Hertie School

Master's Thesis

PREDICTING VOTING BEHAVIOR

COMBINING NEWS URL TRACE DATA AND DEMOGRAPHICS

Authored by

Ray Hossain

Data Science for Public Policy

Supervised by

Dr. Simon Munzert

April 2025

Word Count: 7,747

Table of Contents

Executive Summary

In an era of fragmented media consumption and declining trust in traditional news, digital trace data offers a novel lens to understand political behavior. This study investigates whether individuals' voting preferences in Germany's 2017 federal election can be predicted using demographic data and news browsing behavior (URL visit histories). By combining natural language processing (NLP) with machine learning, the research addresses gaps in prior work, such as reliance on aggregated or domain-level data, while exploring the ethical implications of personal data usage in political contexts.

The analysis of this research is based on two datasets from the same study. The demographic and political preferences are from 1,344 respondents in the Media Exposure and Opinion Formation (MEOF) survey (Munzert et al., 2022a). These respondents collectively visited more than 1.4 million news-related links, which is already a subset of all tracking data that was collected from the survey (Munzert et al., 2022b).

Major techniques used included text analysis using Term Frequency-Inverse Document Frequency (TF-IDF) to explore the importance of words and Latent Dirichlet Allocation (LDA) to explore the diverse topics. Machine learning tools such as gradient boosting, logistic regression, multi-layer perceptron (MLP), support vector machines, and decision trees took demographic and textual aggregation data to classify headlines and individuals. Accuracy and F1 scores for party choice (7 categories) and ideology (left/center/right) are used to evaluate the models.

For party choice, models had lower accuracy ranging in 20 to 30 percent, with linear SVC (feature-selected) peaking at 62 percent. Class imbalances and noise in URL data limited performance. For Ideological prediction, the accuracy improved to 54 percent, most being near 48 percent. Coherent themes like "sports," and "politics" were identifiable through LDA, but the topics alone were not enough for a reliable prediction (F1 = 18 percent for parties).

Individual-level prediction remains challenging, but the method shows promise for analyzing aggregate trends (e.g., shifts in media consumption and polarization). Integration of advanced NLP (e.g., BERTopic), spatial analysis, and cleaner URL processing could enhance accuracy. Despite the results, this research also highlights risks of misuse by actors with access to high-quality trace data (e.g., microtargeting, disinformation). To address this, strengthened privacy regulations (e.g., GDPR) and transparency in data usage are critical.

The code can be found at: https://github.com/RayH-1/Thesis-URL-MEOF-Voting-Ideology

1. Introduction

Freedom of journalism and the press is often considered a cornerstone of a healthy liberal democracy. Legacy media has long held considerable influence over public opinion, with repeated exposure to certain narratives shaping readers' mindsets. However, trust in traditional media has declined in recent years, and news consumption has become increasingly fragmented across digital platforms, podcasts, and social media. At the same time, numerous domestic and foreign actors have gained access to vast amounts of personal data, enabling them to influence political opinions subtly and effectively. While legislation such as General Data Protection Regulation and Digital Services Act aimed at protecting individuals exist, their enforcement remains limited, leaving people vulnerable to targeted disinformation campaigns and algorithmic manipulation, often without their awareness.

Prior research has demonstrated that individuals exhibit strong preferences for news sources that align with their existing beliefs. Since the COVID-19 pandemic, online news consumption has surged, further polarizing media diets. Political preferences have also shifted unexpectedly, even among demographics traditionally loyal to certain parties, as seen in the 2024 U.S. elections and Germany's 2025 *Bundestagswahl*. News organizations play a critical role in shaping narratives, often reinforcing ideological divides through selective framing. These biases, both at the individual and institutional levels, create detectable signals that may be analyzed to gauge political leanings. However, this dynamic also exacerbates polarization, making ideological divisions more pronounced and measurable.

This research seeks to classify voter preferences by analyzing demographic data and online browsing behavior, particularly news consumption patterns. Previous studies have successfully used digital trace data, such as Google Trends and URL visit histories, to predict election outcomes and public sentiment. Similarly, machine learning techniques, including natural language processing (NLP), have been employed to detect media bias and framing in news coverage. However, few studies have combined these approaches to assess how individual news exposure correlates with voting behavior. By bridging this gap, the study aims to provide deeper insights into the relationship between media diets and political preferences.

The study also addresses key limitations in existing research. While prior work has relied on aggregated social media data or domain-level browsing histories, this research focuses on individual-level analysis, offering greater precision. Additionally, it tackles the growing unreliability of traditional polling, which has struggled to capture shifting voter sentiments, as seen in recent elections. By leveraging advanced NLP and machine learning techniques, the study not only enhances predictive accuracy but also opens new avenues for analyzing emerging trends in political behavior. This approach could prove particularly valuable in understanding the impact of AI, algorithmic personalization, and disinformation on electoral outcomes.

This research contributes to a broader understanding of how digital media consumption influences democracy. By examining the intersection of news bias, personal data, and voter behavior, it highlights both the risks of manipulation and the potential for data-driven insights to improve political forecasting. The findings could inform policymakers, tech regulators, and journalists seeking to mitigate polarization and safeguard democratic processes in an increasingly digital and data-driven world.

2. Literature Review

This research is an intersection of various existing research into news media, media consumption, voter preferences, natural language processing, online browsing habits, and political polarization. First, we will discuss the public perception of companies' use of personal information, and the importance of data in the information age. Next, numerous studies will explain the theoretical basis for this research. Then, we will discuss the results of similar research in the past and what it means for the expectations for findings. Finally, the unique value of this study in the context of existing research will be discussed.

2.1 News, Algorithms, and Privacy: Digital Data and Democracy

News Consumptions and Public Opinions

Repeated exposure to certain news headlines can increase the perceived accuracy; in addition, people are inclined to believe headlines if they agree with the viewpoint (Calvillo & Smelter, 2020). Another research studied the dynamic between trust and expectancy and found comparable results. The study found that people who already had preconceived notions about illegal immigration believed headlines that agreed with them, and they believed it even more if it was from a source they do not trust (e.g. conservative seeing CNN saying 'more illegal immigrants') (Blom, 2021). This seems to imply that people prefer seeing news sources that align with their viewpoints, which can potentially allow us to determine someone's viewpoints based on their media diet.

In the 2024 U.S. Presidential elections, most Americans got their news from journalists and news organizations. That still leaves nearly 40 percent who get their news from other sources (Eddy, 2024). Podcasts for example have been on the rise since 2013 (Shearer et al., 2023). Meanwhile, there is no consensus on "the news influencer" (Lipka, 2025). The democratization of news sources and the rise of podcasts is blurring the lines between entertainment and journalism (García De Torres et al., 2025). Fragmentation allows individuals to go to the frames and narratives that best suit them. At the same time, there is no guarantee that all the sources will present news in a factual manner. This has a direct impact on trust in the media and current events.

Election Stakeholders and Risk of Interference

Large tech companies have major influence over people's information and viewpoints. A Pew Research study found over half of Americans get their news partly from social media such as Facebook, YouTube, and X (Twitter) ("Social Media and News Fact Sheet," 2024). In the 2024 U.S. Presidential elections, major tech companies including Meta, Google, SpaceX, and Amazon donated millions to the Trump campaign (*Big Tech Is Donating Millions to Trump's Inauguration*, 2025; *Top Contributors, Federal Election Data for Donald Trump, 2024 Cycle*, n.d.; Yang, 2025). Jeff Bezos, who purchased the Washington Post, exercised editorial control and stopped the endorsement of Kamala Harris while spending money on Trump's campaign (Mangan, 2024). Elon Musk, who purchased Twitter, has been accused of silencing journalists he disagrees with on his platform (Joyella, 2024). Their influence is not limited to the United States, Musk has also meddled in Germany's elections in 2025 (de Graaf, 2025). Big tech companies, and the people behind them, have shown that they are willing to use their assets for political means. They have the capability to influence the public's voting preferences by utilizing the enormous amounts of

data they collect and manipulating the algorithm to promote content, for example news articles, that aligns with the platform owners' views. This does not even take advancements in artificial intelligence (AI) into account, which may exacerbate the situation.

Foreign interference with a concerted effort from nation-states can influence the media narrative. Russia is known to use disinformation campaigns during election years and beyond in many countries. The U.S. Department of Justice stated that Russia has paid influencers to spread propaganda (Simon, 2024). Coupled with Pew Research Center findings where the share of people receiving news from "influencers" and "podcasts" increased, more people may be exposed to disinformation in the future. Blom and Calvillo et al. already established that people will believe in what they agree with, and if they believe in disinformation because it is the only "source" that agrees with them, then one group is functioning "in reality" and the group is not. This results in gridlock and democratic inefficiency. This does not even mention the cybersecurity risks of bad actors getting a hold of people's browsing data to determine how to disseminate disinformation.

Legitimate Data Use Cases and Legislations

AI systems can promote certain ideas by pushing news frames, but they can also help prevent the spread of misinformation. A research paper proposed a way to detect misinformation using AI while being robust, scalable, and fair by having humans in the loop (Demartini et al., 2020). They recommended a mix between AI tools, crowdsourced workers, and fact checking experts.

Companies use the data they collect for non-nefarious purposes, such as advertisements. Public perception regarding such data collection and privacy varies. Studies have shown that user comfort with ad personalization depends on how accurate it is (Dolin et al., 2018). Other studies have shown that people have mixed feelings; some people find ad personalization interesting, while others find it creepy (Ur et al., 2012).

While not all countries have safeguards for privacy and security concerns, the European Union has several laws including the General Data Protection Regulation (GDPR), Digital Services Act (DSA), Digital Markets Act (DMA), and the Artificial Intelligence (AI) Act. The GDPR can prevent the misuse and mis-storage of people's personal information (*What Is GDPR, the EU's New Data Protection Law?*, 2018). The DMA aims to prevent tech companies from monopolizing the market (*About the Digital Markets Act*, n.d.). The DSA highlights the responsibilities of tech companies and holds them accountable for the spread of misinformation and disinformation (*DSA*, n.d.; *The EU's Digital Services Act*, 2022). The AI Act created risk categories for AI systems, and their uses, and aims to prevent misuse of AI (*High-Level Summary of the AI Act | EU Artificial Intelligence Act*, 2024)

2.2 Theoretical Basis: Predicting Political Preferences with Demographics & News Diet

Interest in News Increased

A study into news media framing during the COVID-19 pandemic suggested that the lockdowns imposed by governments across the world were the reason for the increase in citizen consumption of traditional news media (Serrano-Contreras & Díaz-Montiel, 2023). Furthermore, their findings reiterated the idea that the media are not neutral transmitters of information, but they also offer new frames, ways of thinking, to their audiences (Hansen, 2015; Serrano-Contreras & Díaz-Montiel, 2023). Based on this and the fact that people seek out the media they agree with, it can be deduced that people's political preferences and who they vote for can be inferred from the types of news media they consume.

Shifting Demographics

Old voters tend to vote for conservative parties and young voters favor liberal parties. However, in recent years there has been a conservative shift in the younger generation. Donald Trump performed better than Kamala Harris for young males ages 18-29 in the 2024 U.S. Elections (*National Exit Polls*, 2024). This rightward shift is even more apparent in trends over time; both men and women had a considerable rightward shift in political views since 2018 (Cox, 2024). Even the polls were wrong about Kamala's chances of winning the popular vote (Warren, 2024). Similar trends can be seen in other democracies such as Germany's *Bundestagswahl* 2025, where the far-right AfD performed better than the moderate conservative CDU for 18-24 year-olds, and men overall favored the CDU and AfD compared to women (*Bundestagswahl 2025 - Ergebnisse und Analysedaten*, 2025). This highlights the unreliability of polling along with past conceptions, and it demonstrates a need for new methods of gauging voter preferences.

This difference in policy preferences may be explained by the values demographic groups can hold. A study into value-based predictions of election results utilized natural language processing (NLP) on open-ended survey questions and attempted to quantify social values. They found that "the value-based forecasts of Party Vote share investigated matched the final party standing with reasonable accuracy. Consequent to the substantial role that six of Lasswell and Kaplan's eight values had in predicting the final election outcome, it is logical that values be studied more closely in future election polling and analysis" (Parackal et al., 2018). The takeaway is that people's values matter, and machine learning may be used to detect them, so if voters seek out certain values in their news, then those news sources may be used to predict voter preferences.

News Headlines Steering Public Opinions

A study on newspapers on political attitudes towards the 2006 Virginia gubernatorial elections found that "even short exposure to a daily newspaper appears to influence voting behavior and may affect turnout behavior" (A. S. Gerber et al., 2009). More recently, a study in Austria and Germany attempted to determine how saliency and tone in news coverage influenced voter behavior. They found that a positive tone and saliency has an impact on preferences and expectations for coalitions of Austria, but only expectations in the case of Germany (Eberl & Plescia, 2018). From a natural language processing perspective, this provides an opportunity to detect and analyze the language of news articles that voters are exposed to, since there is a possibility that they can be influenced by them.

2.3 Prior Uses of NLP, Digital Trace Data, and User Behavior

Detecting Individuals' Preferences

Studies have used user's digital trace data via Google Trends to predict various topics. This includes demographic information such as unemployment (D'Amuri & Marcucci, 2017) and current events such as COVID-19 cases (Ortiz-Martínez et al., 2020). Some studies question the efficacy of using such methods, highlighting their doubts over the credibility of election forecasting and results at the individual level (Lui et al., 2011). In contrast, predicting election results in Germany using Google Trends data "show[s] that a strong correlation exists between the search preferences of potential voters before the date of the election race and the actual elections results" (Polykalas et al., 2013).

