small and negative. Negative R^2 contradicts its initial definition (Colin Cameron & Windmeijer, 1997). This suggests that the models are unable to capture robust and convincing connections between features and the outcome because the features are not informative enough, and, hence, are affecting their performance on test data. Nevertheless, our results for R^2 are consistent with the existing literature dealing with survey-based feature predictions. In (Stachl et al., 2020), R^2 for all models, the baseline model, Random Forest, Elastic Net are negative for many features. In (Panicheva et al. 2022), the R^2 coefficient for the Elastic Net model predicting subjective well-being is 0.11, although a confidence interval is not provided. Brandenstein (2022) reports $R^2 = 0.17$ for a Random Forest model that predicts beliefs in conspiracy theories. Praet, Guess, Tucker, Bonneau, & Nagler (2021) reports Pseudo $R^2 = 0.28$ but without a cross-validation. On the contrary, both measures of our models' performance r and R^2 are larger for socio-demographic variables, similar to Kosinski, Stillwell, and Graepel (2013), which means that the performance of the selected models is challenged specifically when applied to political attitudes. However, according to Chicco, Warrens, and Jurman (2021), it is still more informative than other metrics used in regression model performance evaluation.

Variable importance

Although, the model performance is not stable and offers suggestive predictions (within 25% and 75% quantiles) and interpretations should be treated with cautious, exploring what website categories are at the front of the predictive model may help offer the direction for the further research. Figure 3 shows the variable importance rank for each of the best random forest models predicting political attitudes from Figure 2. We also assigned predicting variables to higher level topics: Issues (trust in institutes, climate change, EU integration), Democracy, Immigration, and Populism. Our grouping strategy here deviates from the one in Table 2. We separated trust variables from democracy to have a clear group measuring attitudes toward democracy. Trust is only remotely related to attitudes toward democracy.

Variable importance ranks covariates by the contribution each of these covariates makes to predicting the accuracy of each model. Each square represents a covariate, such as visits to a website category from the first row of Table 3 and is colored accordingly. We use a color-coding to visually demonstrate if there are observable predictive patterns and what website categories form those patterns. Since we focused on the top performing categories, we applied the fading visual effect to the plot to reflect the decreasing importance of these categories.

We focus on behavioral patterns across all attitudes of interest. Overall, life and general purposes websites are the most potent variables in models for predicting attitudes toward immigration, populist and issue-related attitudes, and communication and media websites are the most substantial contributors in the models' predicting attitudes toward democracy. Entertainment websites, which include games, gambling, adult content, and humor, are among the weakest predictors.

The observed patterns have two social science implications. First, the variable importance patterns indicate that media and communication websites such as news and social media hold low predicting power in models that are predicting issue-related or populist attitudes and attitudes toward immigration. This finding contradicts the existing literature focusing on the role of new or social media on populist attitudes or attitudes toward immigration, climate change, or EU integration. Our findings suggest that these attitudes are better predicted with lifestyle or general purposes websites such as shopping, business, or search engines, which reflect respondents' social status, financial conditions, and other interests that, when combined, might affect, or even form the attitudes. And second, media and communication websites displayed a suggestive prediction pattern in relation to attitudes toward democracy. However, specific mechanisms behind these

