Predicting Political Attitudes from Web Tracking Data: Machine Learning Approach

Online Appendix

Contents

A Validation of web tracking panel

We 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. For instance, they might have started to visit more news websites to learn more about political issues or the other way around; in short, respondents who are being tracked might be more careful in revealing their political interests and inclinations or vice versa. We correlated visits made by our panel to media websites with ground truth data on the visits of the German general population recorded by the "Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V." (IVW), audit bureau of media circulation in Germany. The correlation between the ranking of news sites visited by our German tracking panelists and the IVW data is strong ($\rho = 0.73$). We report the correlation in the Online Appendix, Figure A.1. These evaluations give us confidence that our tracking data provides a reasonably accurate representation of visited websites by internet users in Germany.

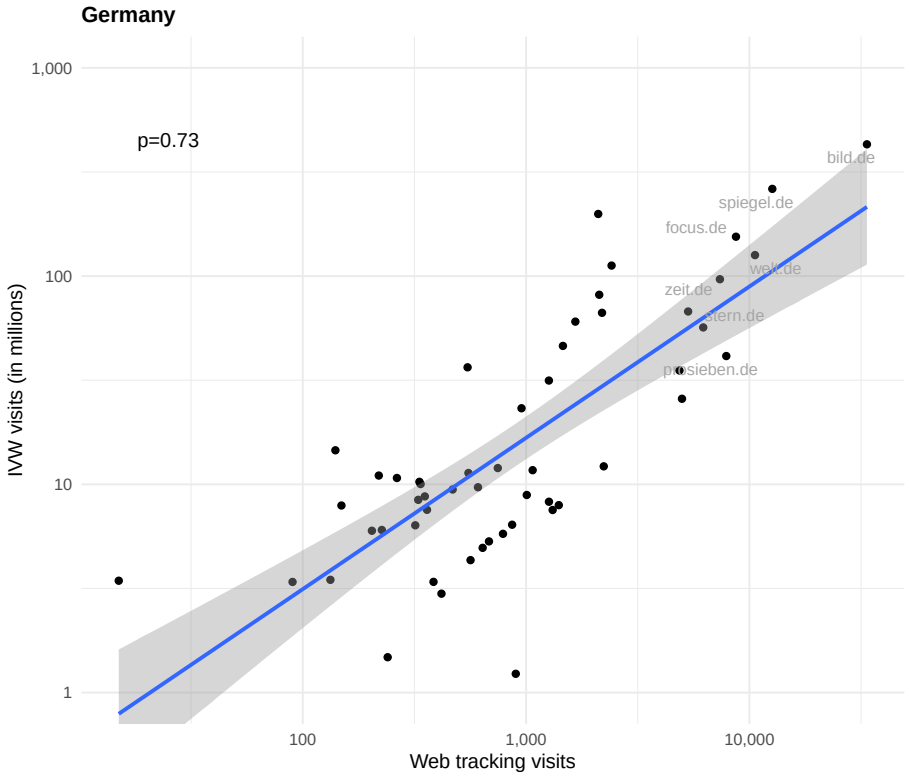

Figure A.1: Comparison of IVW domain visit rankings and news domains visited by our web tracking panelists in Germany. Both axes are logged; p is Spearman's rank correlation.

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. To identify a potential "opt-in bias", we implemented the same privacy attitude battery in a sample of German participants drawn from the regular online access panel of the market research company without web tracking. In total, we sampled 1,000 participants and matched German population margins for gender, age, and education. Respondents have been presented with the following statements and asked about their (dis)agreement on a five-point scale: "Personalized advertising makes me afraid"; "I am concerned about how much data there is about me on the Internet"; and "My privacy on the Internet does not matter to me." Figure A.2. in the Online Appendix presents the results of a comparison of privacy attitudes. The figure shows that there are minor differences in the privacy attitudes of online panelists with and without web tracking technology installed, which brings the results of this paper closer to generalizability.