These findings are not limited to Google Trends data. A study into the 2016 U.S. Presidential election using data from 100 thousand individuals over a 56-day period shortly before the election showed that domain level URL visit history was able to predict election results with accuracy similar to polling (Comarela et al., 2018). The researchers' approach had the added benefit of detecting the shift of voter sentiments after the "Comey Letter". There is value in collecting and analyzing trace data as it can be informative about the real world; this also justifies the scope of monitoring online behavior shortly before elections as it makes it possible to predict party results.

Detecting News Media Biases and Preferences

News sources can contribute to political polarization with their content. Various types of machine learning techniques and algorithms have been used for analyzing political communication (García-Marín & Luengo, 2023). Unsupervised algorithms have been used to analyze large texts from two Spanish media outlets on education issues (Serrano-Contreras et al., 2021). Their results illustrate that news media can report on the same topic differently, and it is detectable using computation. Another study used support vector machines (SVM) to detect various frames (e.g. "security" or "human drama") employed by the Spanish press regarding the refugee crisis (García-Marín et al., 2018). Researchers also have used NLP tools to determine cosine similarities of news headlines to the word "feminism" during the COVID-19 Pandemic; the study aimed to look at temporal fluctuations of the similarities and hypothesized that the media relegated feminist issues in times of social crisis to the background (Serrano-Contreras & Díaz-Montiel, 2023). These studies demonstrate that text themes may also be determined using natural language processing.

Keeping up with modern times, researchers utilized large language models (LLMs) such as ChatGPT-4 to detect political bias from URL web domains (Hernandes & Corsi, 2024). They found that ChatGPT worked best in classifying domains that are clearly far-left or far-right. The researchers highlighted that ChatGPT was unable to classify about 66 percent of the domains, and the classifications were left-leaning. This research highlights the potential limitations of URL-level trace data and advanced classification techniques.

Using News to Predict Voter Preferences

There are no studies directly correlating individuals' preferences and news consumption for voter preference; however, some studies investigate similar topics. There have been studies using the same data as this thesis; they looked at URL-level trace data and various demographic factors to predict party vote and voter turnout (Bach et al., 2021). However, they did not utilize NLP, the respondents were likely different, and they only looked at URL web-domains, like the Chat-GPT study (Bach et al., 2021; Hernandes & Corsi, 2024). Their results were also similar in that "digital trace data do not allow us to accurately identify undecided voters, while we achieve slightly better results for (self-reported) voting and for votes for a right-wing populist party (AfD) and a progressive environmentalist party (Greens). Comparing different feature groups indicates that digital trace data seems to be more informative than sociodemographic information regarding predictions of populist party preferences (voted for AfD)" (Bach et al., 2021). This further corroborates the idea that the "signal" of individuals' preferences may be more detectable with clearly left- or right-wing parties. Other studies at domain-level trace data analysis to predict voter preferences observed that "visits to sites originated from social media are more important to infer candidate preference than those originated from other sources" (Comarela et al., 2018). However, their research was also aggregated at the temporal and spatial levels.

2.4 Strengths of This Research

Addressing Weaknesses of Social Media

Social media text analysis addressed many of the limitations of using Google Trends since it allowed for data to be collected at an individual level without aggregation. However, while social media can give a lot of information, tweets for example, it does have challenges. First, in recent years frequent use of bots has littered the internet with what is known as "AI slop" (Tang & Wikström, 2024). It has gotten to the point where there is research being written about "the Dead Internet Theory", which is the notion that interactions on social media will not be between humans in the future (Muzumdar et al., 2025; Walter, 2025). The underlying problem for analyzing voter preferences is that robots cannot vote yet, and the phenomenon degrades the quality of the data. Second, internet discourse is not limited to text; there are memes and texts can have different meanings. Natural language processing has not yet reached the levels where it can keep up with the pace of human language evolving. Finally, there is the issue of access and availability. Elon Musk's takeover of Twitter has resulted in both a mass exodus from the app and an increase in prices for their API (Geuens, 2025). The exodus introduces a bias in the dataset; if primarily conservatives remain then voter preferences will favor conservatives. In this research, the trace data is obtained at the individual level, and future experiments can allow this information to be collected at the respondents' consent (or purchased from data brokers).

Overcoming Challenges to Polling

The discourse of polls being unreliable comes up every time there is a Presidential election in the United States since 2016. A study has shown that despite these claims, data from the last 70+ years has shown no evidence that poll errors increase over time (Jennings & Wlezien, 2018). This information does not necessarily mean that confidence in polling has returned. As with surveys, respondents can obfuscate their true intentions and beliefs. Polls can also have a bias

from response (or non-response). User's digital trace data overcomes this limitation because it is not inherently political, so there is less of a need to obfuscate; although the respondents can choose to disable tracking, it has not been used in any meaningful way in this dataset. There is also the added benefit of having the data at a granular level, which may be used to determine shifts in media diet and therefore voter preference over time.

Filling Knowledge Gaps and Providing Opportunities

The benefit of digital trace data is that people seek out interests. In this way, the media they seek out should reflect their interests. Past research involving URLs looked at the domain-level, however, URL extensions can be informative because they offer more context. News URL extensions often list the article headlines, which provides an opportunity for natural language processing. Since people also get their news online, this means that digital trace data can capture a sizable portion of the news diet and therefore be a potential predictor of ideological preferences and therefore voting preferences. In the future, the list of valid domains can extend to other types of media that people can get their news from, such as podcasts and blogs.

Appropriate data processing can also potentially inform demographic information, which might forego the need for a separate survey for that information. Research has been conducted on digital trace data to predict an individual's demographic information and the results are somewhat promising (Hu et al., 2007; Pande et al., 2025). It is not unreasonable to suggest that future research will allow even higher accuracies and utilize the same dataset for different information that is fed into the same vote prediction models.

The 'unique selling point' of this research is that it combines demographic survey information with the respondents' digital trace data. This research builds on existing research by utilizing machine learning techniques in addition to natural language processing. Finally, this research aims to detect individuals' political preferences using their news article browsing habits and information about themselves.

3. Data and Description

3.1 Survey and Trace Data

The demographic information in this research comes from the Media Exposure and Opinion Formation (MEOF) multi-wave panel survey that was conducted between 13 July 2017 and 14 October 2019 in Germany (Munzert et al., 2022a). The data was collected by YouGov Germany in a total of nine waves. Each wave consisted of a nationally representative sample of adults in Germany. The 2,579 respondents were also asked questions related to political behavior and attitudes, elections, media use, and hate speech. For this research, the most important wave is the one that has the information regarding the respondent's voting choice in the German 2017 Federal Election (i.e., Wave 5). Other demographic information for the respondents in wave five was acquired from the questionnaire responses in waves 1-4. Information and any new respondents from waves 6 through 9 were not utilized in this research.

Respondents were given the option to opt into the Pulse panel, which is a subset of YouGov's traditional survey panels (Munzert et al., 2022b). The tracking software used ran in the background of the panelists' devices and collected anonymized visit data. The software tracked web traffic for all browsers installed on their computer, without any technical drawbacks, and it was transparent about the data logging. This information included the URLs, timestamps, duration, domains, as well as the respondent ID, which can be used to refer to their survey responses. The respondents also had the option to disable tracking for 15-minute intervals. The total number of respondents in Germany who opted into this was 1,281 in 2017 and 1,433 from 2018 to 2019.

For data processing purposes, only the domains that are classified as news sites from the "German Online News Domains" dataset were used. This is a collection of "1,147 primarily German-language domain names classified as being related to news and current affairs", and it was collected with the assistance of the POLTRACK project (Kulshrestha et al., 2023). This serves as a filter to narrow the scope of this research.

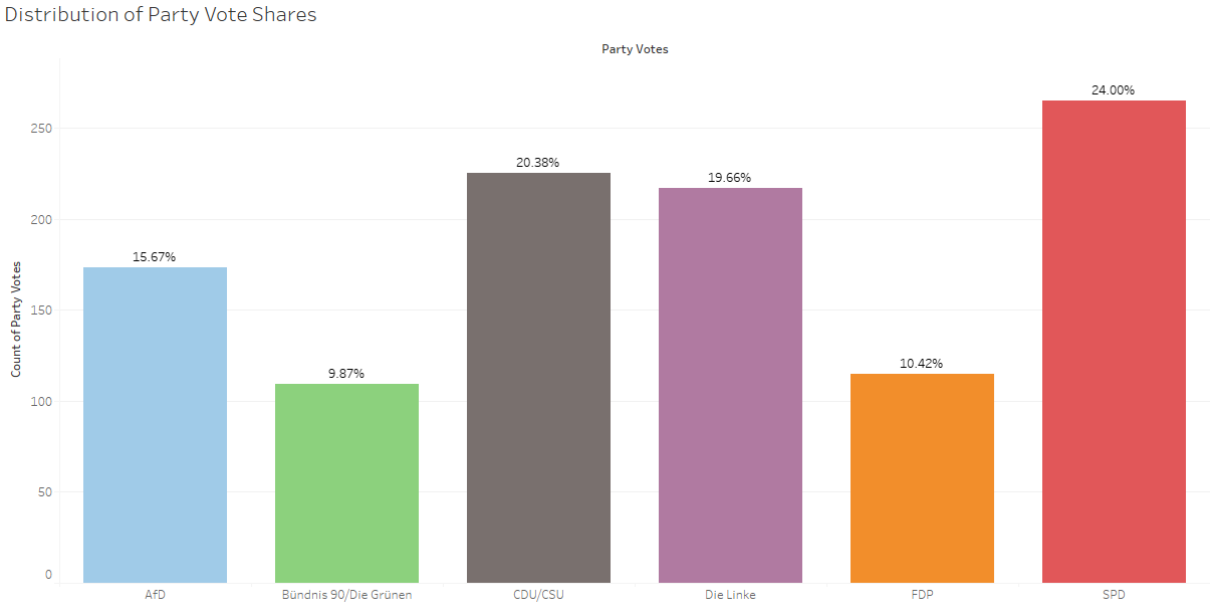
3.2 Descriptive Statistics

Survey Data

SAMPLE: The baseline number of respondents on panel 1 (July through December 2017) is 1,516. After filtering the dataset for wave 5 (i.e., voters), the survey dataset **(N = 1,344)** is composed of the following:

The first version of the outcome variable `secondvote`, refers to the outcome of the survey question: "You have two votes in the federal election. The first vote is for a candidate from your constituency and the second is for a party" (Munzert et al., 2022a). The responses included 'CDU/CSU,' 'SPD,' 'FDP,' 'Bündnis 90/Die Grünen' (Greens), 'Die Linke', 'AfD', 'other party', and 'don't know yet'. Figure 1 shows a class imbalance in the dataset where CDU/CSU, FDP, and Die Linke have a huge portion of the votes compared to that of the FDP and Greens. Interestingly, this doesn't reflect the true distribution of the percentage outcomes from the election; Die Linke and the SPD are overrepresented compared to the election results (*Bundestag Election 2017 Results*, 2017).

Figure 1: Counts of Votes in German Federal Election Excluding "Other" and "Don't Know Yet" Responses.



The second version of the outcome variable condenses the parties into ideologies (i.e., left, right, and center/unknown). This was done to account for potential issues with classification, especially with previous studies which were only able to classify clearly delineated parties (Bach et al., 2021; Hernandes & Corsi, 2024). Figure 2 shows that there is less of an imbalance, but it is important to note that it still does not reflect the true results from the election (i.e., left-wing parties and unknown party voting status are overrepresented). This may be due to a social desirability bias.