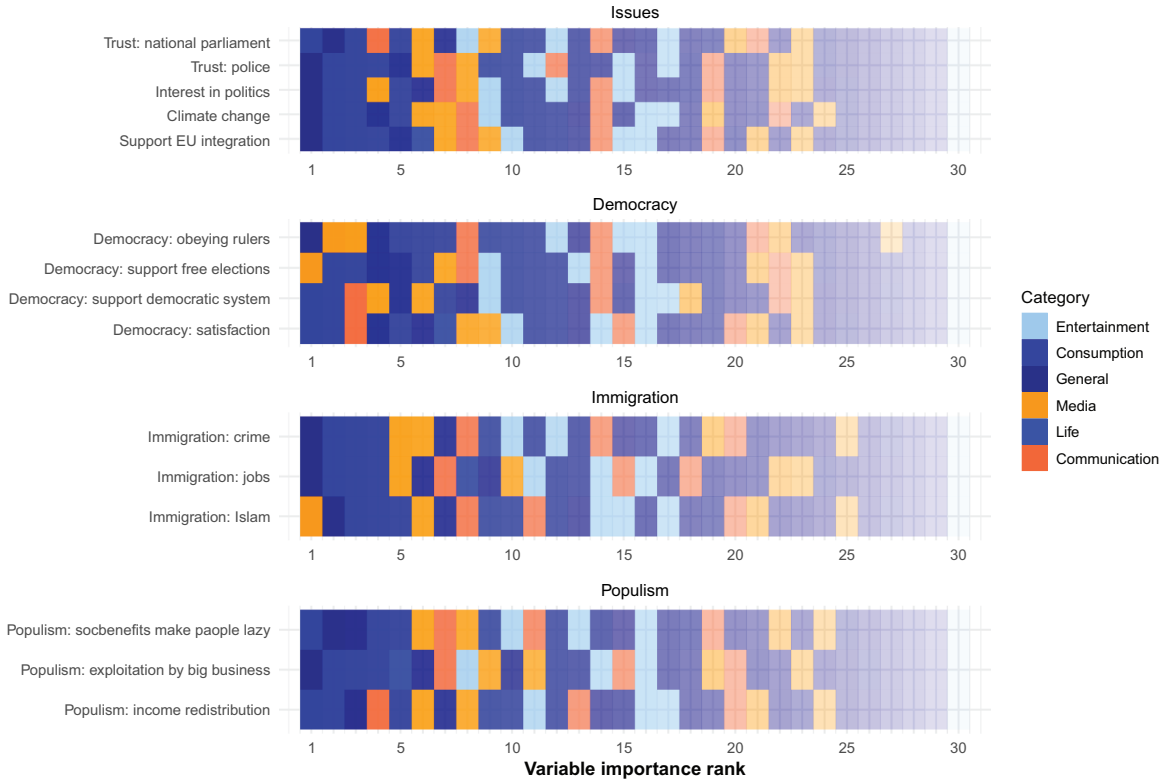

Figure 3. Domain categories ranked by importance in the Random Forest model for attitudes toward policy issues, democracy, immigration, and populism. The fading effect on the plot represents the decrease in the importance of each domain category since the top five domains bring the most significant contribution to prediction accuracy. The color represents two palettes – orange and blue – in order to distinguish between domains related to media/communication and consumption/life-style. To see what specific domain category is behind each square, we made an interactive plot, which can be downloaded from an anonymous OSF repository of this paper: https://osf.io/us4dz/and in the supplementary materials of this manuscript. Additionally, the list of variables for significant models is also available in the online appendix on page 13 and 14.

observed behavioral patterns need further exploration.

Despite offering largely suggestive predictions, Random Forest regression model was able to predict interest in politics. Figure 3 shows that, as expected, media websites play an essential role in predicting interest in politics. We also plot the list of variables ranked by importance in Figure E.3 in the Online Appendix. The websites related to shopping, business, and finance contributed to the prediction accuracy just as much as media websites, suggesting that day-to-day life choices may affect attitudes and media consumption. However, social media, news, and streaming media significantly predict support for a democratic political system, although based on 3×10-fold CV model (see Figure E2 in the Online Appendix). Further research is needed to explore the mechanisms since each category represents specific websites. More granular data analysis will show why visits to business-related websites are associated with populist attitudes and visits to media websites predict attitudes toward democracy.

Discussion

In this paper, we combined surveys with observational data collected from tracking online browsing of 1,003 German individuals. Combining these two types of data, we offer an exploratory analysis of whether big data and ML algorithms can help infer voters' political features, specifically political attitudes measured with surveys. We tested the predictive performance of three ML algorithms: random forest, elastic net, and gradient boosting, supplied with 10-fold repeated cross-validation. Specifically, we built 15 models predicting four groups of political attitudes: Attitudes toward immigration, democracy, the EU, and populist attitudes. We found mixed evidence of the

predictability of political attitudes from web tracking data based on our best-performing random forest model.