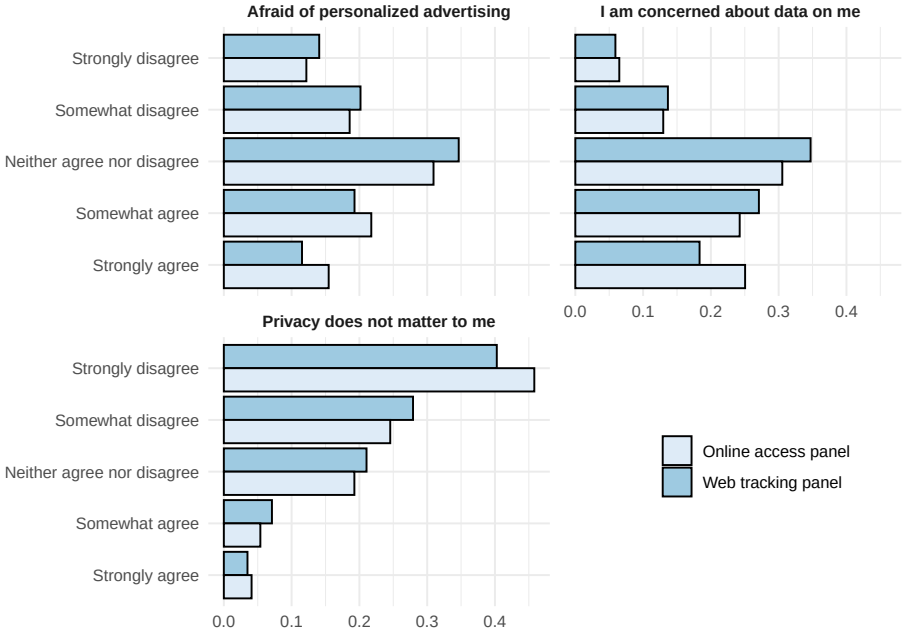

Figure A.2: Comparison of privacy attitudes in a survey of German online access panelists and web tracking panelists.

B Survey data

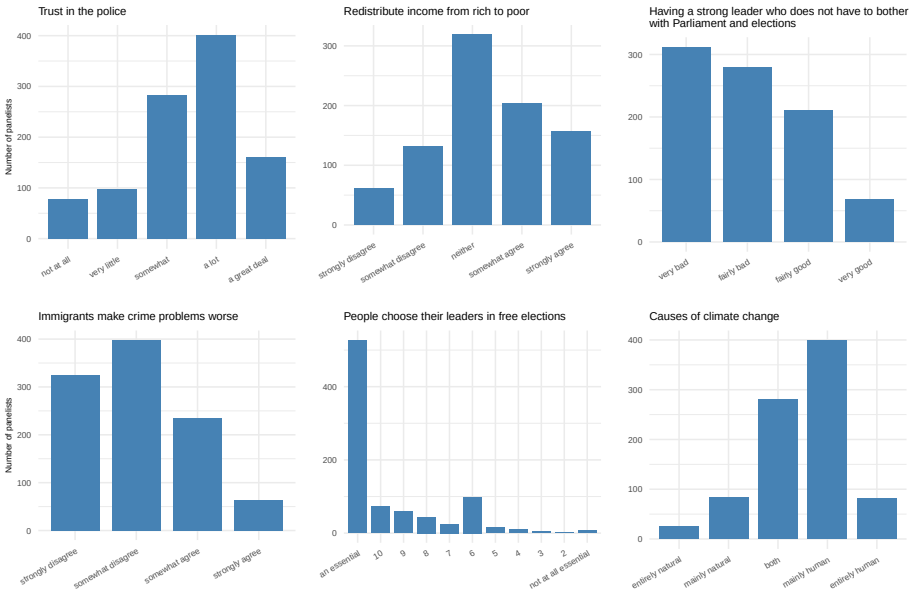

Figure B.1: Distribution of selected survey items measuring political attitudes from each topic: trust in institutions (e.g. the police), populist attitudes (redistribute incomes from rich to poor), democracy (demand in strong leader, and choosing the leader in free elections), immigration (immigrants and crime), climate change.