Figure 2 Counts of Votes Summarized by Ideology

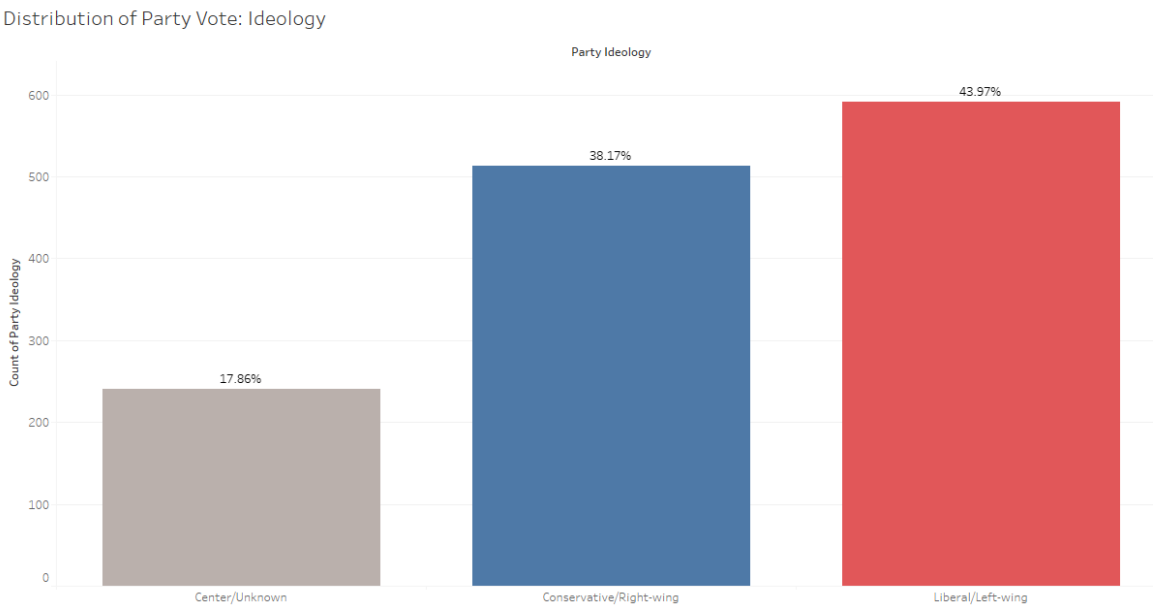


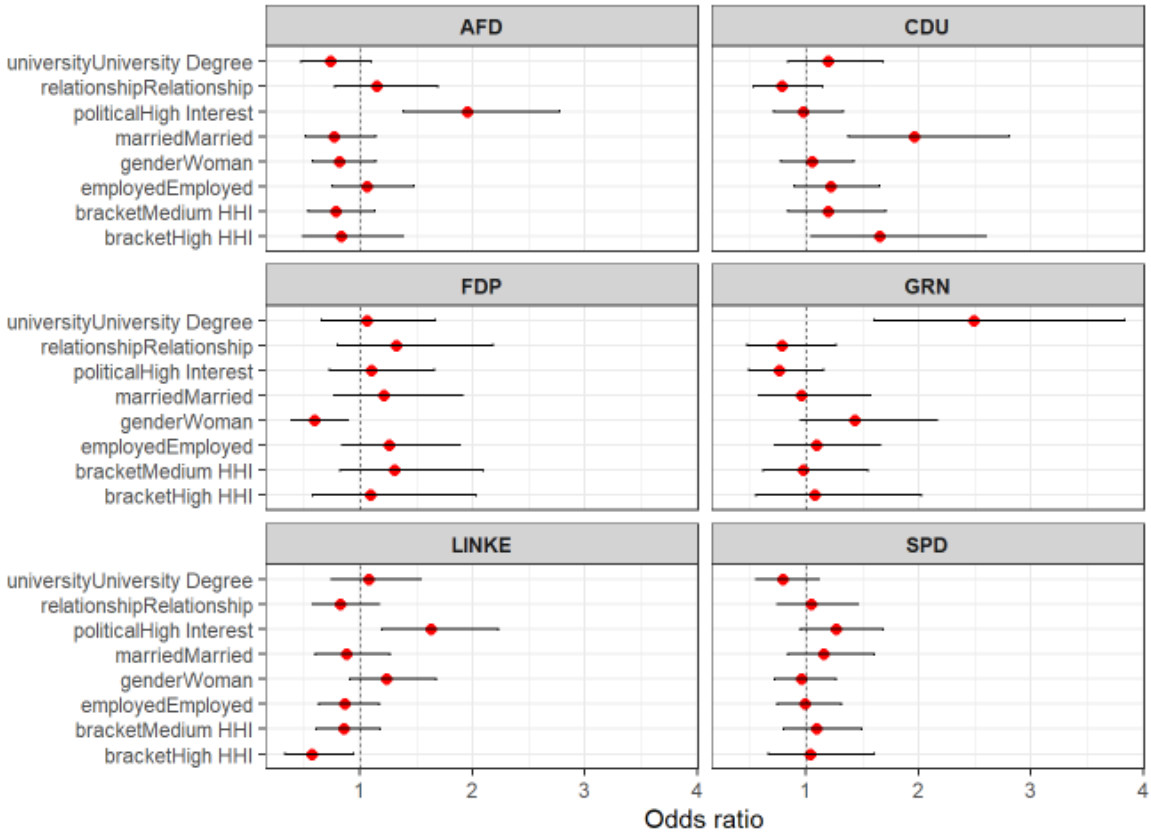
Table 1 shows summary statistics of variables by party ideology; a similar breakdown based on party vote can be found in the appendix [Supplement 5]. In terms of demographic differences, as expected, women are more liberal than men, but it is not much when looking at pure descriptive statistics. The percentage of men-women for liberals was 50-50 while for conservatives it was 57 percent favoring men. Most of the survey respondents do not hold a four-year university degree, and this was consistent throughout the party vote. This is contrary to expectation, since it is said that educated people tend to vote for liberal parties. Most respondents tended to be in a relationship, but not many of them were married. While the entire dataset was balanced about the degree of political interest, it is important to note that those categorized as "Center/Unknown" overwhelmingly responded that they have low interest in politics (74 percent). Meanwhile, a majority of those who responded to the `secondvote` question also indicated a high interest in politics (e.g., ~57 percent for conservatives, 54 percent for liberals).

Table 1: Outcome Variable Broken Down by Features

	Center/Unknown (N=240)	Conservative/Right-wing (N=513)	Liberal/Left-wing (N=591)	Overall (N=1344)
Gender				
Man	107 (44.6%)	294 (57.3%)	299 (50.6%)	700 (52.1%)
Woman	133 (55.4%)	219 (42.7%)	292 (49.4%)	644 (47.9%)
University Degree Status				
No University Degree	207 (86.3%)	395 (77.0%)	452 (76.5%)	1054 (78.4%)
University Degree	33 (13.8%)	118 (23.0%)	139 (23.5%)	290 (21.6%)
Employment Status				
Employed	107 (44.6%)	284 (55.4%)	293 (49.6%)	684 (50.9%)
Unemployed	133 (55.4%)	229 (44.6%)	298 (50.4%)	660 (49.1%)
Marital Status				
Married	67 (27.9%)	230 (44.8%)	220 (37.2%)	517 (38.5%)
Not Married	173 (72.1%)	283 (55.2%)	371 (62.8%)	827 (61.5%)
Relationship Status				
Relationship	138 (57.5%)	341 (66.5%)	353 (59.7%)	832 (61.9%)
Single	102 (42.5%)	172 (33.5%)	238 (40.3%)	512 (38.1%)
Political Interest				
High Interest	62 (25.8%)	292 (56.9%)	319 (54.0%)	673 (50.1%)
Low Interest	178 (74.2%)	221 (43.1%)	272 (46.0%)	671 (49.9%)
Monthly Income Bracket				
High HHI	29 (12.1%)	109 (21.2%)	92 (15.6%)	230 (17.1%)
Low HHI	105 (43.8%)	170 (33.1%)	230 (38.9%)	505 (37.6%)
Medium HHI	106 (44.2%)	234 (45.6%)	269 (45.5%)	609 (45.3%)

Running a generalized linear model and acquiring the odds of party vote against these variables yields comparable results. 'Die Linke' and 'AfD' have higher odds (1.63 and 1.95 times greater) of having high political interest than the reference category of low political interest. This makes sense because they are considered "far-left" or "far-right" which would necessitate a polarization of values that requires the interest to drive it outside the existing "center-leaning" parties. High income is significantly less common among 'Die Linke' voters, with the odds of high income being just 0.57 compared to low income. This may also be reflected in and influenced by the party's socialist platform (*Welcome*, n.d.). Among 'FDP' voters, the odds of being a woman is 0.59 compared to men. The odds of a Green-party voter being university educated is 2.49 times higher than being non-university educated. This finding reiterates the idea that liberals are more educated, but this pattern is not necessarily true for the 'SPD' and 'Die Linke' where their results are not statistically significant at $p < 0.05$. Respondents who voted for the 'CDU/CSU' have higher odds of being married (1.96) and being in the high-income bracket (1.65). This finding makes sense because traditional conservatives support less taxation because they have more money to protect, and the CDU's religious inclinations and average age might influence the odds of marriage being higher [Supplement 7].

Figure 3: Odd Ratios of Feature Variables by Party Vote



Trace Data

SAMPLE: There are approximately 1.4 million unique instances of news-related URLs. After data processing, the trace dataset **(N = 145,875)** retains the unique URL ID, domain, duration (in seconds), and the cleaned URL extensions containing the "article headline". There are approximately one million words with a total of 106,220 unique words.

After filtering out URL and German filler words, the word cloud in Figure 4 highlights the difficulty of using words alone to gauge political leanings. The biggest words are "campaign" and "Deutschland" which makes sense due to the elections. There are also words like "online" which could be a byproduct of the URL extension, or it could be in reference to news articles discussing something in the context of "online." Other exceptionally large words include "fussball" and "politik" which can refer to assorted topics of news.

Figure 4: Word Cloud of Article Headlines (URL Extensions)

Term Frequency-Inverse Document Frequency (TF-IDF) is a measure used in natural language processing to evaluate the "importance" of a word in a document relative to the corpus (*Understanding TF-IDF*, 2025). In this context, the "Term Frequency" (TF) refers to the number of times a given word appears in each headline divided by the total number of words in the headline. "Inverse Document Frequency" (IDF) is the natural log of the total number of headlines (in this case over one hundred thousand), divided by the number of times the word appears in all the headlines. The scikit-learn python package uses natural-log and normalizes the values based on the number of documents (Future Mojo, 2022).

$$\text{tf}(t,d) = f_{t,d}$$

$$\text{idf}(t,D) = \ln\left(\frac{N+1}{n_t+1}\right) + 1$$

$$w_{t,d} = \text{tf}(t,d) \times \text{idf}(t,D)$$

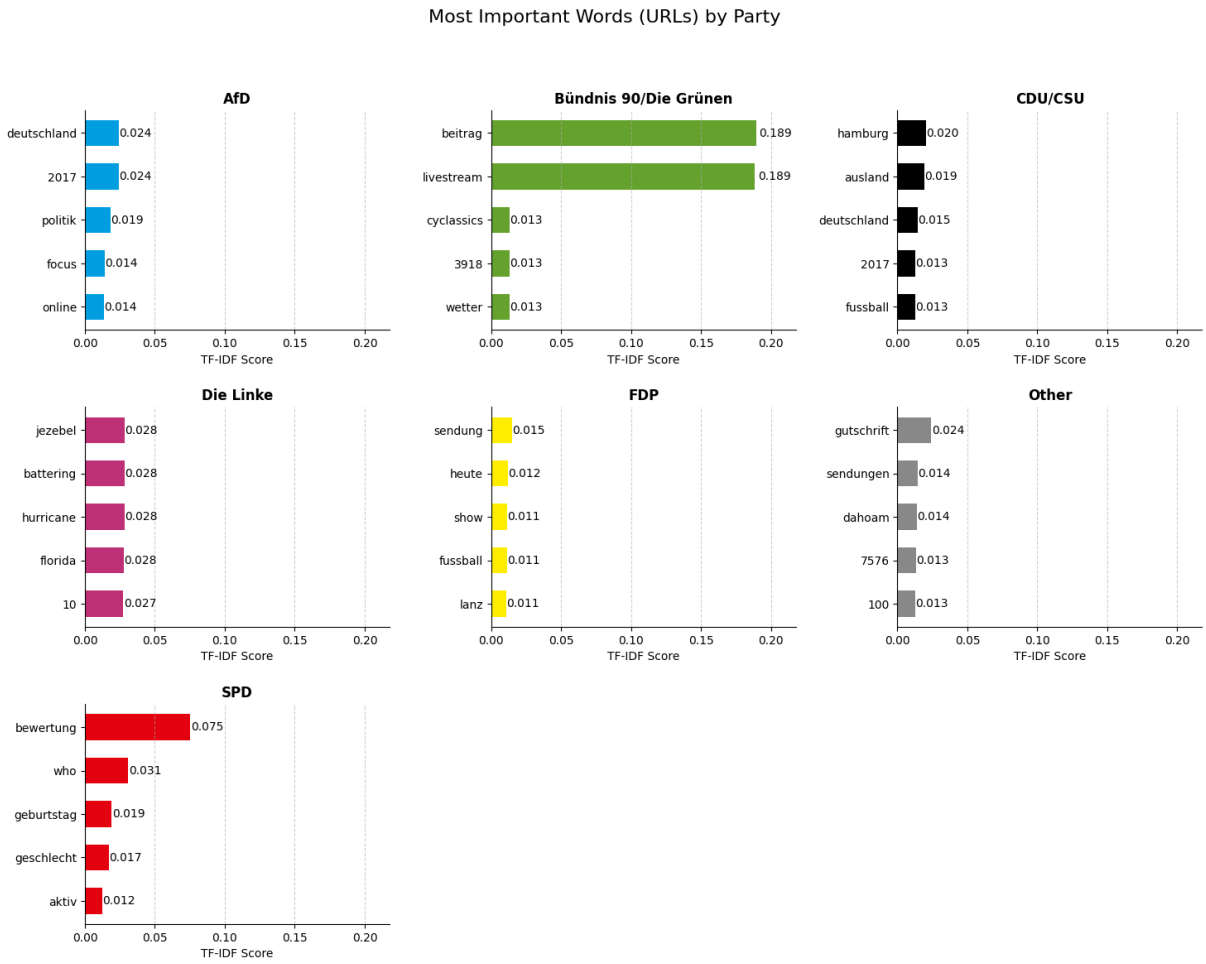
Where t = word, d = headline, D = all headlines, and w = TF-IDF score

Something to note is that TF-IDF scores for individual words will vary depending on whether the data is in long or wide format and when stratifying it by the outcome variable. Having the data in long format also introduces confounding effects, since there are duplicates for party votes. When looking at long format data (i.e., at the URL level where every row is a unique headline), there are some interesting findings:

- The Green Party has high TF-IDF scores for "livestream" and "beitrag" (contribution). This may be explained by URL artifacts and the fact that not many survey respondents voted for the Greens.