The model predicted interest in politics and attitudes toward democratic systems. Despite the limitations of our data and measurements, the results are compatible with previous studies of individuals' personalities with larger samples. Our highest predictions for interest in politics and attitudes toward democracy vary from $r = 0.09$ to 0.15 compared to 0.17 for "satisfaction with life" also measured on a 5-point scale in Kosinski, Stillwell, and Graepel (2013), [0.20, 0.40] average estimation in Stachl et al. (2020) and in Funder and Ozer (2019). The predictability of interest in politics can be explained by more specific website domain visits, which can be associated with it, such as media outlets and other political content. Trust toward political institutions, however, are more abstract and cannot be attributed to specific websites.

We also explored what model features impact the prediction of political attitudes. Two main categories of websites demonstrated observable patterns: General-purpose and consumption websites (e.g., business and shopping) and media and communication websites.

General-purpose websites (e.g., search engines) and consumption websites (e.g., shopping, real estate, finance, etc.) display a suggestive predictive pattern for issue-related (e.g., climate change, the EU integration, immigration) and populist attitudes. One potential reason for these associations is that it is consistent with the nature of these attitudes since they are related to social benefits, business, and income in case of populist attitudes, taxes, and other economic changes in case of climate change policies and EU integration, as well as trust in the police. Attitudes toward immigration could also be affected by social status and life circumstances reflected in consumption-related websites, primarily if immigration is associated with crime and jobs. This is something that respondents might experience rather than receive information from news or social media. Further in-depth exploration of the web tracking data is needed to understand what kind of websites, including web search queries or YouTube video topics, drive the predicting effects.

The second group that stands-out in the models is media and communication. Visits to these websites are correlated to attitudes toward democracy. Media websites are also the top websites that are predicting two attitudinal items, such as perception of Islam and support for free elections, while they are ranked fourth in predicting interest in politics. The role of media domains in predicting some political attitudes, specifically attitudes toward democracy, adds to the literature on media effects and the role of news in politics. This finding contradicts the literature arguing that media have limited effect on political behavior or attitudes. The finding also shows methodological potential of ML models: These advanced ML methods can help to learn about political behavior or attitudes from large amount of data and avoid manual website labeling. Nevertheless, as mentioned before, an in-depth exploration of website domains and the mechanisms that each domain might uncover is needed.

The third category of websites we anticipated would exhibit significant effects in the models –– entertainment and lifestyle websites –– ultimately did not emerge as strong predictor. This does not confirm hypotheses in the existing literature that economic frustration (if we associate gambling with economic hardship) could be responsible for populist attitudes. Our findings are consistent with Praet, Guess, Tucker, Bonneau, and Nagler (2021) that political orientations are moderately reflected in lifestyle choices. One potential reason for null effect of this group is that these websites could represent the opposite mechanisms. Respondents may visit gambling websites because of economic hardship or, the opposite, because of excessive financial sources and, therefore, the effect may not be as sounding as if the group represent a single-meaning mechanism. This, in turn, raises the issue of mechanisms in the observed associations between political attitudes and website visits measured based on web tracking data. Further in-

depth website categorization is needed to ensure consistency in the mechanism that each website domain accounts for.

In general, this paper broadens the scope of political science literature concerning the methodology and utilization of predictive modeling within the discipline. The paper additionally presents an algorithm for implementation of predictive modeling based on the combination of web tracking and survey data. Moreover, it provides theoretical foundations and suggests for potential directions for explanatory research. Lastly, the findings of the paper have policy making and normative implications.

Initially, from a broader perspective within political science, this paper's findings indicate the challenge in identifying political attitudes from web tracking data. This has two implications: (1) putting attitudes on a latent left-right ideology scale, we did not find observable differences in website visits among respondents with pro- or anti-immigration attitudes, pro- or anti- climate change policies, which is consistent with Praet, Guess, Tucker, Bonneau, and Nagler (2021) suggesting that political polarization is not reflected in the lifestyle but rather limited to partisan news preferences; (2) contrary to Kosinski, Stillwell, and Graepel (2013), which shows that Facebook likes could be used to measure users' personality traits, our advanced ML models were able to retrieve only suggestive signals about what attitudes individuals might have based on their website visits' patterns, which implies that surveying is still the most reliable method to measure attitudes. However, our data is bounded by a specific timeframe and can potentially show different results over time. We made the replication materials available on an OSF repository for testing predictive capabilities of web tracking data in different time frames and political contexts.