C Political attitudes and visits to website categories: Distributions

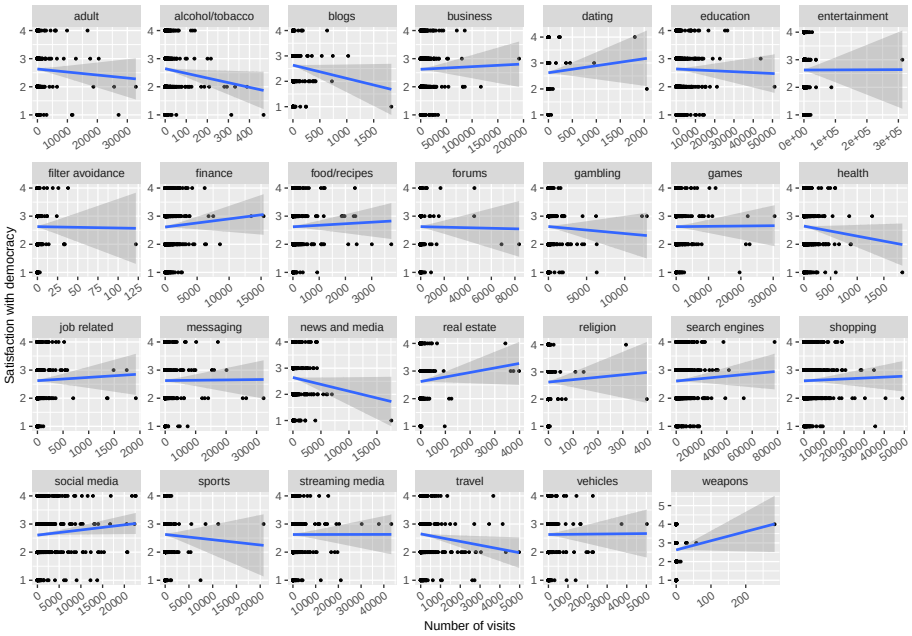

Figure C.1: Association of visits to different categories of websites and satisfaction with democracy measured on the scale from 1 - very unsatisfied to 4 - very satisfied. Black dots represent data points, the blue line is a linear function and gray area is 95% confidence interval.

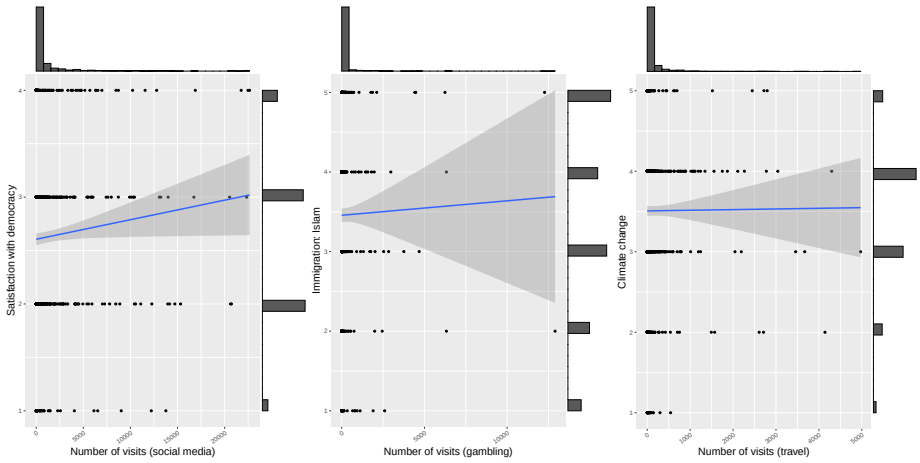

Figure C.2: Association of selected political attitudes and visits to social media. Black dots represent data points, the blue line is a linear function and gray area is 95% confidence interval.