- The top five words for Die Linke are more-or-less consistent with most words having a score of 0.28.
- There is a remarkably high TF-IDF score for “bewertung” (evaluation) for SPD voters.

Figure 5: TF-IDF Scores of News Headlines at the URL Level by Party Vote

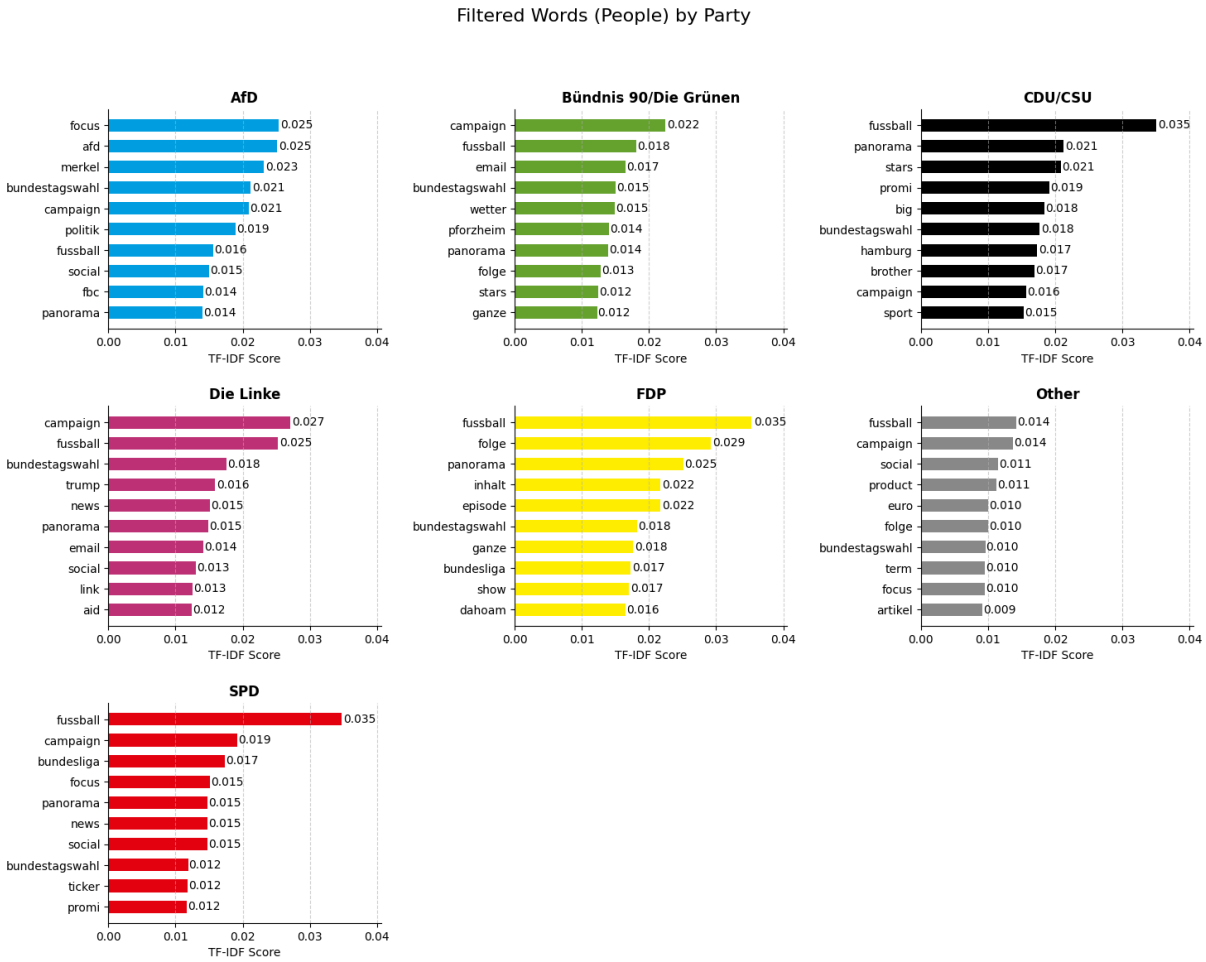


Since TF-IDF scores at the URL level were not particularly informative about individual voters’ interests, the data was converted to a wide format. This meant that every row was a unique individual, and all the headlines were appended to a single cell for each person. In this instance, the t in the TF-IDF formula remained as individual words, but d refers to all the headlines associated with each person, and D refers to all the headlines across all survey respondents. The limitation of using a wide format is that the effect of individual headlines is not noticeable, but the findings are much more interpretable.

The findings are significantly different from the long format and are also more informative:

- "Fussball" (football) is the top word for most of the party vote shares (i.e., CDU, SPD, FDP, and Other); CDU/CSU and SPD voters encounter the word far more than other words. This might allude to the fact that these parties, especially the SPD and CDU, are much more center-leaning than their counterparts.
- The more left-wing parties Die Linke and Greens have "campaign" as their highest-scored word, while the right-wing AfD has the word "focus" which might allude to the news site focus.de.
- "AfD" is high for AfD voters, which is unsurprising. However, AfD voters encountered the word "merkel," which refers to Angela Merkel.
- The term "Bundestagwahl" (i.e., German Federal elections) is a term in every voting group's top ten, which means they were keeping up with the elections.
- FDP voters have a lot more entertainment-related words like "episode," "show," or "dahoam" which is in reference to a German TV show.
- Green party voters have 'outdoorsy' words in their top ten like "wetter" (weather), "Pforzheim," and "cyclassics" (a bike race) from the URL level headlines.

Figure 6: TF-IDF Scores of News Headlines at the Person Level by Party Vote (Filtered)



4. Methodology

4.1 Target Variable

The objectives of this research are to:

[1] Determine if it is possible to classify users' browsing behavior and demographic information into specific political parties in Germany. The outcome would potentially allow prediction of the political party an individual voted for, based on their news browsing habits and their circumstances.

[2] If predicting the party vote is not possible, then this research aims to determine if ideologies (i.e., left-center-right) can be ascertained instead of party votes.

For the first objective, the party vote is determined by the answer to the following question:

> Bei der Bundestagswahl können Sie ja zwei Stimmen vergeben. Die Erststimme für einen Kandidaten aus Ihrem Wahlkreis und die Zweitstimme für eine Partei. Was [werden Sie auf Ihrem Stimmzettel ankreuzen][haben Sie auf Ihrem Stimmzettel angekreuzt]{if likelihoodvote dk==997}?
>
> 1. CDU/CSU
> 2. SPD
> 3. FDP
> 4. Bündnis 90/Die Grünen
> 5. Die Linke
> 6. AfD
> 7. Andere Partei, und zwar: [other] {open}
> 8. Weiß ich noch nicht

To simplify the list of parties, respondents who selected options 7 (Other) and 8 (Do not know yet) along with non-respondents were categorized as "Other".

To determine ideologies, respondents who indicated 'SPD,' 'Die Linke,' and 'Bündnis 90/Die Grünen' are categorized as "Liberal/Left-wing". Respondents who indicated 'CDU/CSU,' 'FDP,' and 'AfD' are categorized as "Conservative/Right-wing." The "Other" category containing 'Andere Partei,' 'Weiß ich noch nicht,' and non-responses are categorized as 'Center/Unknown.' Please see Appendix III for more details on data processing decisions.

4.2 Features

Survey Data

Demographic information was acquired from the MEOF survey:

- Age
- Gender
- Marital Status
  - Only those who selected "verheiratet" are categorized as "Married."
- Employment Status
  - Only those who selected "Erwerbstätigkeit/Berufstätigkeit" are categorized as being "Employed."
- Relationship Status
  - Respondents who selected "In einer Beziehung, mit Partner/in zusammenlebend" and "In einer Beziehung, aber nicht mit Partner/in zusammenlebend" are categorized as "In a Relationship".

- Political Interest
    - Respondents who indicated a 4 or a 5 in political interest (out of a 5-point scale) were categorized as having "High political interest".
- Degree Status (from Vocational Education)
    - Only those who selected "Universität- oder Fachhochschulabschluss" were categorized as having a University Degree.

The Household Income category used median imputation to fill non-responses. Due to re-coding for the other categorical variables into binary variables, any non-responses are automatically placed in the opposite group.

Trace Data

While there is other information in the trace data such as domains, URLs, and timestamps, the only two vectors of information that was utilized in the machine learning are the following:

- Duration: total number of seconds on a website.
- News proportion: total visits to news domains for a given respondent divided by all the sites they visited.

TF-IDF Means: Due to the machine learning models usually needing data at the same dimensions, we need to reduce the dimensions of the data. There are a couple of approaches attempted (for details on implementation, see the GitHub repository):

- Mean (i.e., Average)
- Entity-weighted identifies the named entities in the text (person, places, organizations, etc.) and doubles the weights.
- Syntactic weighted analyzes the grammatical structure of the sentences and assigns higher weights to subjects, objects, and verbs.

Topics: Latent Dirichlet Allocation, an unsupervised learning model for text, was applied to the corpus at the person level. Based on the results, the topics are assigned* as the following:

- Current Events
- Economy
- Elections
- Entertainment
- General
- Law and Order
- Politics
- Sports
- Unassigned

*Only the dominant topic is assigned to each person, along with its associated probability and topic words.

4.3 Timeframe

The demographic information comes from the most recent wave available before wave five, which is the last wave before the election (see GitHub for details). The digital trace data comes from the date respondents activated tracking (around July 2017) until the day before the election on September 24th, 2017.

4.4 Models

A variety of traditional machine learning models and a rudimentary neural network were used to classify the data. The goal is to evaluate a breadth of methods to gauge their performance and determine which models yield the greatest accuracy.

Traditional Approaches

All categorical variables were one-hot-encoded, especially the income category. The train-test split is 80-20, and it is stratified by the outcome variable(s). Since all models rely on supervised learning, the hyperparameters are cross-validated (k = 10) on the test set. The code for all the models was completed on Jupyter Notebook using Python.

Demographic data and TF-IDF means for each person were input in the following models:

- Gradient Boosting
- Random Forest
- Logistic Regression
- Decision Tree
- Support Vector Machine: Linear SVC was assessed on the text data itself to find any other patterns of word distribution on both outcome variables.

Non-Traditional Approaches

Latent Dirichlet Allocation (LDA): Words are tokenized with the Spacy German model and only the nouns, verbs, and adjectives are selected. Next, a bag-of-words (BOW) approach is used to create the LDA model. Ten-topics was selected as the parameter because it gave the most "coherent" word groups; other topic amounts such as 12, 15, and 20 were attempted, but the results were less coherent. Since this is an unsupervised classification model, it is up to our judgment on what the potential categories may be. More information on topic words can be found in the appendix [Supplement 10].

Naïve Bayes: The topics at the URL level were split 80-20 into training and test sets and predicted the outcome variables.

Neural Network: Demographic information and the topics at the individual level are input into the feed-forward neural network (multi-layer perception/MLP). MLPs are often called universal approximators because one hidden layer is enough to approximate any smooth function. These tools are powerful because they function similarly to a human brain, which allows them to find patterns that may be more difficult for simpler models.

This research uses the Keras package in Python to build the neural network. The activation function of choice is the ReLU. The output activation function is soft-max. There is early stopping if the function runs 10 epochs without improvement. Adam is used as an optimizer.

4.5 Evaluation

METRICS: There are class imbalances among some of the feature variables and outcome variables. As a result, the models are evaluated based on their accuracy and F1 scores.

5. Results

5.1 LDA Topic Verification

To check that the topics are somewhat accurate to the text they are from, they were trained with a naïve Bayes classifier against the TF-IDF feature vectorizer. The F1 score was 0.81 percent at the URL level. Then, multinomial naïve Bayes was used at the URL level to predict whether a certain URL (headline) belongs to a certain party. The URLs were down-sampled to have the exact number per outcome variable. The F1 score was low (around 18 percent). The exact same thing was tried with ideology, and the F1 score was around 38 percent.

While there is some justification for using Topic LDA to determine if the topics of the text are accurate, the topics alone are not enough to predict the party choice or ideology of a certain headline.