From a normative perspective, our study reveals that despite the vast amount of available data, only a limited amount of information related to political attitudes can be harvested from individuals' browsing histories. Hence, contrary to recent developments in digital privacy policies, our findings do not substantiate the assumption that sensitive political information can be extracted from digital trace data. This also challenges the notion that such data could be utilized by advertising distributors like Google or by politicians for political microtargeting.

Although we performed several robustness tests of our models, the study has several limitations that could affect the results. Our findings represent a conservative estimation of the predictive power of web tracking data. Our estimation is based on bounded ordinal variables standard in political science to measure political attitudes, but only sometimes informative for predictive ML models (Seveso, Campagner, Ciucci, & Cabitza, 2020). We also do not use data from mobile devices, which could potentially reveal more patterns from individuals' daily life. With larger samples, better representations of URLs that are not limited to domains, alternative continuous instead of categorical measures of attitudes, and various model specifications, including hyperparameters configurations beyond the ones considered in our grid search that improves the model performance, we expect the findings to gain more accuracy and robustness. Our findings also might change through time. Therefore, in this paper, we offer the algorithm to replicate this analysis for future research. All materials for the replication of this paper with new data can be found in repositories available on open-source platforms OSF.

Notes

1. https://www.nytimes.com/interactive/2020/10/27/upshot/biden-trump-poll-quiz.html
2. https://whotargets.me/en/
3. https://blog.google/technology/ads/update-our-political-ads-policy
4. https://about.fb.com/news/2020/01/political-ads/
5. https://business.twitter.com/en/help/ads-policies/prohibited-content-policies/political-content.html
6. A full literature review is available in the Online Appendix.
7. https://www.pewresearch.org/politics/2014/06/12/section-3-political-polarization-and-personal-life/.
8. https://topepo.github.io/caret/variable-importance.html.
9. We provide the exploratory analysis of base-line OLS regressions in the Online Appendix.

Acknowledgments

We would like to express our gratitude to the editors and anonymous reviewers for their valuable suggestions which

significantly helped improve the manuscript. We thank the faculty of GESIS CSS Department and the TUM School of Social Sciences and Technology for helpful comments.

Disclosure statement

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

Funding

The Volkswagen Foundation funded this project, Grant #94758

Notes on contributors

Nora Kirkizh, M.A., is a PhD student in Political Science, at the School of Social Sciences and Technology at the Technical University of Munich.

Roberto Ulloa is a Post-Doctoral Researcher at GESIS — Leibniz Institute for the Social Sciences.

Sebastian Stier is a Scientific Director of the Department Computational Social Science at GESIS — Leibniz Institute for the Social Sciences and Professor of Computational Social Science at the School of Social Sciences, University of Mannheim.

Jürgen Pfeffer, Full Professor of Computational Social Science at the School of Social Sciences and Technology at the Technical University of Munich.

References

Al Halabi, W. S., Kubat, M., & Tapia, M. (2007). Time spent on a web page is sufficient to infer a user's interest. In *Proceedings of the Third IASTED European Conference on Internet and Multimedia Systems and Applications*, USA: ACTA Press, (p. 41–46).

Althoff, T., Nilforoshan, H., Hua, J., & Leskovec, J. (2022). Large-scale diet tracking data reveal disparate associations between food environment and diets. *Nature Communications*, *13*(267). doi:10.1038/s41467-021-27522-y

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? understanding and improving the accuracy of self-reported measures of internet use. *Communication Methods and Measures*, *11* (3), 173–190. doi:10.1080/19312458.2017.1317337

Brandenstein, N. (2022). Going beyond simplicity: Using machine learning to predict belief in conspiracy theories. *European Journal of Social Psychology*, *52*(5–6), 910–930. doi:10.1002/ejsp.2859

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Cerina, R., & Duch, R. (2020, March). Measuring public opinion via digital footprints. *International Journal of Forecasting (To Appear)*, *36*(3), 987–1002. doi:10.1016/j.ijforecast.2019.10.004

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination r-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *Peer Journal of Computer Science*, *7*(623), e623. doi:10.7717/peerj-cs.623

Colin Cameron, A., & Windmeijer, F. A. (1997). An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, *77*(2), 329–342. doi:10.1016/S0304-4076(96)01818-0

DellaPosta, D., Shi, Y., & Macy, M. (2015). Why do liberals drink lattes? *American Journal of Sociology*, *120*(5), 1473–1511. doi:10.1086/681254

Dixit, R., Chinnam, R. B., & Singh, H. (2020). Artificial intelligence and machine learning in sparse/inaccurate data situations. In *2020 IEEE Aerospace Conference*, Big Sky, MT, USA, 1–8.