D Political attitudes and domain categories: Exploratory analysis

One of the caveats of social science concepts like political attitudes is the scale they are measured on. Political attitudes are predominantly measured on ordinal scales such as the four-point agree-disagree scale, five-point Likert scale, and ten-point scale. Even though numerical and ordered, these scales are less informative for a machine learning (Ml) algorithm than continuous variables like age. The ML community has limited application of ML algorithms for predicting self-reported features such as attitudes (Salganik et al., 2020; Mõttus et al., 2020; Azucar et al., 2018; Panicheva, Polina et al., 2022; Brandenstein, 2022; Kim et al., 2022; Han, 2022; Leist et al., 2022; Pargent & Albert-von der Gönna, 2018), which are very popular in the social science.

Figure B.1 illustrates the distribution of selected political attitudes with an example of each type of scale. Except for a few questions, the respondents generally placed their answers in the middle of the scale, which allows us to expect the normality of residuals in further regression analysis. To explore the descriptive association between political attitudes and domain categories, we plot "satisfaction with democracy" against each domain category with no control variables (see Figure C.1). We also describe this association with linear function represented by a blue line and 95% confidence interval represented by the grey area. Each plot of the figure shows that most of the data points are gathered on the left-hand side of the x-axis because only a few respondents made more than 50 visits to websites of a specific category. The skewed distribution of domain visits per respondent is observable in Figure C.2.

We focus on selected domain categories and "satisfaction with democracy", attitudes toward immigration, and attitudes toward climate change. The distribution of website visits is significantly skewed towards zero, with very few respondents who made more than 100 visits to a domain category. Nevertheless, Figure C.1 displays a change in the slope of a linear regression depending on the number of website visits. However, confidence intervals increase with the increase of website visits because the function has fewer observations when moving towards more significant numbers of website visits. Large confidence intervals signify the statistical significance of the associations. Even if statistically significant, the associations in Figure C.1. or C.2. do not consider other domain categories as control variables or can be sensitive to outliers. We run linear regression models for each political attitude against all categories to identify domain categories with significant explanatory power.

We run OLS regressions, where each attitudinal question from Table 2 in the main manuscript is placed against all domain categories in Table 3. Estimates of average effect size for each domain category and 95% confidence intervals are plotted in Figure D.1 for selected attitudes from each group (climate change, democracy, immigration, populism, and the EU) and in Figure E.3 for the rest of the questions. Statistically significant effects are in red, and nonsignificant — in black. Running OLS models with all domain categories in one model helps

to identify features with the most significant explanatory power. As a result, this allows preliminary assessment of selected features' potential to predict variables of interest. Figure D.1 shows that very few domain categories "survive" the test when placed together into one model. On average, two out of 34 domain categories per model are statistically associated with a political attitude of interest. In contrast, some models have a single or no statistically significant domain category. The difference in the number of statistically significant features across models signifies different levels of the explanatory power of domain visits to predict specific political attitudes. Attitudes related to democracy, immigration, and interest in politics appear to be the most favorable attitudes to be identified from categorized website visits. Trust and the EU attitudes are the least promising, with no or a single statistically significant domain category.

Although statistically significant, some estimates have large confidence intervals. For instance, domain category weapon or religion, which have a limited amount of visits in the data. Among nonsignificant features, there are many domain categories with large confidence intervals. Large confidence intervals point to a possibility that outliers might drive some effects. This observation is also consistent with patterns of slope change in Figure C.1 and C.2. We address this potential issue below by applying more robust ML algorithms, which are not sensitive to outliers and multicollinearity.