Figure 7: Topics Accuracy (TF-IDF) Comparison

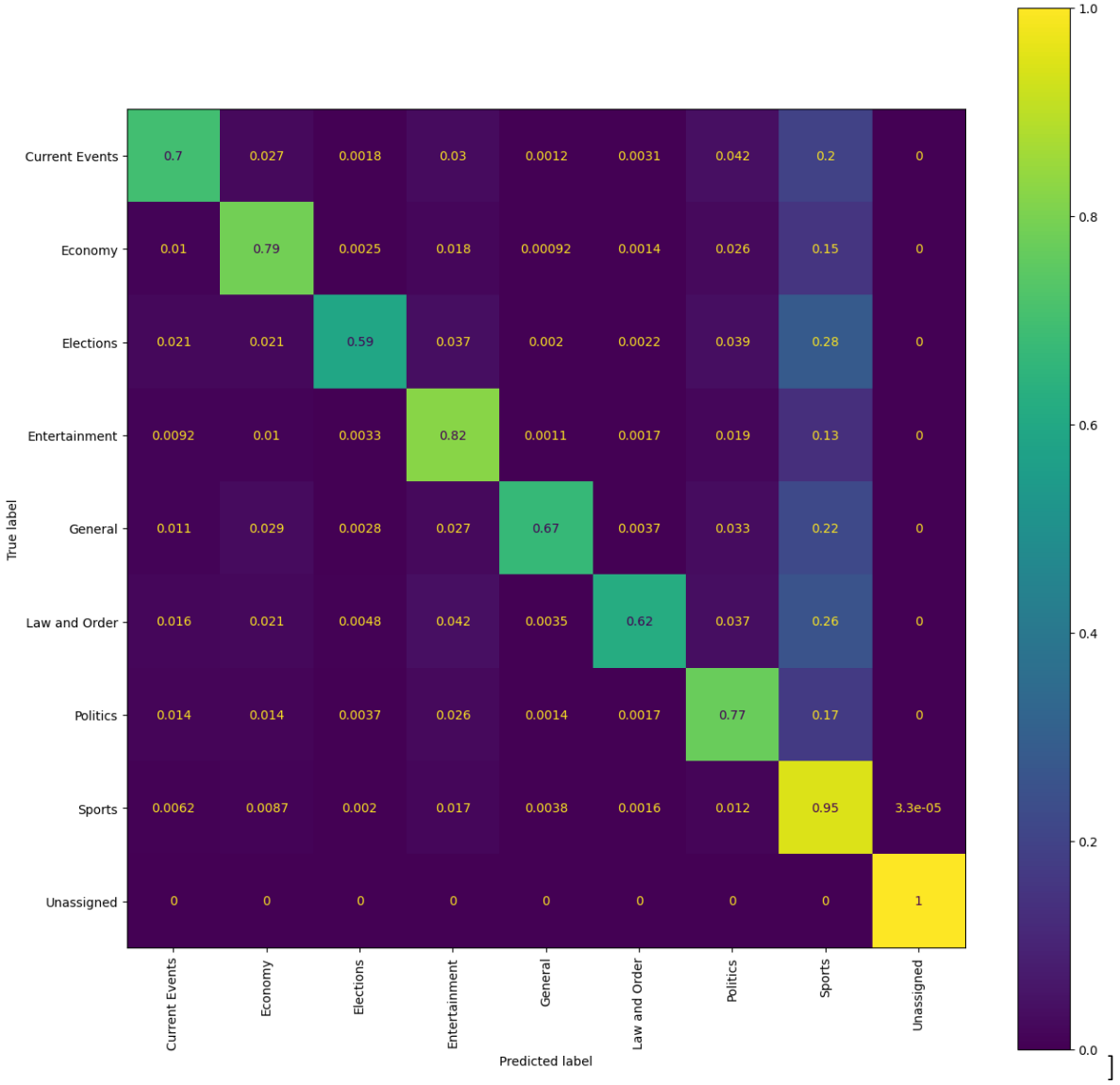
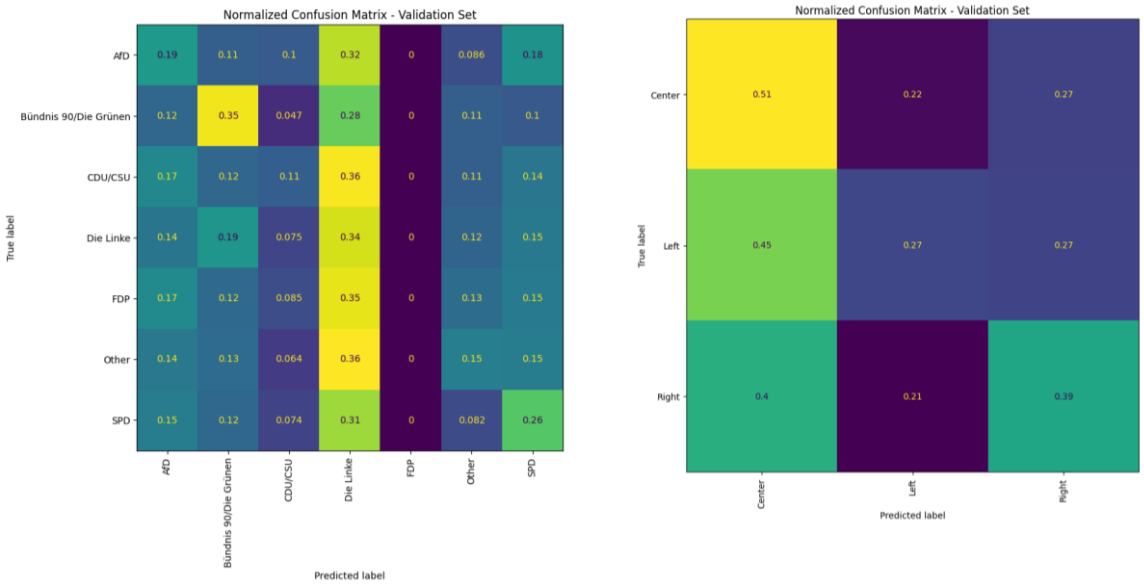


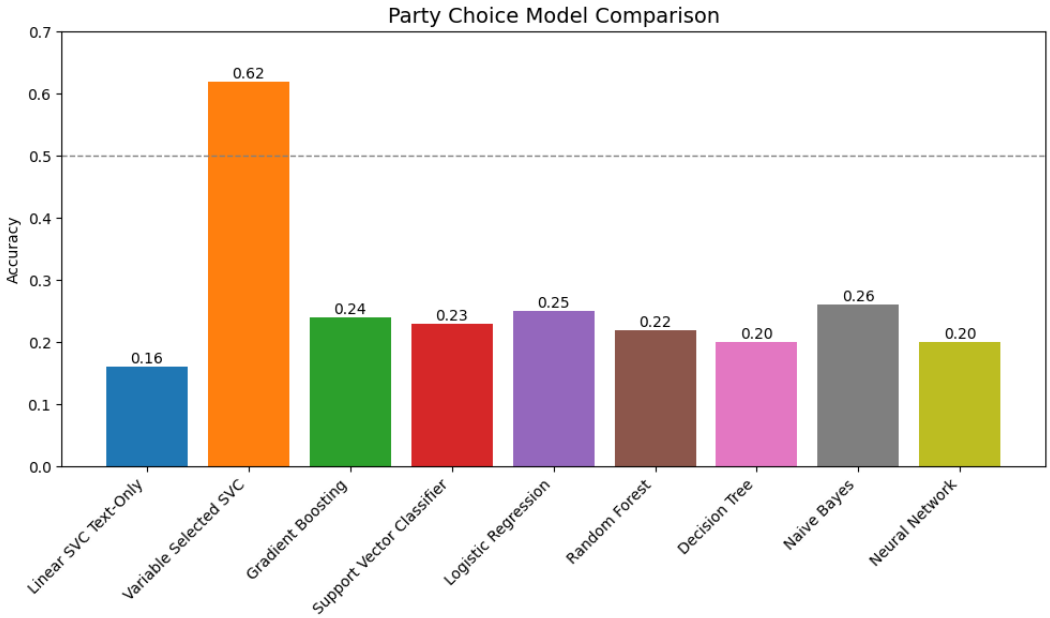
Figure 8: Multinomial Naive Bayes and Topic Headlines at URL level



5.2 Party Choice

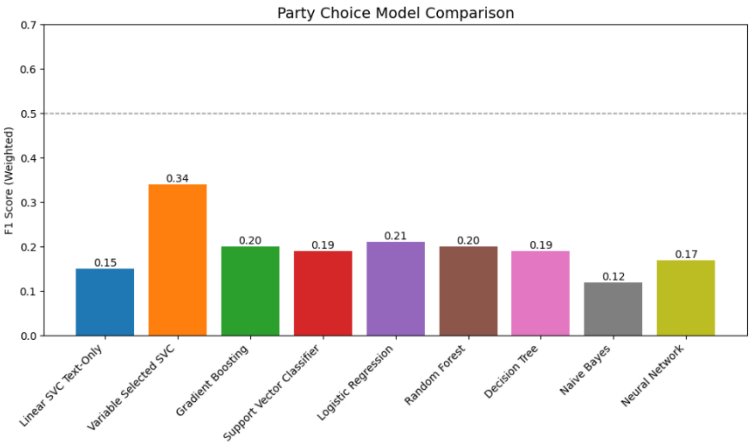
The answer to whether it is possible to predict a respondent's political party preferences based on their demographic and digital trace information is not promising. In terms of accuracy, most of the models performed in the 20 to 30 percent range. The best-performing model was Linear SVC, which selected specific features and output a result. The combinations of features that resulted in that accuracy are income, degree status, marital status, political interest, relationship, news proportion, and the TF-IDF entity weighted means. However, the result was still only 62% accuracy.

Figure 9: Party Choice Accuracy Comparison



Because of class imbalances, accuracy is not a particularly good metric to rely on. Linear SVC, with variable selection, also far outperforms the other models by a margin of ten percentage points. The other models' F1 scores are similar to their regular accuracy. Feature selection was not done on the other models, and it is reasonable to surmise that doing so would increase their accuracy as well.

Figure 10: Party Choice F1 Score Comparison



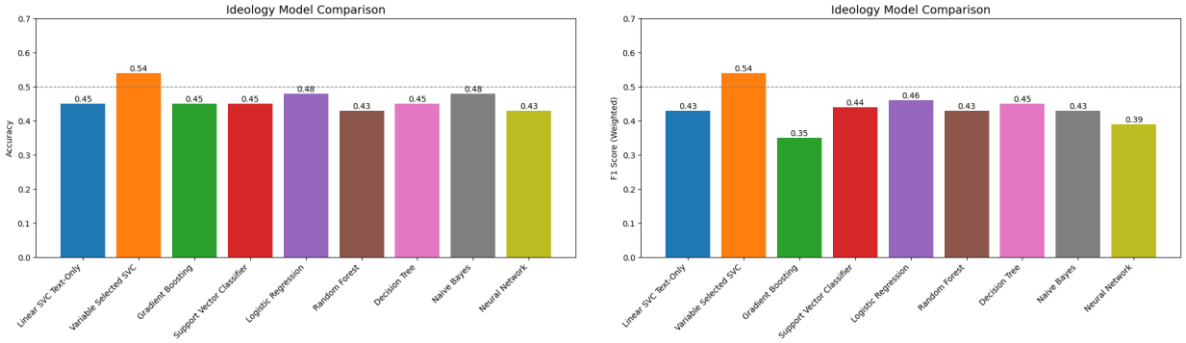
5.3 Ideology

While the results for predicting individual parties did not have promising results, the chance of predicting the ideology correctly is more optimistic. The top-performing model is the variable selected Linear SVC, which performed at an accuracy and F1 score of 54 percent. The accuracy is barely over the 50 percent threshold, and it is significantly worse than the party choice performance of 52 percent. However, there are a couple of key facts to consider.

First, "ideology" here is essentially three different classes. This means that removing the noise of the "center" class may yield greater results. Remember, the "center" class is essentially the non-respondents (i.e. those who responded they voted but did not indicate party) and those who indicated "other" and "don't know". This would likely make the demographic factors that align with the "true" ideology or party vote more apparent.

Second, these results are pretty low for three classes, but they may be higher for binary classification. This is based on the fact that going from a 7-class classification above to a 3-class classification yielded nearly 30 percentage points of accuracy and F1 scores for most models. Furthermore, this suggests that the results somewhat reflect what results may look like in a three-party system and can be more accurate in a two-party system.

Figure 11: Ideology Accuracy & F1 Score Comparison



6. Conclusion, Limitations, and Implications

The objective of this research was to determine whether it is possible to predict individual voters' party preferences based on their personal information such as demographics and their digital trace data. In the case that party preferences were not conclusive, then the goal is to determine the ideological lean. Several computational tools were employed, ranging from traditional machine learning models to natural language processing and neural networks. This research builds upon existing research in political classification based on digital trace data (both of people and news sources).

Feature-selected linear SVC performed the best both for party and ideology classification. Most models performed at similar levels in the same accuracy category, due to them using the exact same features to maintain consistency. The only models that did not use the same features were the neural network, naïve Bayes, and feature-selected SVC. Overall, the results for predicting party votes are less promising than for predicting ideology.

Data quality of digital trace data presented some of the biggest challenges. The news sources are not always in German, which reduces the quality of the trace data. Working with URL data creates sparsity; there are a lot of duplicates and empty values depending on browsing habits. Data cleaning is especially difficult because every website uses different filler extensions (e.g., "pid" and "rid" are both identification extensions from two different sites). Sometimes the news sites do not have news articles in the headlines; a potential explanation is advertisements. Other limitations included the choices of models and unavailability of resources; for example, BERTopic is an advanced LLM for topic modeling however the enormous number of words made this difficult to run on our devices. Using ChatGPT to predict the URL similar to the implementation by (Kotzé & Senekal, 2025) was considered, but it would be very expensive.

As alluded to previously, there are several directions for further research and improvement. First, using BERTopic to get higher-quality topic labels may result in classification improvement for both the Naïve Bayes method and the MLP implementation. A more natural language processing approach with named-entity recognition such as introducing weights to (or annotating) political keywords such as people, terms, etc. can address the noise limitations of digital trace data. This research did not explore any spatial aspects, nor did it investigate what the predictions would yield in terms of vote shares; future research can attempt to determine the vote shares per region. At the very least, this research could be replicated with more streamlined data processing and removing ambiguity (e.g., "center/unknown" class).