Engel, D., Hüttenberger, L., & Hamann, B. (2012). A survey of dimension reduction methods for high-dimensional data analysis and visualization. In C. Garth, A. Middel, & H. Hagen (Eds.), *Visualization of large and unstructured data sets: Applications in geospatial planning, modeling and engineering - proceedings of IRTG 1131 workshop 2011* (Vol. 27, pp. 135–149). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Englehardt, S., & Narayanan, A. (2016). Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (p. 1388–1401). New York, NY, USA: Association for Computing Machinery.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20* (177), 1–81.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. doi:10.1177/2515245919847202

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, *4*(5), 472–480. doi:10.1038/s41562-020-0833-x

Hofgesang, P. I. (2006). Relevance of time spent on web pages. In *Proceedings of KDD Workshop on Web Mining and Web Usage Analysis, in conjunction with the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Kerna, M. L., McCarthyb, P. X., Chakrabartyb, D., & Rizoiud, M.-A. (2019). Social media-predicted personality traits and values can help match people to their ideal jobs. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(52), 26459–26464. doi:10.1073/pnas.1917942116

Kirkizh, N., Froio, C., & Stier, S. (2022). Issue trade-offs and the politics of representation: Experimental evidence from four European democracies. *European Journal of Political Research*, *62*(4), 1009–1030. doi:10.1111/1475-6765.12558

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802–5805. doi:10.1073/pnas.1218772110

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. doi:10.18637/jss.v028.i05

Leist, A. K., Klee, M., Kim, J. H., Rehkopf, D. H., Bordas, S. P. A., Muniz-Terrera, G., & Wade, S. (2022). Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Science Advances*, *8*(42), eabk1942. doi:10.1126/sciadv.abk1942

Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2020). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*, *38*(5), 616–632. doi:10.1177/0894439319828012

Panicheva, P., Mararitsa, L., Sorokin, S., Koltsova, O., & Rosso, P. (2022). Predicting subjective well-being in a high-risk sample of Russian mental health app users. *EPJ Data Science*, *11*(1), 21. doi:10.1140/epjds/s13688-022-00333-x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Praet, S., Guess, A. M., Tucker, J. A., Bonneau, R., & Nagler, J. (2021). What's not to like? Facebook page likes reveal limited polarization in lifestyle preferences. *Political Communication*, *39*(3), 1–28. doi:10.1080/10584609.2021.1994066

Scharkow, M. (2016). The accuracy of self-reported internet use—a validation study using client log data. *Communication Methods and Measures*, *10*(1), 13–27. doi:10.1080/19312458.2015.1118446

Seveso, A., Campagner, A., Ciucci, D., & Cabitza, F. (2020). Ordinal labels in machine learning: A user-centered approach to improve data validity in medical settings. *BMC Medical Informatics and Decision Making*, *20*(142). doi:10.1186/s12911-020-01152-8

Shi, F., Shi, Y., Dokshin, F. A., Evans, J. A., & Macy, M. W. (2017, April). Millions of online book co-purchases reveal partisan differences in the consumption of science. *Nature Human Behaviour*, *1*(4), doi:10.1038/s41562-017-0079

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., & Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, *117*(30), 17680–17687. doi:10.1073/pnas.1920484117

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503–516. doi:10.1177/0894439319843669

Stier, S., Kirkizh, N., Froio, C., & Schroeder, R. (2020). Populist attitudes and selective exposure to online news: A cross-country analysis combining web tracking and surveys. *The International Journal of Press/politics*, *25*(3), 426–446. doi:10.1177/1940161220907018

Waldmann, P. (2019). On the use of the Pearson correlation coefficient for model evaluation in genome-wide prediction. *Frontiers in Genetics*, *10*. doi:10.3389/fgene.2019.00899

Wiedemann, A. (2023). The electoral consequences of household indebtedness under austerity. *American Journal of Political Science*. doi:10.1111/ajps.12708

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x