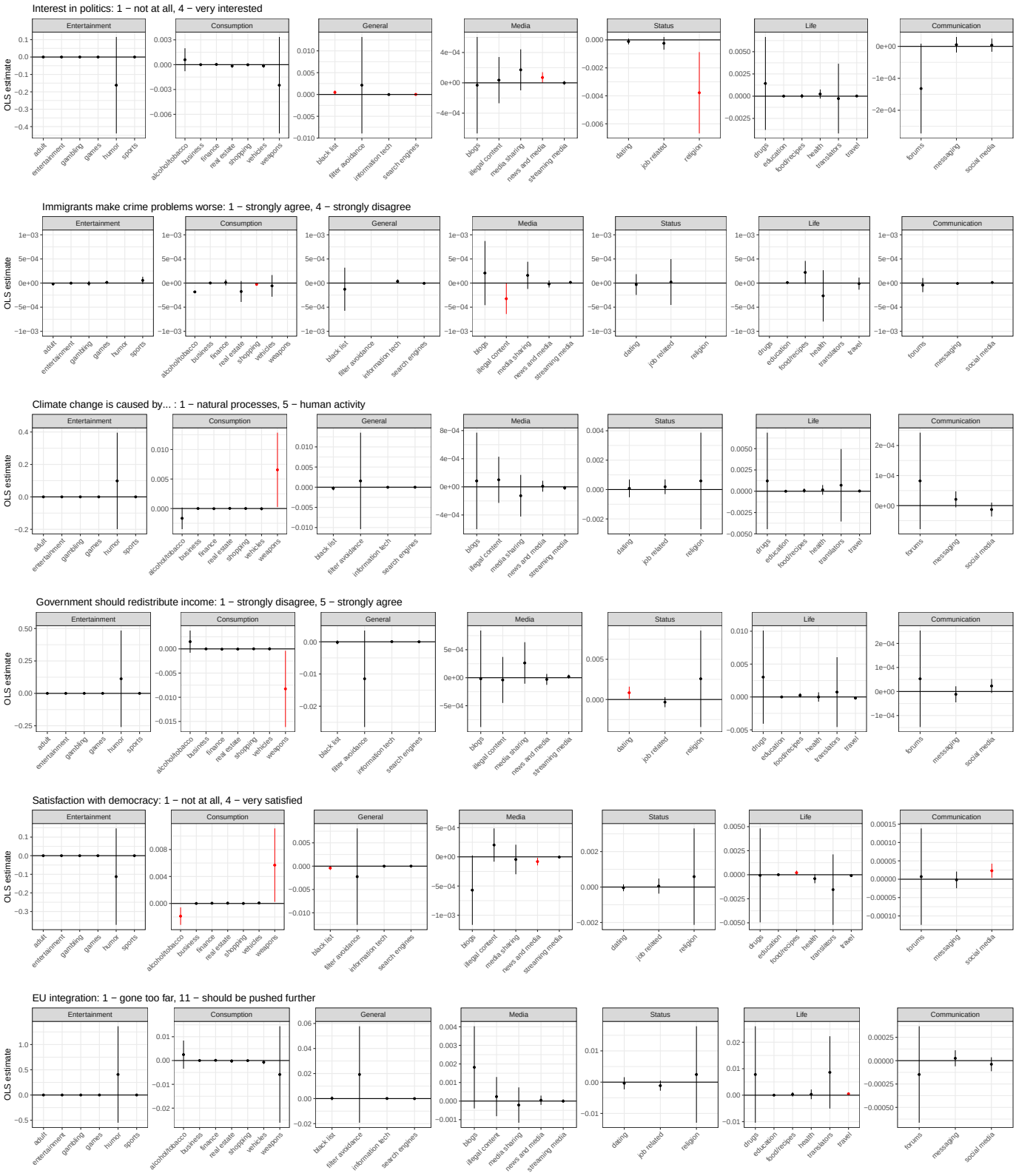

Figure D.1: OLS estimates for each domain category plotted against selected political attitudes

The bars around estimates on Figure D.1 represent confidence intervals. Statistically significant effects are in red (p-value < 0.05), and nonsignificant effects are in black (p-value > 0.05). Figure D.1. in the Online Appendix shows estimates for the rest of the political attitudes. Estimates without confidence intervals mean that they are too small to be visible on the plot, and estimates that are missing on the plot mean that their confidence intervals are too large to fit into the plot but are nevertheless not statistically significant. The range for the y-axis prioritizing statistically significant estimates, and if they are small, the plotting requires a small range for the y-axis. The interpretation of statistically significant estimates is that respondents visiting websites related to alcohol/tobacco are not satisfied with democracy (the model in the second row from the bottom) because the domain category alcohol/tobacco has a negative association with satisfaction with democracy. Importantly, this is a descriptive, not causal, relationship. These OLS models do not suffice for causal interpretation — for instance, ten visits to alcohol/tobacco websites decrease satisfaction with democracy from somewhat satisfied to not satisfied — because in this paper, we aim to predict or identify political attitudes based on websites visited by respondents rather than to explain what websites have effects on political attitudes.