There are some major implications from the findings of this research. First, while the results were not particularly good, this is the worst it will ever be. Technology advances at a rapid pace, and more advanced methods will continue to increase accuracy. Second, this information is not something that only researchers have access to. Major corporations collect incredibly substantial amounts of data from "tracking cookies." They have access to data that is significantly higher quality and resources to create far more accurate models. If the tools are transparent and democratized, this information benefits policy stakeholders in track public sentiment. If it is not, then bad actors would be able to covertly sway elections or have the upper edge in manipulating public sentiment. Policymakers need to stay ahead of the curve, not only in terms of legislation but also in promoting research in the field of artificial intelligence and data sciences in the policy context.

Bibliography

About the Digital Markets Act. (n.d.). Retrieved April 19, 2025, from https://digital-markets-act.ec.europa.eu/about-dma_en

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2021). Predicting Voting Behavior Using Digital Trace Data. *Social Science Computer Review*, *39*(5), 862–883. https://doi.org/10.1177/0894439319882896

Big Tech is Donating Millions to Trump's Inauguration. (2025, January 10). Common Cause. https://www.commoncause.org/articles/big-tech-is-donating-millions-to-trumps-inauguration/

Blom, R. (2021). Believing false political headlines and discrediting truthful political headlines: The interaction between news source trust and news content expectancy. *Journalism*, *22*(3), 821–837. https://doi.org/10.1177/1464884918765316

Bundestag election 2017 results. (2017). The Federal Returning Officer. https://www.bundeswahlleiterin.de/en/bundestagswahlen/2017/ergebnisse/bund-99.html

Bundestagswahl 2025—Ergebnisse und Analysedaten. (2025). tagesschau.de. https://www.tagesschau.de/wahl/archiv/2025-02-23-BT-DE/

Calvillo, D. P., & Smelter, T. J. (2020). An initial accuracy focus reduces the effect of prior exposure on perceived accuracy of news headlines. *Cognitive Research: Principles and Implications*, *5*(1), 55. https://doi.org/10.1186/s41235-020-00257-y

Comarela, G., Durairajan, R., Barford, P., Christenson, D., & Crovella, M. (2018). Assessing Candidate Preference through Web Browsing History. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 158–167. https://doi.org/10.1145/3219819.3219884

Cox, D. (2024, November 7). *2024 Election Edition: Young Men Swing Toward Trump - The Survey Center on American Life*. https://www.americansurveycenter.org/newsletter/2024-election-edition-young-men-swing-toward-trump/, https://www.americansurveycenter.org/newsletter/2024-election-edition-young-men-swing-toward-trump/

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*(4), 801–816. https://doi.org/10.1016/j.ijforecast.2017.03.004

de Graaf, A. (2025, February 21). *Fact check: How Elon Musk meddled in Germany's elections*. Dw.Com. https://www.dw.com/en/how-elon-musk-meddled-in-germanys-elections/a-71676473

Demartini, G., Mizzaro, S., & Spina, D. (2020). *Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities*.

Dolin, C., Weinshel, B., Shan, S., Hahn, C. M., Choi, E., Mazurek, M. L., & Ur, B. (2018). Unpacking Perceptions of Data-Driven Inferences Underlying Online Targeting and Personalization. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3173574.3174067

DSA: Code of Practice on Disinformation. (n.d.). [Text]. European Commission - European Commission. Retrieved April 19, 2025, from https://ec.europa.eu/commission/presscorner/detail/en/ip_25_505

Eberl, J.-M., & Plescia, C. (2018). Coalitions in the news: How saliency and tone in news coverage influence voters' preferences and expectations about coalitions. *Electoral Studies*, *55*, 30–39. https://doi.org/10.1016/j.electstud.2018.07.004

Eddy, K. (2024, May 28). More than half of Americans are following election news closely, and many are already worn out. *Pew Research Center*. https://www.pewresearch.org/short-reads/2024/05/28/more-than-half-of-americans-are-following-election-news-closely-and-many-are-already-worn-out/

Future Mojo (Director). (2022, April 27). *NLP Demystified 6: TF-IDF and Simple Document Search* [Video recording]. https://www.youtube.com/watch?v=fIYSi41f1yg

García De Torres, E., Legorburu, J. M., Parra-Valcarce, D., Edo, C., & Escobar-Artola, L. (2025). Intimacy in Podcast Journalism: Ethical Challenges and Opportunities in Daily News Podcasts and Documentaries. *Media and Communication*, *13*, 8994. https://doi.org/10.17645/mac.8994

García-Marín, J., Calatrava, A., & G. Luengo, Ó. (2018). Debates electorales y conflicto. Un análisis con máquinas de soporte virtual (SVM) de la cobertura mediática de los debates en España desde 2008. *El Profesional de La Información*, *27*(3), 624. https://doi.org/10.3145/epi.2018.may.15

García-Marín, J., & Luengo, Ó. G. (2023). New Methodological Perspectives in Political Communication Research: Machine Learning and Algorithms. In M. Musiał-Karg & Ó. G. Luengo (Eds.), *Digital Communication and Populism in Times of Covid-19* (pp. 13–28). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33716-1_2

Gerber, A. S., Karlan, D., & Bergan, D. (2009). Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions. *American Economic Journal: Applied Economics*, *1*(2), 35–52. https://doi.org/10.1257/app.1.2.35

Geuens, R. (2025, February 19). *How many monthly active users does X (Twitter) have?* SOAX. https://soax.com/research/twitter-active-users

Hansen, F. S. (2015). Framing yourself into a corner: Russia, Crimea, and the minimal action space. *European Security*, *24*(1), 141–158. https://doi.org/10.1080/09662839.2014.993974

Hernandes, R., & Corsi, G. (2024). *LLMs left, right, and center: Assessing GPT's capabilities to label political bias from web domains* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2407.14344

High-level summary of the AI Act | EU Artificial Intelligence Act. (2024, February 27). https://artificialintelligenceact.eu/high-level-summary/

Hu, J., Zeng, H.-J., Li, H., Niu, C., & Chen, Z. (2007). Demographic prediction based on user's browsing behavior. *Proceedings of the 16th International Conference on World Wide Web*, 151–160. https://doi.org/10.1145/1242572.1242594

Jennings, W., & Wlezien, C. (2018). Election polling errors across time and space. *Nature Human Behaviour*, *2*(4), 276–283. https://doi.org/10.1038/s41562-018-0315-6

Joyella, M. (2024, January 9). *Elon Musk Accused Of 'Silencing His Critics' As X Suspends Journalists*. Forbes. https://www.forbes.com/sites/markjoyella/2024/01/09/elon-musk-silencing-his-critics-as-journalists-are-suspended-by-x/

Kotzé, E., & Senekal, B. A. (2025). Benchmarking Political Bias Classification with In-Context Learning: Insights from GPT-3.5, GPT-4o, LLaMA-3, and Gemma-2. In A. Gerber, J. Maritz, & A. W. Pillay (Eds.), *Artificial Intelligence Research* (Vol. 2326, pp. 161–175). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-78255-8_10

Kulshrestha, J., Stier, S., Puschmann, C., Merten, L., Rauxloh, H., & Schultz, C. (2023). *German Online News Domains* [Dataset]. OSF. https://doi.org/10.17605/OSF.IO/S5UHB

Lipka, E. T., Luxuan Wang, Jacob Liedke, Galen Stocking and Michael. (2025, February 19). No consensus on who comes to mind when Americans are asked to name a news influencer. *Pew Research Center*. https://www.pewresearch.org/short-reads/2025/02/19/no-consensus-on-who-comes-to-mind-when-americans-are-asked-to-name-a-news-influencer/

Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). On the predictability of the U.S. elections through search volume activity. *Department or Program: Computer Science Wesley College*.

Mangan, D. (2024, October 25). *Jeff Bezos killed Washington Post endorsement of Kamala Harris, paper reports*. CNBC. https://www.cnbc.com/2024/10/25/jeff-bezos-killed-washington-post-endorsement-of-kamala-harris-.html

Munzert, S., Barberá, P., Guess, A. M., & Yang, J. (2022a). *Media Exposure and Opinion Formation in an Age of Information Overload (MEOF) – Survey GermanyMedia Exposure and Opinion Formation in an Age of Information Overload (MEOF) – Survey Germany* (Version 1.0.0) [Dataset]. GESIS. https://doi.org/10.4232/1.13979

Munzert, S., Barberá, P., Guess, A. M., & Yang, J. (2022b). *Media Exposure and Opinion Formation in an Age of Information Overload (MEOF) – Webtracking on-siteMedia Exposure and Opinion Formation in an Age of Information Overload (MEOF) – Webtracking on-site* (Version 1.0.0) [Dataset]. GESIS. https://doi.org/10.4232/1.13981

Muzumdar, P., Cheemalapati, S., RamiReddy, S. R., Singh, K., Kurian, G., & Muley, A. (2025). *The Dead Internet Theory: A Survey on Artificial Interactions and the Future of Social Media*. https://doi.org/10.48550/ARXIV.2502.00007

National Exit Polls: Election 2024 Results. (2024, November 5). https://www.nbcnews.com/politics/2024-elections/exit-polls

Ortiz-Martínez, Y., Garcia-Robledo, J. E., Vásquez-Castañeda, D. L., Bonilla-Aldana, D. K., & Rodriguez-Morales, A. J. (2020). Can Google® trends predict COVID-19 incidence and help preparedness? The situation in Colombia. *Travel Medicine and Infectious Disease*, *37*, 101703. https://doi.org/10.1016/j.tmaid.2020.101703

Pande, P., Kulkarni, A. K., P, B., S, B., Ramalingam, V., & R, R. (2025). Big Data Analytics in E-commerce Driving Business Decisions Through Customer Behavior Insights. *ITM Web of Conferences*, *76*, 05001. https://doi.org/10.1051/itmconf/20257605001

Parackal, M., Mather, D., & Holdsworth, D. (2018). Value-based prediction of election results using natural language processing: A case of the New Zealand General Election. *International Journal of Market Research*, *60*(2), 156–168. https://doi.org/10.1177/1470785318762234

Peterson, D., Rooduijn, M., Hopp, F. R., Schumacher, G., & Bakker, B. N. (2025). Loneliness is positively associated with populist radical right support. *Social Science & Medicine*, *366*, 117676. https://doi.org/10.1016/j.socscimed.2025.117676

Polykalas, S. E., Prezerakos, G. N., & Konidaris, A. (2013). An algorithm based on Google Trends' data for future prediction. Case study: German elections. *IEEE International Symposium on Signal Processing and Information Technology*, 000069–000073. https://doi.org/10.1109/ISSPIT.2013.6781856

Serrano-Contreras, I.-J., & Díaz-Montiel, A. (2023). Feminist Framing in Times of Pandemic: An Analysis of the Spanish Case. In M. Musiał-Karg & Ó. G. Luengo (Eds.), *Digital Communication and Populism in Times of Covid-19* (pp. 29–40). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-33716-1_3

Serrano-Contreras, I.-J., García-Marín, J., & Luengo, Ó. G. (2021). Coberturas mediáticas, polarización y reformas educativas en España. *Revista de Ciencia Política (Santiago)*, *ahead*. https://doi.org/10.4067/S0718-090X2021005000109

Shearer, E., Liedke, J., Matsa, K. E., Lipka, M., & Jurkowitz, M. (2023). *Podcasts as a Source of News and Information*.

Simon, S. (2024, September 7). DOJ says Russia paid right-wing influencers to spread Russian propaganda. *NPR*. https://www.npr.org/2024/09/07/nx-s1-5101895/doj-says-russia-paid-right-wing-influencers-to-spread-russian-propaganda

Social Media and News Fact Sheet. (2024, September 17). *Pew Research Center*. https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/

Tang, J., & Wikström, P. (2024, September 19). *'Side job, self-employed, high-paid': Behind the AI slop flooding TikTok and Facebook*. The Conversation. http://theconversation.com/side-job-self-employed-high-paid-behind-the-ai-slop-flooding-tiktok-and-facebook-237638

The EU's Digital Services Act. (2022, October 27). https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

Top Contributors, federal election data for Donald Trump, 2024 cycle. (n.d.). OpenSecrets. Retrieved April 19, 2025, from https://www.opensecrets.org/2024-presidential-race/donald-trump/contributors?id=N00023864

Understanding TF-IDF. (2025, February 7). GeeksforGeeks. https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

Ur, B., Leon, P. G., Cranor, L. F., Shay, R., & Wang, Y. (2012). Smart, useful, scary, creepy: Perceptions of online behavioral advertising. *Proceedings of the Eighth Symposium on Usable Privacy and Security*, 1–15. https://doi.org/10.1145/2335356.2335362

Walter, Y. (2025). Artificial influencers and the dead internet theory. *AI & SOCIETY*, *40*(1), 239–240. https://doi.org/10.1007/s00146-023-01857-0

Warren, J. (2024, November 13). *Were the 2024 election polls wrong?* UC Riverside News. https://news.ucr.edu/articles/2024/11/13/were-2024-election-polls-wrong-ucr-expert-weighs