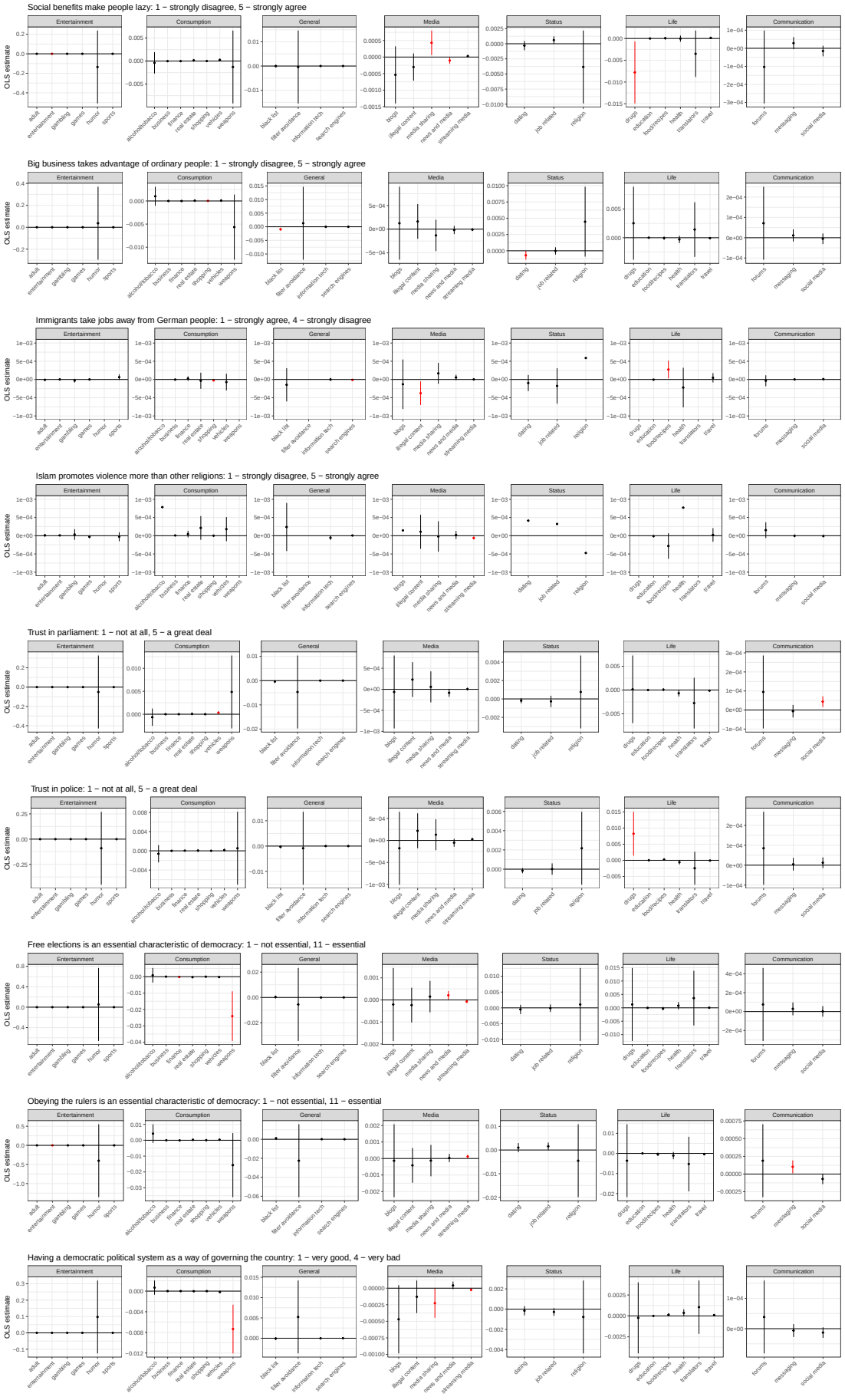

Figure D.2: OLS estimates for each domain category plotted against selected political attitudes. (Cont.)

E Variable importance

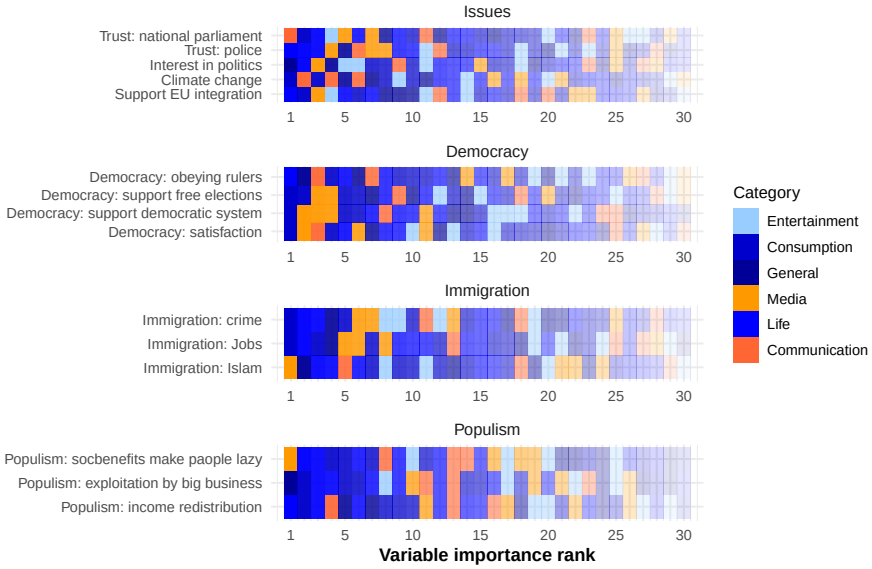

Figure E.1: Variable importance rank from linear regression models.

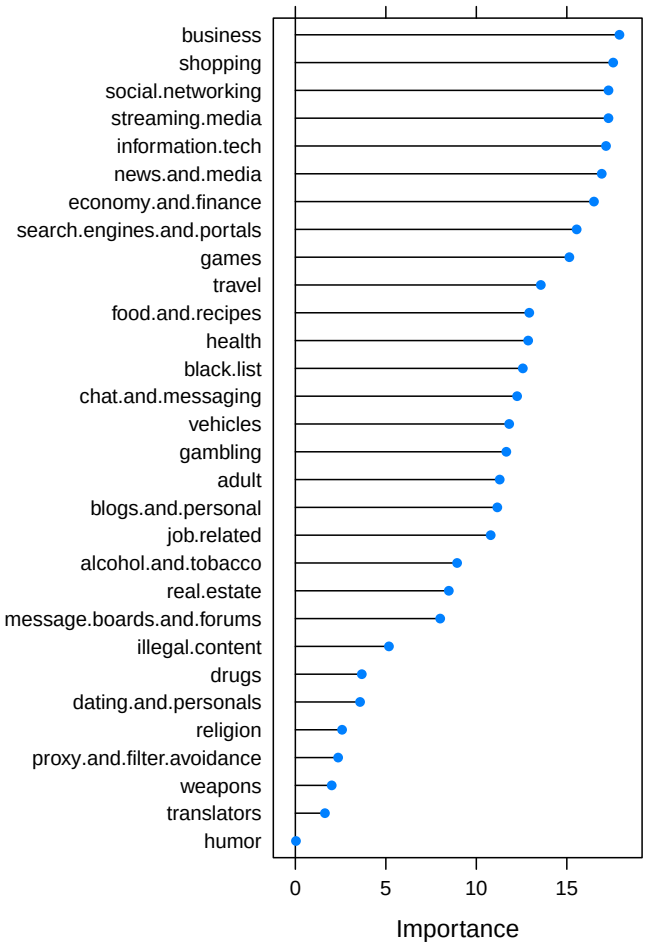

Figure E.2: Variable importance rank from Random Forest model for support for democratic political system.

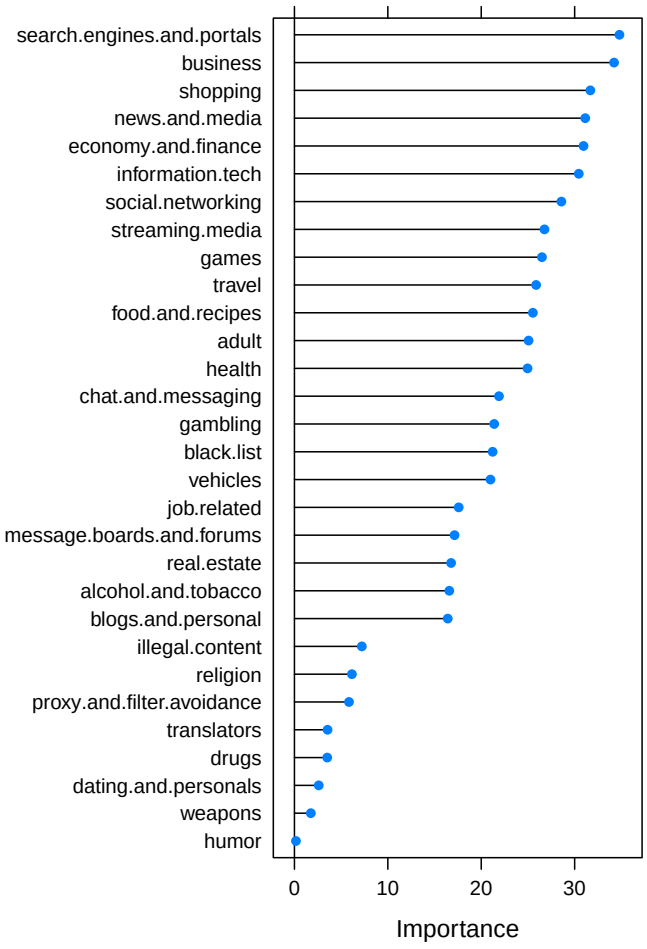

Figure E.3: Variable importance rank from Random Forest model for interest in politics.

References

Azucar, D., Marengo, D., & Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personality and Individual Differences*, *124*, 150-159. doi: 10.1016/j.paid.2017.12.018

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: https://doi.org/10.1002/ejsp.2859

Han, S. (2022). An analysis of koreans' attitudes towards migrants by application of algorithmic approaches. *Heliyon*, *8*(8), e10087. doi: https://doi.org/10.1016/j.heliyon.2022.e10087

Kim, D., Chung, C. J., & Eom, K. (2022). Measuring online public opinion for decision making: Application of deep learning on political context. *Sustainability*, *14*(7). doi: 10.3390/su14074113

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

Mõttus, R., Wood, D., Condon, D. M., Back, M. D., Baumert, A., Costantini, G., ... Zimmermann, J. (2020). Descriptive, predictive and explanatory personality research: Different goals, different approaches, but a shared need to move beyond the big few traits. *European Journal of Personality*, *34*(6), 1175-1201. doi: 10.1002/per.2311

Panicheva, Polina, Mararitsa, Larisa, Sorokin, Semen, Koltsova, Olessia, & Rosso, Paolo. (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

Pargent, F., & Albert-von der Gönna, J. (2018). Predictive modeling with psychological panel data. *Zeitschrift für Psychologie*, *226*(4), 246–258. doi: https://doi.org/10.1027/2151-2604/a000343

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., ... McLanahan, S. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. doi: 10.1073/pnas.1915006117