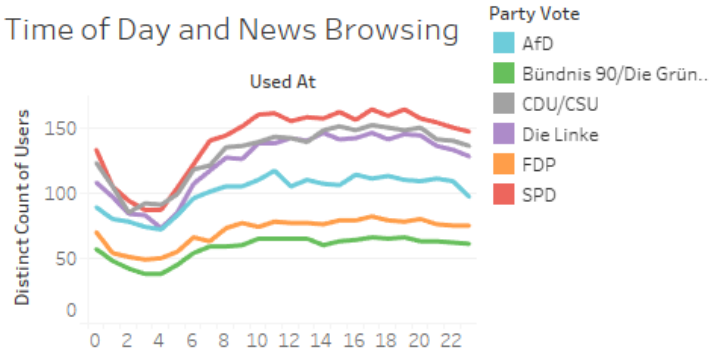
Welcome: Die Linke English pages. (n.d.). DIE LINKE. English Pages. Retrieved April 21, 2025, from https://en.die-linke.de/welcome/

What is GDPR, the EU's new data protection law? (2018, November 7). GDPR.Eu. https://gdpr.eu/what-is-gdpr/

Yang, M. (2025, January 9). Google and Microsoft donate $1m each to Trump's inaugural fund. *The Guardian*. https://www.theguardian.com/technology/2025/jan/09/google-microsoft-donate-trump-inaugural-fund

Appendix

I. Supplementary Materials



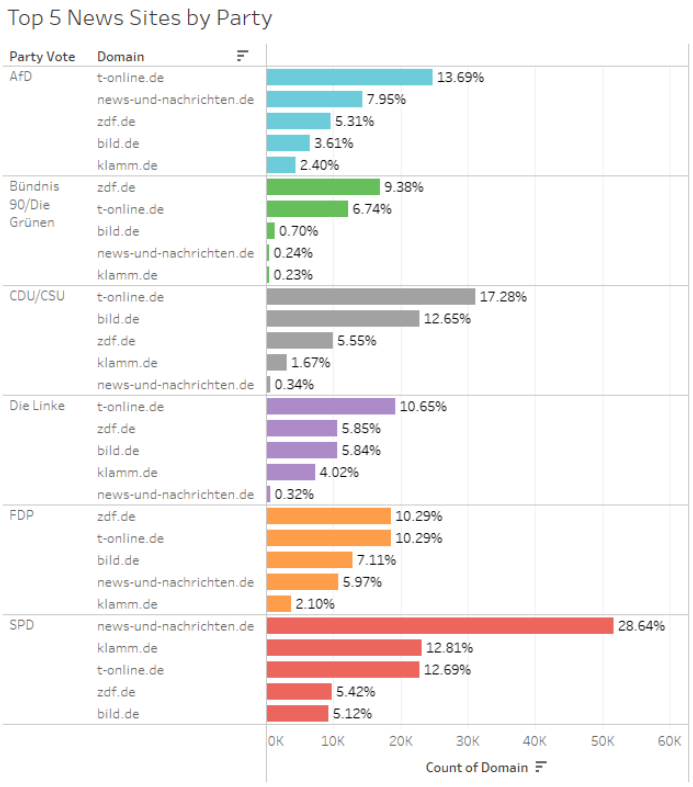
Avergate Time of Day

Party Vote	Avg Time of..
AfD	03:06:46 PM
Bündnis 90/Die Grünen	03:28:56 PM
CDU/CSU	02:59:04 PM
Die Linke	03:08:37 PM
FDP	03:29:36 PM
SPD	02:43:19 PM

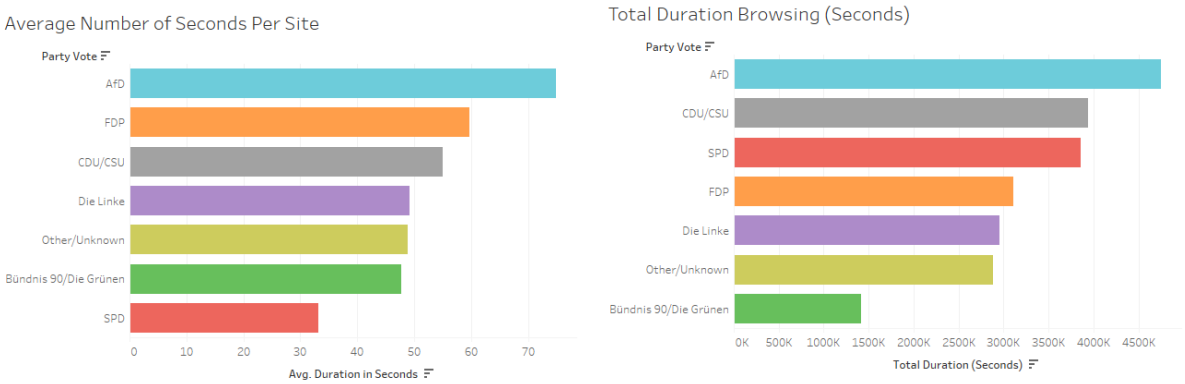
Supplement 1: Browsing Habits and Time of Day

The time of day for browsing might reflect the demographic characteristics of the person. For example, there may be spikes in news browsing after someone wakes up in the morning or during their lunch break on a workday. Alternatively, anyone who is unemployed or working from home may look at the news at any time of the day. The news browsing behavior by party does not show anything unexpected. Most browsing activity occurs between 8 am and 10 pm. The average time of day varies between 2:43 PM and 3:29 PM.

Past research has suggested that certain types of ideologies, and therefore voters who prefer certain parties, visit news sites that align with their views. In the United States, conservatives prefer "Fox News" and liberals prefer "CNN." Based on Supplement 2, this is not the case in Germany. The top five visited news domains for each party vote are the same. However, this also highlights the benefit of this research approach compared to past research. Observing only domains does not explain the content of the actual article. The URL extension which has the news headlines may lean in certain ways that the voters find preferable. Alternatively, it may be that other factors like the total distribution of all domains have a bigger impact on behavior.

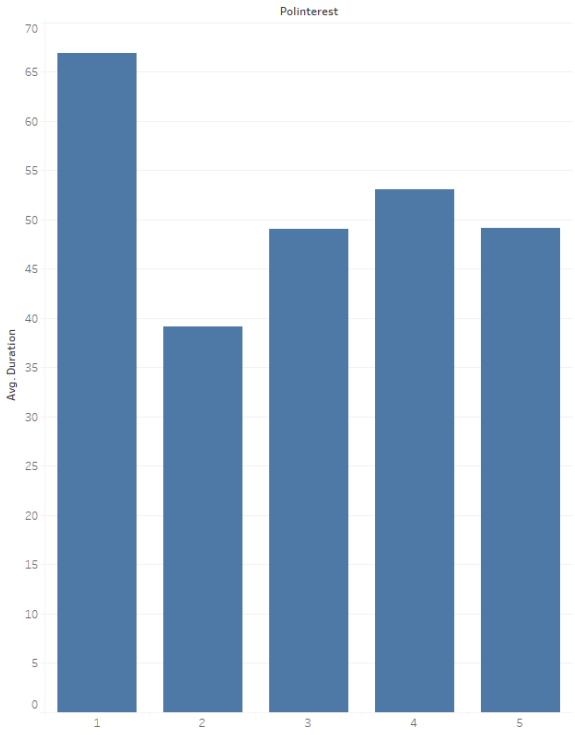


Supplement 2: Top 5 News Domains by Party Vote



Supplement 3: Average Duration and Total Duration on News Sites and Party Vote

AfD voters have the highest average duration per URL visit, and they also spend the most cumulative time on news sites. Interestingly, SPD voters spend the least average duration on a particular site, but they are among the top three in cumulative time on news sites. What this implies is that they are likely to visit a lot more news sites to make up for the difference. Green party voters come among the last in terms of average duration and cumulative duration



Supplement 4: Average Duration by Political Interest

Another interesting finding is that the average duration on the news site does not follow the logical pattern; a reasonable expectation is that the higher the interest in politics, the longer the average duration on a news site. Supplement 8, which is sorted from 1 through 5 representing increasing levels of political interest, shows that the average duration does increase for levels 2 to 4, then drops off at level 5. A potential explanation is that it follows the distribution of the responses [Supplement 4]. However, this does not make much sense since it is looking at average duration, and not total duration. Another explanation may be that a high interest means higher quality engagement, which allows the respondents to read faster. While level 1 interest is the highest average duration, not many people indicated a one, so the average is higher due to a lack of data.

	AfD (N=173)	Bündnis 90/Die Grünen (N=109)	CDU/CSU (N=225)	Die Linke (N=217)	FDP (N=115)	Other/Unknown (N=240)	SPD (N=265)	Overall (N=1344)
Gender								
Man	103 (59.5%)	48 (44.0%)	117 (52.0%)	108 (49.8%)	74 (64.3%)	107 (44.6%)	143 (54.0%)	700 (52.1%)
Woman	70 (40.5%)	61 (56.0%)	108 (48.0%)	109 (50.2%)	41 (35.7%)	133 (55.4%)	122 (46.0%)	644 (47.9%)
University Degree Status								
No University Degree	140 (80.9%)	69 (63.3%)	168 (74.7%)	169 (77.9%)	87 (75.7%)	207 (86.3%)	214 (80.8%)	1054 (78.4%)
University Degree	33 (19.1%)	40 (36.7%)	57 (25.3%)	48 (22.1%)	28 (24.3%)	33 (13.8%)	51 (19.2%)	290 (21.6%)
Employment Status								
Employed	89 (51.4%)	59 (54.1%)	128 (56.9%)	99 (45.6%)	67 (58.3%)	107 (44.6%)	135 (50.9%)	684 (50.9%)
Unemployed	84 (48.6%)	50 (45.9%)	97 (43.1%)	118 (54.4%)	48 (41.7%)	133 (55.4%)	130 (49.1%)	660 (49.1%)
Marital Status								
Married	59 (34.1%)	38 (34.9%)	117 (52.0%)	70 (32.3%)	54 (47.0%)	67 (27.9%)	112 (42.3%)	517 (38.5%)
Not Married	114 (65.9%)	71 (65.1%)	108 (48.0%)	147 (67.7%)	61 (53.0%)	173 (72.1%)	153 (57.7%)	827 (61.5%)
Relationship Status								
Relationship	107 (61.8%)	63 (57.8%)	152 (67.6%)	119 (54.8%)	82 (71.3%)	138 (57.5%)	171 (64.5%)	832 (61.9%)
Single	66 (38.2%)	46 (42.2%)	73 (32.4%)	98 (45.2%)	33 (28.7%)	102 (42.5%)	94 (35.5%)	512 (38.1%)
Political Interest								
High Interest	111 (64.2%)	49 (45.0%)	116 (51.6%)	125 (57.6%)	65 (56.5%)	62 (25.8%)	145 (54.7%)	673 (50.1%)
Low Interest	62 (35.8%)	60 (55.0%)	109 (48.4%)	92 (42.4%)	50 (43.5%)	178 (74.2%)	120 (45.3%)	671 (49.9%)
Monthly Income Bracket								
High HHI	30 (17.3%)	20 (18.3%)	56 (24.9%)	25 (11.5%)	23 (20.0%)	29 (12.1%)	47 (17.7%)	230 (17.1%)
Low HHI	72 (41.6%)	41 (37.6%)	65 (28.9%)	96 (44.2%)	33 (28.7%)	105 (43.8%)	93 (35.1%)	505 (37.6%)
Medium HHI	71 (41.0%)	48 (44.0%)	104 (46.2%)	96 (44.2%)	59 (51.3%)	106 (44.2%)	125 (47.2%)	609 (45.3%)

Supplement 5: Summary Statistics by Party Vote

The figure above shows the breakdown of the survey data by party votes. Overall, the number of supports for any individual statistic is quite low. This makes any skewness in the data extremely apparent. For example, there are only 33 respondents from the Other/Unknown category that have a university degree, compared to 200+ who do not.

	Left			Green			SPD			FDP			CDU			AfD		
Characteristic	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-value**	**OR**	**95% CI**	**p-valu**
Gender																		
Man	—	—		—	—		—	—		—	—		—	—		—	—	
Woman	1.23	0.91, 1.68	0.2	1.43	0.95, 2.17	0.088	0.96	0.72, 1.27	0.8	0.59	0.39, 0.89	0.014	1.05	0.78, 1.43	0.7	0.81	0.58, 1.14	0.2
university																		
No University Degree	—	—		—	—		—	—		—	—		—	—		—	—	
University Degree	1.08	0.74, 1.54	0.7	2.49	1.61, 3.83	<0.001	0.79	0.55, 1.11	0.2	1.06	0.66, 1.67	0.8	1.20	0.84, 1.69	0.3	0.73	0.48, 1.10	0.15
employed																		
Unemployed	—	—		—	—		—	—		—	—		—	—		—	—	
Employed	0.86	0.64, 1.17	0.3	1.09	0.72, 1.66	0.7	0.99	0.75, 1.31	>0.9	1.26	0.84, 1.89	0.3	1.22	0.90, 1.65	0.2	1.06	0.76, 1.48	0.7
married																		
Not Married	—	—		—	—		—	—		—	—		—	—		—	—	
Married	0.88	0.61, 1.27	0.5	0.96	0.58, 1.57	0.9	1.16	0.84, 1.61	0.4	1.21	0.77, 1.91	0.4	1.96	1.38, 2.81	<0.001	0.77	0.52, 1.14	0.2
relationship																		
Single	—	—		—	—		—	—		—	—		—	—		—	—	
Relationship	0.83	0.58, 1.17	0.3	0.78	0.48, 1.27	0.3	1.05	0.75, 1.46	0.8	1.32	0.80, 2.19	0.3	0.79	0.54, 1.15	0.2	1.15	0.78, 1.70	0.5
political																		
Low Interest	—	—		—	—		—	—		—	—		—	—		—	—	
High Interest	1.63	1.20, 2.23	0.002	0.76	0.50, 1.16	0.2	1.27	0.95, 1.69	0.10	1.10	0.73, 1.66	0.6	0.98	0.72, 1.33	0.9	1.95	1.38, 2.78	<0.001
bracket																		
Low HHI	—	—		—	—		—	—		—	—		—	—		—	—	
High HHI	0.57	0.33, 0.94	0.033	1.08	0.56, 2.02	0.8	1.04	0.67, 1.61	0.9	1.09	0.58, 2.03	0.8	1.65	1.05, 2.60	0.030	0.83	0.49, 1.38	0.5
Medium HHI	0.85	0.61, 1.18	0.3	0.98	0.62, 1.55	>0.9	1.09	0.80, 1.50	0.6	1.30	0.82, 2.10	0.3	1.19	0.84, 1.71	0.3	0.78	0.54, 1.13	0.2

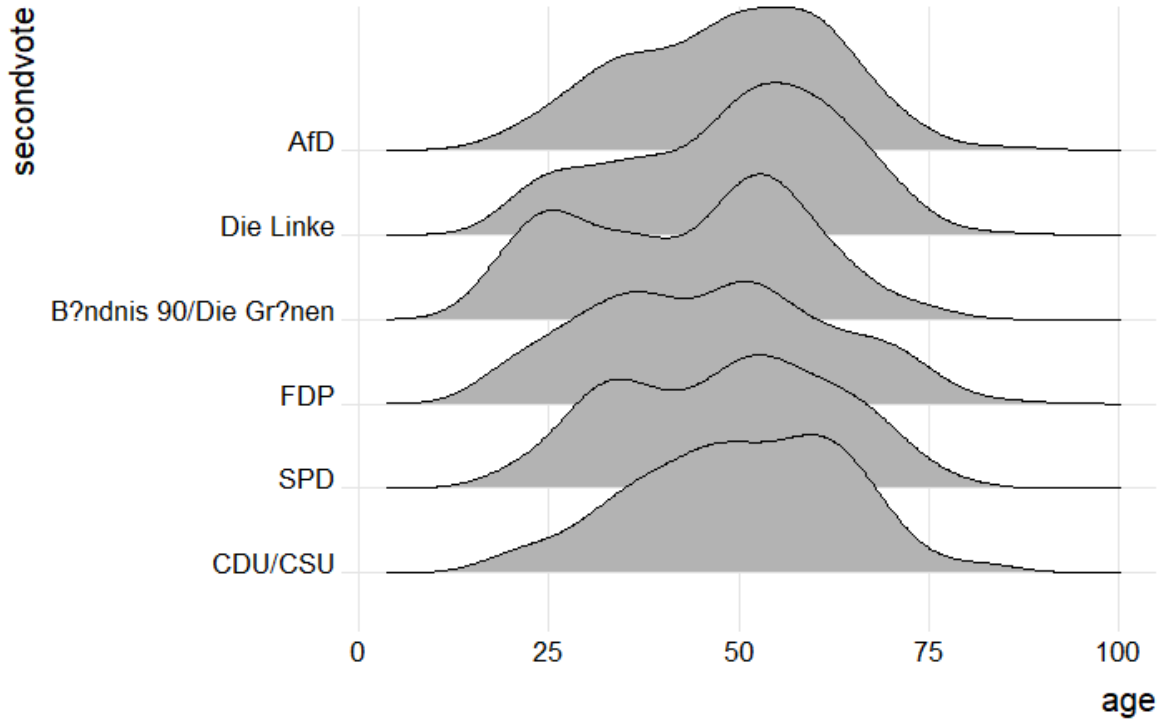
Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Supplement 6: Odds Table by Party Vote

Most results from a generalized linear model did not yield statistically significant results. The only few that did are the following ($p<0.05$):

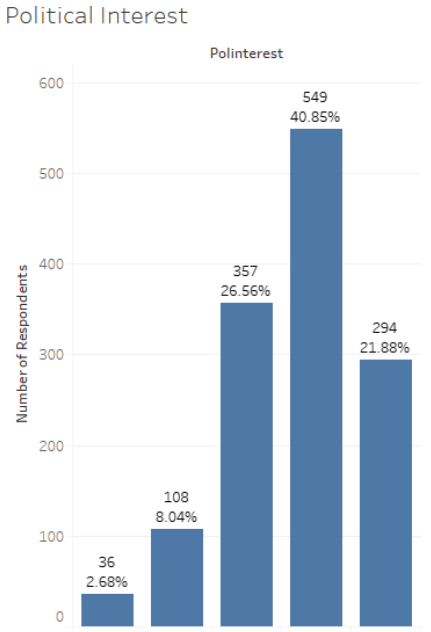
- Lower odds for women in FDP (0.01)
- Higher odds for university degree Greens (0.001)
- Higher odds for political interest for Die Linke (0.002)
- Higher odds for political interest for AfD (0.001)
- Higher odds for marriage for CDU (0.001)
- Lower odds for high income for Die Linke (0.03)
- Higher odds for high income for CDU (0.03)

Considering there is a combination of 50+ options (i.e., of party × categorical features), then a larger sample size is probably needed for more statistically significant results.



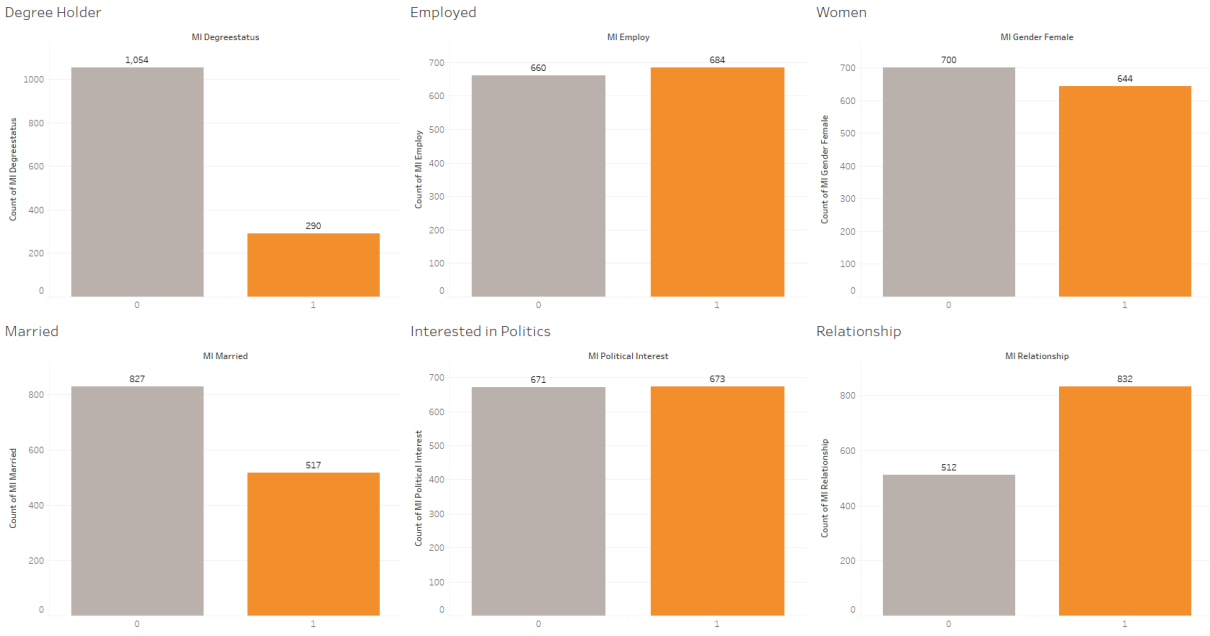
Supplement 7: Ridgeline Plot of Ages by Party Vote

The figure above shows that the age variable is not normally distributed. The distributions of ages are either skewed, favoring over age, like CDU/CSU, AfD, and Die Linke, or the distribution has twin peaks as with SPD, FDP, and Greens. The more left-wing parties seem to have twin peaks suggesting that they garnered support from younger voters. This was likely true at the time of the 2017 elections before the noticeable trend of younger voters voting for conservatives. The more right-wing parties, CDU/CSU and AfD are older, which is to be expected.



Supplement 8: Political Interest Levels by Respondents

Most respondents said that they are level three or four in political interest. Since level three is not particularly informative, we decided to group levels one through three with low political interest, while four and five remained high interest. This also had the added benefit of having sized classes [Supplement 5]. Note: non-respondents are also put in the not-interest category.



Supplement 9: Demographics Breakdown

The reason these continuous variables are binarized is so that machine learning models can easily read them. Luckily, the employment, gender, and political interest features are nearly balanced. The marital status and relationship features are not particularly balanced, but they are kept because they are crucial factors in a person's livelihood. For example, some would attribute the current rise in males voting for far-right parties to the "male loneliness epidemic" (Peterson et al., 2025). This can be reflected in relationship status. Meanwhile, marriage represents a more long-term commitment, and people who are married receive privileges from the government that a normal relationship does not. The most unbalanced feature class is the university degree status. The reason we kept this is that education correlates with higher support for left-wing parties. Future research should change university degree status to something else separating education categories, like high school graduation. In Germany, there is a sizeable number of people who do not complete a university degree and do an "Ausbildung" (e.g. trade-school) instead.

Supplement 10: LDA Topic Results

Unsupervised classification does not automatically label the outputs; that is up to the researcher to decide. While there is no quantifiable metric we used to decide topic labels, we used heuristics to say what the given topic is thematically. The "size" of the words in terms of presence and the commonality between the words determined the topic label (note the figure above has it in German and translated to English):

0. **Economy**: Investment, Securities, Boss
1. **Sport:** Football, Ticker, Play
2. **Elections:** Party, Candidate, *Schulz*
3. **Politics:** Euro, Policy, Money
4. **Sports:** Broadcast, Sport, Comment
5. **Current Events:** Earthquake, Opinion, Finance, *Bundestag*
6. **Entertainment:** Fashion, Film, Program, TV
7. **Entertainment:** Soaps, Episode, *Folge*, *Staffel*
8. **Law and Order:** Article, Police, Region, Protest
9. **General:** The words were small and diverse

Note the duplicate topics (highlighted). The total number of topics is 8. However, there is also an unassigned topic that is initialized, so the total number of "topics" is 9. This approach increased the accuracy of neural networks from 41 percent to 48 percent, although the reason for this is unclear. It may be recommended to simplify the topics even further for better results, especially because of overlap.

II. Data Challenges

Data cleaning occupied the majority of the time in this research. Because URLs are uniquely messy due to every site using a different URL extension scheme, methods of cleaning needed to be developed for this project. The first major decision was to decide from what part of the URL to drop. For example, some news sites have their news headline after the first slash (/). This also introduces a lot of irrelevant information for the URLs that have their headlines after the second slash (e.g. "/politics/article-headline"). Ultimately, it was decided to use the second method to reduce data sparsity.

While not employed in this research, future studies looking into URL-level data after the extension should remove any back-to-back consonants (exact implementation varies). This helps to remove nonsense URL extensions, especially from websites that do not show their headlines but some random ID. Alternatively, researchers should use a lexicon of real words and remove any words that do not exist.

III. Data Processing Decisions

- Everyone who did not respond to the question voted_2017 was removed. This made the dataset of respondents go down from 1516 to 1344. The true number of respondents who voted in the election is 854, but some of them still answered who they would vote for (e.g., they wanted to vote, but they could not). This means that voters' preferences are still reflected.
- "Firstvote" was not used because it asked about the candidate's party.
- Initially, the ideology category ranged from -3 to +3. However, it is not particularly easy to decide which party was more left- or right-wing than another party. Therefore, the range was changed to -1 to 1. Since many machine learning models do not like negative numbers, the values were made into a categorical variable (from 0 to 2).
- For the sake of simplicity, the parties were divided into a pseudo-left-right binary. However, in Germany, this delineation can be difficult to determine, and some parties overlap on issues. The FDP is especially difficult since they are conservative in some respects and liberal in others (like the decriminalization of cannabis). In addition, the definition of "liberal" is used in the American sense, since in Germany, "liberal" is used in the "neoliberal" sense (e.g. "liberale politik").
- Numbers that typically come in pairs of two and four were kept because they reflect how dates are written (so DD-MM or DD-YYYY or YYYYY-MM). All other numbers are removed. This does mean that for some article-headlines, quantities are removed.
- German "stopwords" like "der", "aus", etc. were removed.
- Unique words that end with "id" were searched using regular expressions (REGEX) and removed manually based on whether it was thought to refer to a word or a URL code.
- Common URL-related phrases like "html", "pdf", "logout", etc. were removed.
- News websites themselves like "bild" or "zdf" were removed because they are not informative of the article headlines. Note that this was not an exhaustive process, there is a chance that many news sites were overlooked because they were not detected

Statement of Authorship

I hereby confirm and certify that this master thesis is my own work. All ideas and language of others are acknowledged in the text. All references and verbatim extracts are properly quoted and all other sources of information are specifically and clearly designated. I confirm that the digital copy of the master thesis that I submitted on April 26th is identical to the printed version I submitted to the Examination Office on April 28th.

DATE: April 16, 2025

NAME: Ray Hossain

SIGNATURE: