# Executive Summary

# Introduction

Technological change and a swift growth in data availability are leading to a crossroads for democracy. Individuals’ online behavior can be traced and used to create personalized profiles to strategically target them with advertisements aimed at influencing everything from their consumer behavior to political preferences. These Big Data advancements have resulted in a modern form of gerrymandering, building digital boundaries around electoral constituencies and communities (Gurumurthy & Bharthur, 2018), resulting in increasing polarization both online and offline.

Strategic digital targeting to achieve political aims was brought to the public’s attention with the Obama campaign, and solidified as a major public issue during the 2016 Presidential election. In particular, the Cambridge Analytica scandal which came to light in 2018 illustrated how technology had allowed for individuals’ Facebook data to be harvested and used to target users with personalized messaging based on their psychographic profiles in an effort to influence the election outcome (Gurumurthy & Bharthur, 2018; Kruschinski, 2017).

Pro-technology perspectives highlight how this micro-targeting and enhanced forms of digital communication can in fact bolster democracy. For example, digital intelligence can facilitate improved grassroots organizing, bringing in those who may be less involved in the political process and providing smaller causes with the opportunity to be competitive with wealthy, entrenched power structures (Kruschinski, 2017). Technology has also played a major role in facilitating organization in major social and civic movements, such as the Arab Spring or the contemporary protests in Hong Kong (Howard & Duffy, 2011; Shao, 2019).

However, many argue that while there are clearly benefits for grassroots movements with data-based electioneering, the technological platforms that enable them are controlled by the elite, who may not always have the best interest for democracy in mind (Gurumurthy & Bharthur, 2018). Critics fear an erosion of privacy, and manipulation of the citizenry such that electoral outcomes no longer reflect a “democratic mandate or informed choice” (Gurumurthy & Bharthur, 2018; Kruschinski, 2017). Not only can electoral outcomes be influenced by the elite within a country, but democracies are also increasingly facing challenges in the form of data-driven interference on the part of international actors (Jamieson, 2018). These trends influence erosion of trust in institutions, increasing social and political polarization, and heightened cynicism in the political process (Dimock, 2019).

The ability to effectively mitigate these negative impacts is being impeded by both the modern narrative of the “ungovernability” of major technology platforms, as well as a general lack of understanding about what can and is being done with AI and Big Data. These developments have been cast as “neutral tools of economic progress and social advancement” that all nations must embrace as a necessary path to “technomodernity” (Gurumurthy & Bharthur, 2018). Additionally, technological change is taking place at break-neck speed, and often requires a high-level of knowledge to comprehend thoroughly enough to make reasonable, future-proof policy decisions.

The inherent complexity surrounding these issues also limits the public’s understanding of major issues such as how companies harvest their data and what can be learned from it. Concerns are often voiced over a lack of transparency regarding who uses these data, and how the algorithms that make targeting decisions actually function (Bach et al., 2019).

For this reason, it is critical that individuals understand how their personal data can be used to infer private details which can be used to create the profiles necessary to enable personalized micro-targeting. Prior research has shown that online behavior patterns may indeed offer insight into individual’s political preferences, though results have been mixed. For example, Bach and colleagues (2019) were not able to meaningfully predict political preferences based on individual-level browsing behavior, while Pennacchiotti & Popescu (2010) were able to determine political leanings based on the vocabulary used in individuals’ Twitter posts.

Most prior studies have relied on social media posts and web browsing history, like the previous two examples. One particularly interesting avenue of research has focused on somewhat of a combination of these: search engine queries. Search engine queries have the potential to be particularly illustrative because 1) they may provide insight into *both* the topics one is interested in online as well as the language used to search for them and 2) individuals are known to be particularly honest when conducting web searches, often treating the search query box as something of a confidant. For example, search engine queries are often formatted linguistically like a sentence, with individuals asking deeply personal questions about their health, relationships, or anxieties (Stephens-davidowitz, 2017).

Prior research on the predictive power of search engine queries has primarily relied on aggregate-level Google Trends data. This paper seeks to augment the existing literature by analyzing individual-level search engine query data over many search engine platforms, utilizing both keyword-methods as employed by prior researchers as well as modern text analysis methods in an attempt to uncover the intention behind a search, not just the words used.

**More on findings etc.**

# Literature Review

To place this research in context, a brief overview of the current state of the literature follows. First discussed are public perceptions of the acceptable uses of online data and the potential ways these data could be exploited. Next, a theory-driven social science perspective on why digital data may be predictive of political party preference is defined, followed by a brief review of prior studies that have successfully used such data to achieve these means. The benefits of search engine data in particular are then discussed, as well as the specific contribution of this research, particularly regarding the individual-level nature of the data.

## Digital Data: Public Perceptions and Implications for Democracy

Beginning with Obama’s first presidential campaign, data-driven microtargeting has been a major theme for many major political campaigns, featuring prominently in the 2016 Trump and Clinton campaigns, for example (Kruschinski, 2017). Digital data can be used to develop profiles of individuals in an effort to curate and deliver tailor-made messaging that is as efficient as possible for eliciting a desired behavioral response. The Cambridge Analytica scandal brought the issue to the public’s attention, with critics warning of the potential to manipulate voters and erode privacy, and supporters pointing to the potential benefits microtargeting has for mobilizing specific target groups, or those who may be naturally less inclined to vote (Kruschinski, 2017).

Ur and colleagues investigated non-technical users’ attitudes towards Online Behavioral Advertising (OBA), which relies on users’ browsing history to deliver custom ads (both political or commercial). They found that participants felt OBA was both useful and simultaneously “creepy,” expressing concerns about privacy. While participants were aware that contextual targeting was taking place, many were surprised to know that not only could online behavior theoretically be used to tailor advertising, but that this is already common practice. Concerns were particularly pronounced when participants felt that the profiles generated to target them were inaccurate: For example, if they felt stereotyped or that they were receiving ads that weren’t representative of their true interests (Ur, Leon, Cranor, Shay, & Wang, 2012; Dolin et al., 2018).

Several studies have found that individuals lack a fundamental understanding of how such targeting takes place, and that users are even less comfortable with targeted advertising once they gain a fuller understanding of how the data are gathered (Dolin et al., 2018; Ur et al., 2012).

While in theory such data-driven targeting could be a boon to grassroots campaigners and allow organizers to bring more individuals into the political sphere, critics have expressed concern about the ability for the political elite to exert undue influence on the political process. Gurumuthy and Bharthur warn against a future where the unethical use of Big Data and AI “allows political influence to move from public campaigns to private sentiment, a shift that repositions electoral politics from a spectacle that is overt to a script that is covert,” where voter behavior is manipulated such that outcomes no longer reflect informed decision-making or the democratic will (Gurumurthy & Bharthur, 2018).

The opportunity to exploit these new micro-targeting capabilities also exists for foreign powers, not just the elite within a particular country. Indeed, the 2016 U.S. Presidential election saw probable evidence of Russian interference that relied on data from social media to target particularly-relevant constituencies in an effort to bolster the Trump campaign (Jamieson, 2018).

In this context where the availability of digital data is growing at an unprecedented rate, machine learning is becoming an ever-more powerful tool, and the public lacks a general awareness about and comfort with the modern targeting applications, it is critical to examine the extent to which digital data, such as search engine queries, can actually be used to efficiently target individuals to influence political processes.

## Theoretical Basis for Search Engine Query Data as a Predictor of Political Preferences

Prior studies relying on digital data to predict political preferences have been criticized for lacking a solid theoretical basis to qualitatively explain their results. This can lead to poor reproducibility, and increases the chances of incorrectly interpreting results they may be simply due to chance (C. Lui, Metaxas, & Mustafaraj, 2011; Yasseri, 2016). This paper argues that differences in browsing behavior and linguistic choices are meaningfully related to differences in demographics, which also share strong associations with political preference.

Browsing behavior has been shown to vary on the basis of demographic characteristics, which are in themselves associated with differences in political affiliation. For example, Hu and colleagues describe how women are more likely to seek medical or religious information online than men are (Hu, Zeng, Li, Niu, & Chen, 2007). In turn, gender is also correlated with party identification, with women in the United States being more likely to favor the Democrats (Pew Research Center, 2016). Similar relationships hold true for other characteristics such as age and level of education, which have also been shown to vary with ideology (Hu et al., 2007; Weber & Castillo, 2010).

Weber & Castillo showed similar findings for web search behavior in particular, based on factors such as the length of queries and the web pages visited after a search. They determined that “demographic factors have a measurable influence on search behavior:” For example, queries beginning with the first name “Hal” in low-education areas typically ended the search with the last name “Lindsey,” in contrast to those in higher-education areas where “Higdon” was the more common last name (Weber & Castillo, 2010). Query language has also shown to be able to predict age and gender (Jones & Tomkins, 2007).

Differences in query language varying systematically based on demographic characteristics is not surprising in the larger context of linguistic variety. Indeed, several studies have found that blogger age and gender are inferable on the basis of linguistic choices such as length of a post and the words contained, punctuation, capitalization, and general prose style (Burger & Henderson, 2006; Nowson & Oberlander, 2005). Formal written texts have also been found to vary in a meaningful way on the basis of age and gender (Argamon, Koppel, Fine, Shimoni, & Science, 2003; Koppel, Argamon, & Gan, 2000).

Even smaller strings of text in the form of Tweets have been able to predict demographics and political preference. For example, Democrats and Republicans “tend to use a specific vernacular (‘obamacare’) when discussing issues of interest to both sides (healthcare reform)” (Pennacchiotti & Popescu, 2010). Rao and colleagues also work on Twitter data, but emphasize the importance of sociolinguistic cues. For example, character repetition (e.g., “that’s soooo crazy”), is often indicative of a female writer, as are the use of emoticons or multiple exclamation points (“!!!”). Like Pennacchiotti and Popescu, Rao et al. note particular vocabulary as being particularly illustrative, with certain terms like “dude” or “bro” being strongly associated with younger writers (Rao, Yarowsky, Shreevats, & Gupta, 2009).

Thus, web browsing behavior generally, as well as linguistic decisions even in short text (such as search queries), have been shown to be able to illustrate differences in demographics, which are also clearly associated with differences in political preferences (Pew Research Center, 2016).

In addition to how people search and behave online, the simple condition of whether or not they make a politically-oriented query or the volume of such queries can be a meaningful for explaining the importance of search queries from a theoretical perspective. Yasseri and Bright elaborate:

*“We base this theory on a rational choice approach to explaining voting behavior, which conceptualizes voters as similar to consumers in a market, seeking to vote for the political party who offers them the greatest “pay-off” in terms of policies​. Online information seeking, from this rational choice perspective, can be explained in terms of voters looking for more information about the election: perhaps about practical matters such as how to vote, or perhaps about substantive matters such as which political party might suit them best. Such information seeking is rational in that it increases the chance that the voter will vote for the party which represents them best.”* (Yasseri, 2016)

## Prior Uses of Digital Data for Prediction of Political Preferences

The possibility for harnessing the predictive power of online behavior data is not particularly new, with prior studies utilizing information on Facebook and Twitter posts (including the volume of posts and the sentiment of the related text), web browsing data, and aggregate level Google Trends search query information, with mixed success.

Early work by Tumasjan, Sprenger, Sandner, and Welpe claimed that Tweet volume could be used as an alternative to traditional polling, and that the sentiment of politician’s and parties’ Twitter messages “closely corresponds to political programs, candidate profiles, and evidence from the media coverage of the campaign trail” (Tumasjan, Sprenger, Sandner, & Welpe, 2010). Tweet sentiment analysis has also been found to correlated to presidential job approval polls (O’Connor, Balasubramanyan, Routledge, & Smith, 2010).

However, these works and others have faced criticism for their lack of reproducibility and disregard for sample representativeness. For example, Chung and Mustafaraj applied the same methods employed by Tumasjan et al. (2010) and O’Connor et al (2010) to a new dataset and found it was unable to result in an accurate prediction on a new sample (Chung & Mustafaraj, 2010). Similarly, Schoen, Jungherr, and Ju showed that even using the same case as the Tumasjan et al (2010) study, vastly different results were achieved through the inclusion of a different set of parties or timeframe, both of which appeared to be arbitrary choices in the original paper, thus also placing doubt on the ability of this method to generalize to future elections (Schoen, Jungherr, & Ju, 2012).

The inability for Twitter to consistently predict election outcomes or political preferences is unsurprising given that social media users differ in meaningful ways from the electorate at large. Indeed, even the prevalence of bots and spam accounts should make one question the reliability of such a sample (Gayo-avello, 2011). Additionally, there is likely to be a significant influence of self-selection bias, as those who are active on social media are likely to be the most politically-oriented and perhaps ideologically extreme (Gayo-avello, 2012).

Gayo-avello (2012) outlines several key ways to overcome the challenges perceived in prior studies utilizing social media data. Namely, a credible baseline should be established (as discussed in section XX), the study timeframe should be clearly specified and justified, only data from eligible voters should be included, state of the art sentiment analysis should be employed over simplistic methods, spam should be removed, and bias in the data should be analyzed and acknowledged (Gayo-avello, 2012).

Another avenue of research focused on the predictive power of online browsing history as a whole, thus overcoming several of these issues, in particular due to the fact that internet users generally are more representative of the electorate than users of a particular social media platform.

Comarela, Barford, Christenson, and Crovella (2018) found that web browsing history was able to predict candidate preference at rates in line with modern polling techniques. They focus on a state-by-state and day-by-day analysis, comparing the web browsing data of 100,000 individuals over a 56 day period shortly before the 2016 US Presidential election to statewide polling data, thereby overcoming the common challenge of missing individual-level “ground-truth” labels. Their results showed that domain-level URL visit history was able to predict election results with a comparable accuracy to polling (with linear correlation of 0.94), and that the fine-tooth nature of the method allows for the analysis of the impact of a specific event, such as the release of the “Comey letter” on a day-by-day and state-by-state basis (Comarela, Barford, Christenson, & Crovella, 2018).

Bach and colleagues conducted a similar study, though with the advantage of having survey data on political preference to augment their corpus of browsing data for 2,000 German adults eligible to vote in the 2017 federal election, for four months before and after the vote. However, they found that online browsing behavior was not a strong predictor of self-reported voting behavior in their sample. In particular, their model struggled to identify undecided voters, though performed better for parties at the political periphery, such as the Greens and AfD (Bach et al., 2019).

This addition of individual ground-truth labels is quite rare in most prior studies utilizing digital data for predicting political preferences. Indeed, the few studies that have prioritized predicting individual-level characteristics have primarily relied on differences in linguistic features (Pennacchiotti & Popescu, 2010; Rao et al., 2009), though Hu and colleagues also took an individual focus when analyzing differences in web browsing behavior (Hu et al., 2007).

At the present time no studies that have taken an individual-level approach to using search queries as a predictor of political preferences are known. However, collective applications of search query data both for illustrating current events (“predicting the present” – as coined by Choi & Varian, 2011) as well as forecasting future outcomes. Google Trends data has been successfully used to forecast topics such as unemployment (D’Amuri & Marcucci, 2017), housing prices (Wu & Brynjolfsson, 2015), consumer purchasing behavior (Goel, Hofman, Lahaie, Pennock, & Watts, 2010), and the spread of influenza (Ginsberg et al., 2009).

Google Trends data have also successfully been applied to elections. Stephens-davidowitz showed that search volume for the terms “vote” or “voting” in a particular geographic area was strongly correlated with the electoral turnout in the region in the 2008, 2010, and 2012 US elections (Stephens-davidowitz, 2013). Street, Murray, Blitzer, and Patel (2015) also showed that queries related to voter registration were strongly correlated with actual voter registration rates.

Polykalas, Prezerakos, and Konidaris (2013) applied a similar method to study the outcome of three German elections, also relying on a pre-defined list of keywords that were determined to be relevant for electoral outcomes and measuring their relationship to election results. The algorithm was able to accurately predict the election outcome of all three of the studied elections (Polykalas, Prezerakos, & Konidaris, 2013).

Lui, Metaxas, and Mustafaraj (2011) cast doubt on the applicability of Google Trends data to forecasting elections, however. They argue that such data were not successful at predicting the 2008 and 2010 US elections compared to incumbency, polls, or even chance They point out that this could be due to limitations on simply using search volume for a particular candidate’s name, since this does not adequately illustrate *why* an individual may be searching for a candidate. For example, if a candidate is particularly well-known, they may not be searched for at all, which is actually a good sign for their election prospects. The researchers therefore recommend employing sentiment analysis to get an understanding for the driving forces behind a user’s query (C. Lui et al., 2011).

It is also important to note that aggregate-level Google Trends data have some notable flaws for forecasting: For example, there is no way of knowing who is using Google to confirm that they are eligible voters, or how often – the same individual may Google a candidate name many times, for instance. Thus, it is difficult to assume a “one person, one vote” scenario is represented with Google Trends (Chung & Mustafaraj, 2010).

## Strengths of Search Query Data

Much of the interest in utilizing digital data to forecast political preferences is to overcome the drawbacks inherent in traditional polling. For example, polls require significant time and monetary resources, and hence “cannot give insight into the short-term dynamics of vote choice, especially on a per-state level.” Results can also be sullied through interviewer effects, word choice, question order, or even reticent respondents (Comarela et al., 2018).

Additionally, self-reported vote forecasts have been shown to often be misleading. Rogers and Aida (2014) examined seven pre-election surveys with post-election vote validation and discovered that many predicted voters do not vote after all, and many who say they won’t vote actually do. Additionally, self-predicted voters differ significantly from actual voters, though there is little difference between self-predicted voters and non-voters, thereby showing that, “Vote self-prediction is “biased” in that it misleadingly suggests that there is no participatory bias” (Rogers & Aida, 2014).

Another concern with polling is that participants may be untruthful about their voting intentions when they hold views they believe to be socially undesirable, such as racial animus, or even one’s intention to vote for a polarizing candidate like Donald Trump (Brownback & Novotny, 2016). There is some evidence to suggest that search query data may be able to combat this issue. For example, Google searches are unlikely to exhibit major social censoring, because users are typically acting alone, and online (Stephens-davidowitz, 2012). Additionally, in a study on user perceptions of web-based information disclosure, participants expressed that they are typically honest when conducting web searches (Conti, Point, York, & Sobiesk, 2007).

Therefore, search query data – if it is able to successfully predict political preferences on an individual level – may be more accurate than self-reported voting intention, both because it avoids common polling issues, as well as the impact of social desirability bias.

## Contributions of this Research

This research builds on prior studies that have relied on digital data to predict political preferences in several meaningful ways:

1. The unique nature of the dataset means that individual-level search query history is able to be compared to ground-truth, individual-level survey data on political affiliation and voting history. This also ensures that bots are not polluting the data, and the demographics of the sample can be confirmed and compared to those of the electorate at large.
2. While much work has already been done using Google Trends, prior studies have relied exclusively on pre-defined lists of terms, and measured their search volume. However, there is a clear opportunity available in utilizing the complete search history, since perhaps there are specific words, topics, or themes that are unexpectedly predictive of political preferences.
3. Similarly, sentiment analysis will be implemented to augment the keyword-only approach taken by previous researchers. This will help to create a fuller understanding of *why* someone is searching for a particular term (C. Lui et al., 2011).
4. No recent prior studies have considered contemporary search engines other than Google, despite the fact that users of Bing.com or DuckDuckGo.com may differ in meaningful ways from Google users.

# Dataset

## Survey and Search Query Data

The data for this research come from the YouGov Pulse panel survey *Paying Attention to Attention: Media Exposure and Opinion Formation in an Age of Information Overload,* running from April 2018 through November 2018 over five waves. Each wave was comprised of a nationally-representative sample of US adults (for further information, see Appendix A). The original sample was 1,339 individuals, and once filtered for those with adequate search history data and a response on the outcome measures, the analytic sample totals 708. Survey data for this analysis comes from the fifth survey wave, which took place between December 12th, 2018, and January 7th, 2019. Survey questions included a variety of demographics, as well as voting intention before the 2018 midterm elections, and whether and for whom the participant voted for after the election.

In addition to the panel survey data, web tracking data (which includes all search engine queries) was passively collected through YouGov Pulse. Participants consented to installing Reality Mine, software which tracks web browsing history in real time – with the exception of sensitive items such as passwords and financial transactions.

An important note is that the data faced one major challenge: Only a subset of the participants had full URL information available, and thus information on the search query text used. This led to a significant decrease in the available sample size. Luckily, those with and without full URL information do not differ in statistically significant ways. For an overview of how potential differences were analyzed, see Appendix B.

## Target Variable

The goal of this research is to determine if it is possible to predict 1) whether an individual will vote in an upcoming election and 2) if so, for which party for the House of Representatives election (the Democrats or Republicans).

For Question 1 regarding turnout, the following question was re-coded as binary (with positive responses to option 5 “I definitely voted in the midterm election on November 6” coded as voted, and all others as did not vote), and all NA answers were removed:

Which of the following statements best describes you?

1. I did not vote in the election this November
2. I thought about voting this time, but didn’t
3. I usually vote, but didn’t this time
4. I attempted to vote but did not or could not
5. I definitely voted in the midterm election on November 6

Those who responded that they voted were further asked:

For whom did you vote for the U.S. House of Representatives?

1. The Republican candidate in my congressional district
2. The Democratic candidate in my congressional district
3. The Independent candidate in my congressional district
4. I did not cast a vote for the U.S. House

If the respondent answered with option 4 “I did not cast a vote for the U.S. House” or the response was coded NA, these responses were removed from the dataset. Similarly, Independent voters (option 3) were removed given that they totaled only 11 respondents in this dataset. Thus, the remaining two options for Question 2 were that the respondent voted for the Republican or Democrat.

## Features

Intro

**KEYWORDS** In line with prior research that has focused on the volume of queries for certain keywords (e.g. (D’Amuri & Marcucci, 2017 or Stephens-davidowitz, 2013), an initial line of inquiry will focus on:

1. Whether or not an individual searched for terms related to registering to vote.
2. If they searched for terms related to politics generally (i.e. “congress,” or “democrat”), or the names of any member of the House of Representatives, Senate, or well-known political figures (including those not serving as of the time period of this study, such as Hillary Clinton).
3. Whether or not they searched for the candidate names of those running for the House of Representatives in their state.
4. A regularized list of all keywords, whether political or not.

**TOPIC**

**SEARCH BEHAVIOR** Given that individuals’ online behavior is known to vary on the basis of socio-demographic characteristics that are associated with voter status and party preference, users’ search behavior (without concern for the text of the query itself) will also be analyzed. This includes:

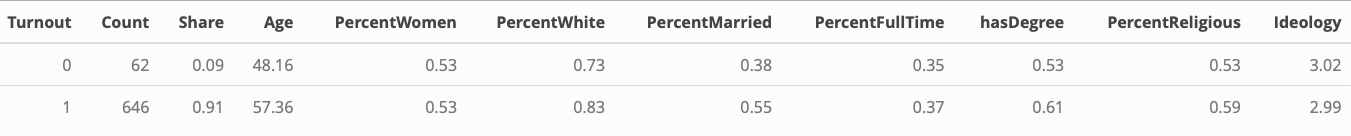
1. The search engine used
2. Mean number of queries per day
3. Average query length
4. Etc.

## Descriptive Statistics

### Sample

After filtering the dataset for individuals with complete URLs who also answered whether or not they turned out in the 2018 U.S. midterm election, the final dataset (N = 708) is composed as follows.

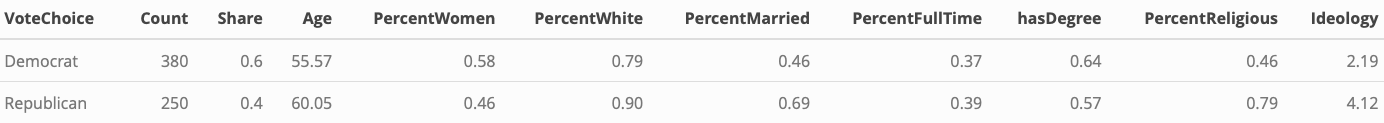
First, in terms of outcome variable 1 – whether or not an individual turned out to vote – there is a notable class imbalance in this dataset, with 91% stating that they voted in the 2018 election. This is, of course, not in line with the actual 2018 midterm election in the United States, which – while high for such an election – was only 53% (US Census Bureau, 2019). This discrepancy could be due to errors in sampling, panel conditioning, or – as was discussed in the literature review – the result of social desirability bias. It is notable that while online behavior data may present an avenue to avoid some of these challenges, the research process necessarily involves relying on self-reporting, a necessary limitation of the setup.



In terms of other demographic differences, it is notable that voters are 9.2 years older than non-voters, on average. The share of women is the same in both groups, as is ideology (which is measured from 1 for those who identify as “very liberal” to 5, “very conservative”). Among voters, 61% have at least a two-year college degree, in contrast to only 53% of those who did not vote, and 55% are married, unlike 38% of those who did not vote. Voters are also marginally more religious, on average.

Taking a look at only those who claimed to have voted for a Democrat or Republican candidate in the 2018 election, it is evident that this dataset skews towards the Democrats, with 60% of the sample favoring that party. Only 40% voted for the Republican candidate in their congressional district.

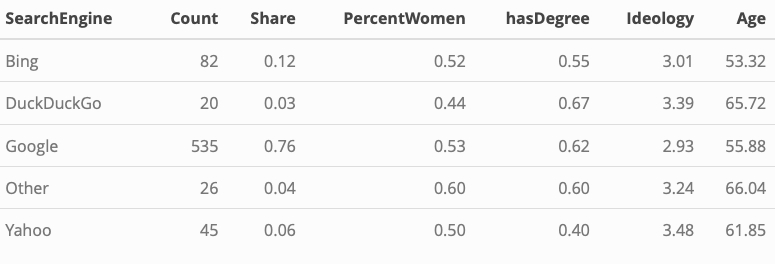
There are also notable differences in the demographic makeup of each sample. In particular, unlike the distinction between voters and non-voters, gender differences appear to be more pronounced, with women making up a much greater share of Democrat voters versus Republican. Republican voters are also approximately 4.48 years older than Democrat voters in this sample, and are more likely to be white, married, and religious. Unsurprisingly, ideology is clearly associated with party preference, with Democrat voters identifying towards the liberal end of the spectrum (values closer to 1), and Republicans towards the conservative end (with values closer to 5).



These findings align well with the general academic understanding of party coalitions in the United States, which describe the Democrats as younger, more diverse, less religious, and more likely to be female than their Republican counterparts (Pew Research Center, 2016).

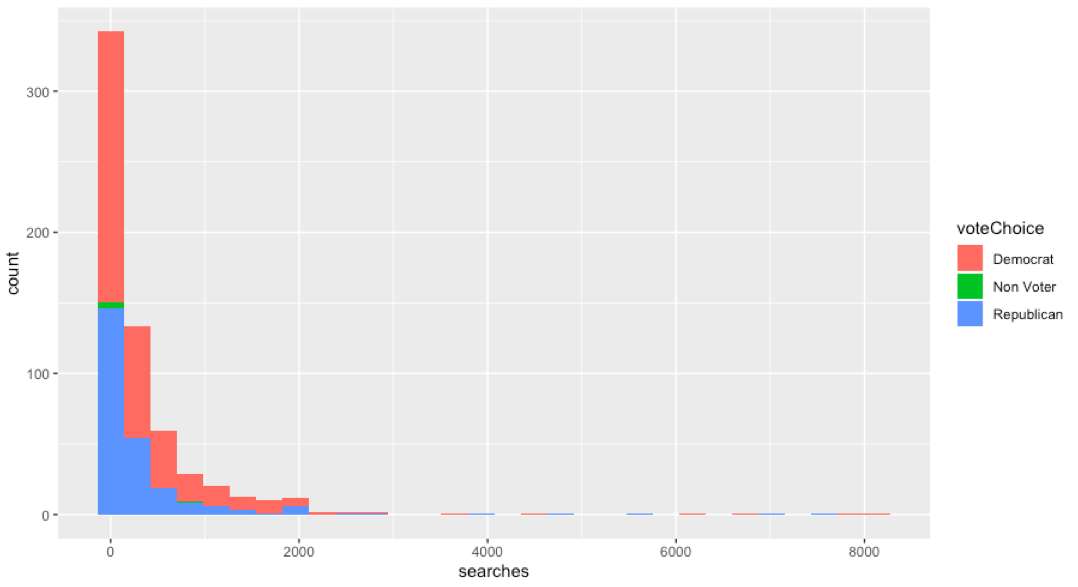
### Search Behavior

**SEARCH ENGINE** One advantage of this research in comparison to prior studies is the ability to leverage query data from other outlets other than Google. This is relevant, because the user groups for different search engines vary in meaningful ways. Additionally, while Google is the market leader (comprising 54% of the individual search queries and 76% of the users in this dataset), Bing in particular makes up a significant portion of the other queries, at 40%, though only 12% of the users. Also, interestingly, there is absolutely no overlap among the users of the various search engines – no participant made queries using both Google and Bing during the entire timeframe of the study, for example.

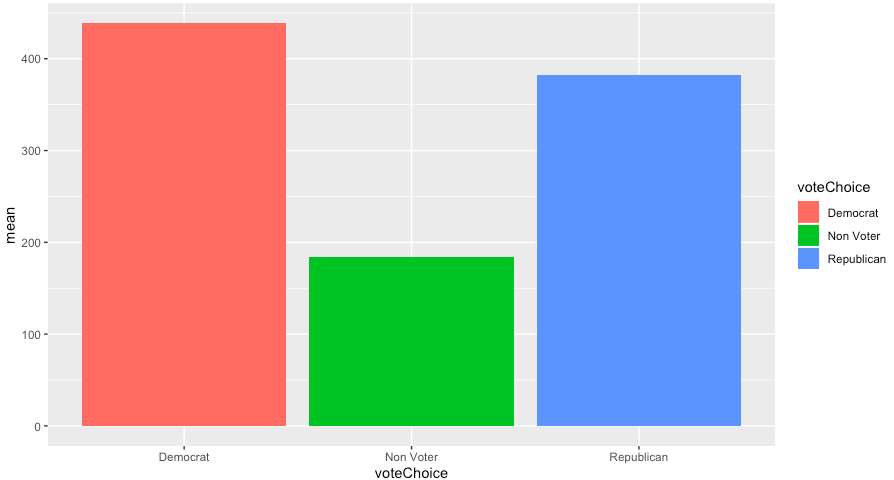


Demographically, there are some interesting, though subtle differences. Women make up an approximately even share of the user base of Bing, Google, and Yahoo, though are less likely to use DuckDuckGo, and more likely to use fringe (“Other”) platforms. Education levels are highest among DuckDuckGo users, followed by Google, and other platforms. The percentage of individuals with a college degree is notably lower among Yahoo users. DuckDuckGo and Other users are also quite a bit older, followed by Yahoo users, with Bing having the youngest audience – about two and a half years younger than Google users. Ideological differences are perhaps the most interesting, with Google users being the most centrist, and DuckDuckGo and Yahoo users the most conservative.

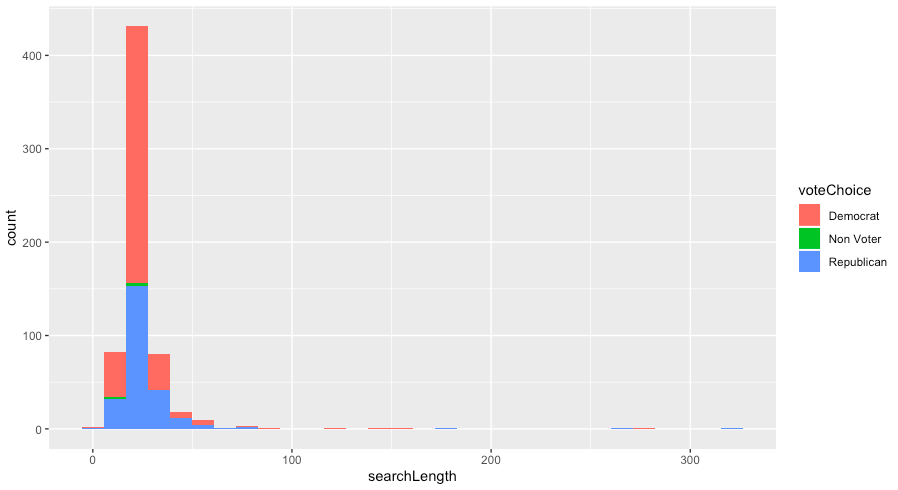
**NUMBER OF QUERIES** For the entirety of this project, only the first sequential search query is considered. For example, if someone input the same query multiple times in the same date, all queries after the first were dropped from the dataset. After this pre-processing, throughout all users over the entire dataset, the mean number of queries was 408.6, though with significant left skew (demonstrated by a median of 93.5). This ranged from only 1 search to 8,139. Remove these?



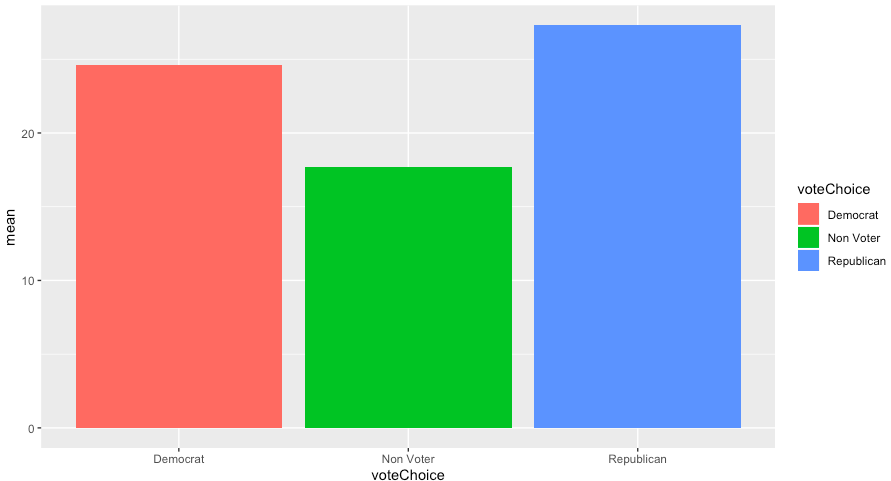
There are notable partisan differences on this metric, with Democrats having the highest mean number of searches (over 400 per person), closely followed by Republicans. There is a notable drop with non-voters having less than 200 search queries on average per person in total.



**QUERY LENGTH** Query text length ranged from a low of 1 character, to a maximum of 322, with a mean character length of 25.7 (and median of 22.12).

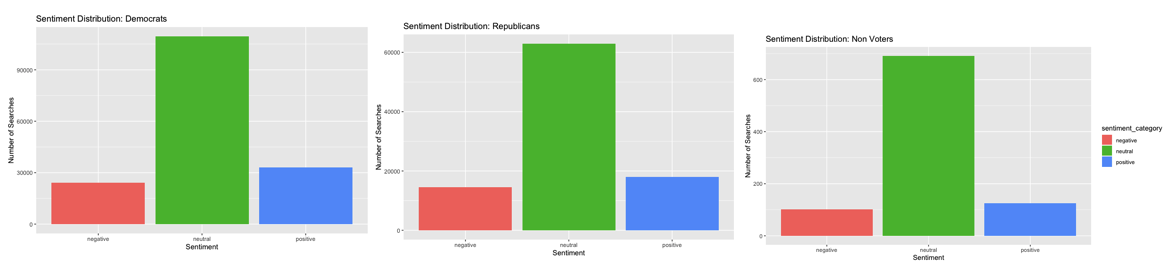


There is a significant left-hand tail, though more minor variation on a partisan basis. Republicans lead Democrats in terms of query length, but not by a considerable amount. Non-voters have the shortest mean query length at under 20 characters.

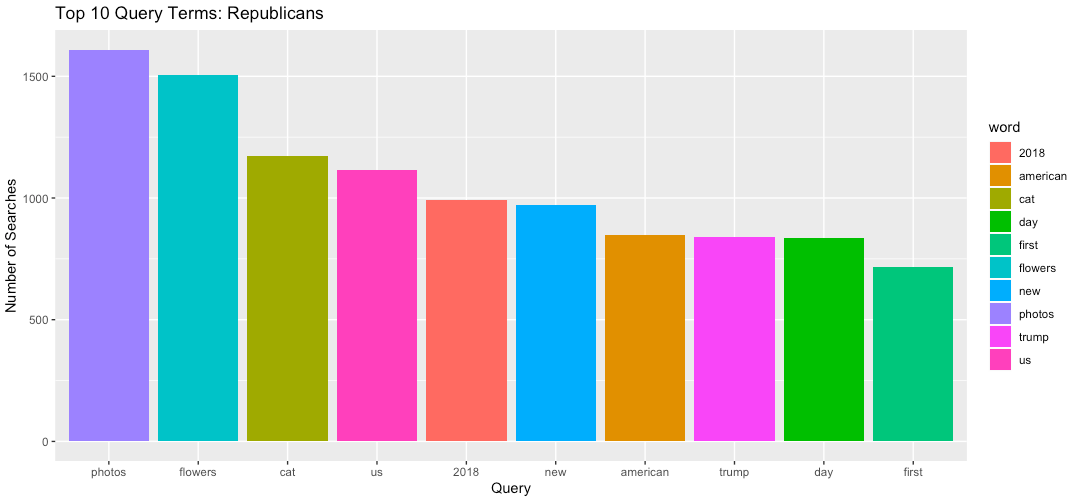
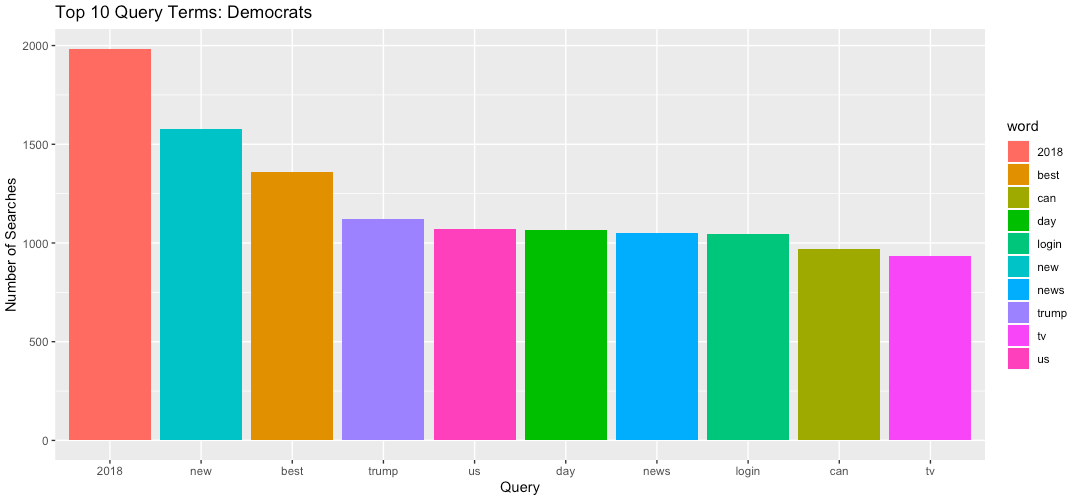


**SENTIMENT** Following recommendations from Lui et al. (2011), the sentiment behind a query text is explored in an attempt to uncover the intention motivating the participant’s query. Several sentiment analysis methods were evaluated, including Bing (B. Lui, n.d.), which utilizes a simple positive-negative structure, the NRC Emotion Lexicon (Mohammad, 2016) that associates vocabulary with eight different emotional keywords, and sentence-level analyses that can account for items such as negation. For example, the phrase “today is *not* a good day” would be evaluated as negative, despite the presence of a positive word, “good.”

However, the findings were not promising from a predictive standpoint from any of the methods employed, as very little variation existed based on ideology or voter status (for details on each method and results, see Appendix C). Therefore, sentiment is not employed as a feature in further models.

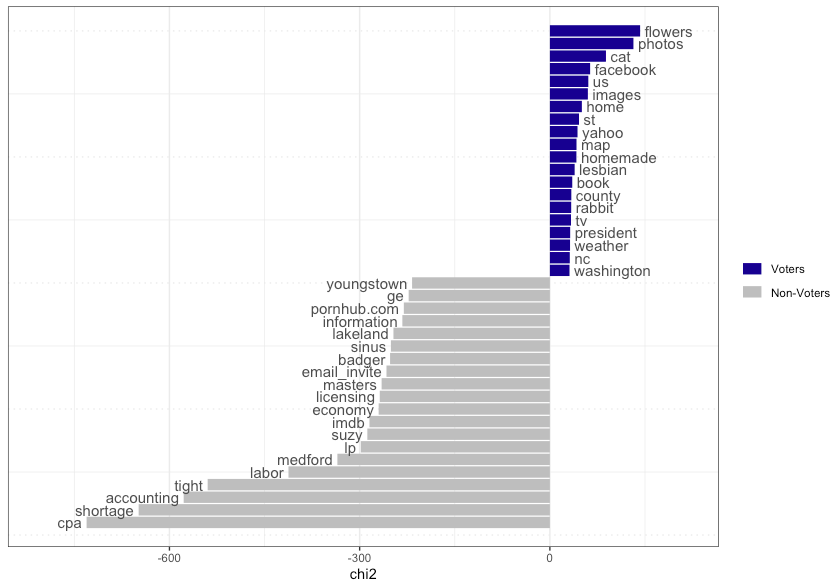


**VOCABULARY** Perhaps unsurprisingly, partisanship has a notable relationship to the top search terms used by an individual, though with some overlap. For example, Democrats and Republicans both have the terms “2018,” “day,” “new,” “trump,” and “us” in their top 10 queries. However, Republicans are much more likely to search for the words “photos,” “flowers,” and “cat,” while the terms “best,” and “news” are more associated with Democrats.

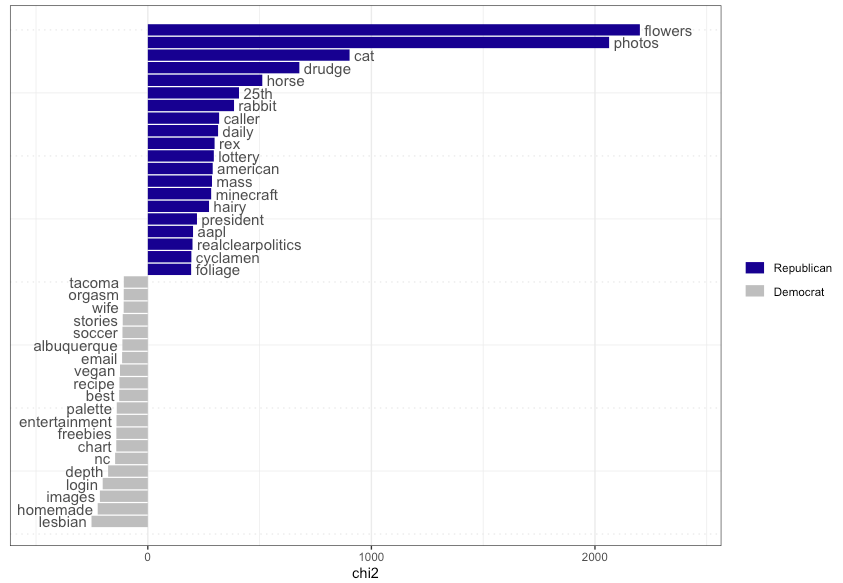


Differences in vocabulary can be more analyzed using *keyness* analysis, which relies on chi-squared to compare the relative frequency of terms between two documents to identify the terms that are most strongly associated with each group.

Running such a model, it is evident that there are clear differences in language used by voters vs. non-voters and between Democrats and Republicans.



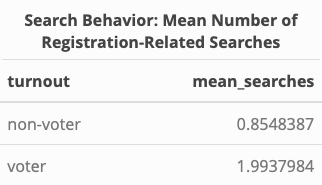
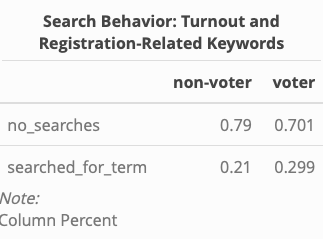
In comparing voters and non-voters, certain terms are very heavily associated with non-voters in this dataset (such as “cpa,” “shortage,” and “accounting”). Terms like “flowers” and “images” are more associated with voters, which is logical given that these words feature heavily in the keyness analysis comparing Democrat and Republican voters.

****

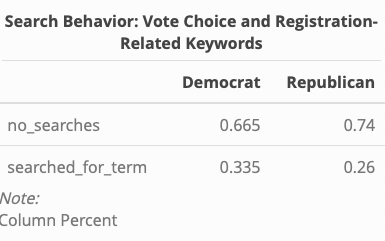
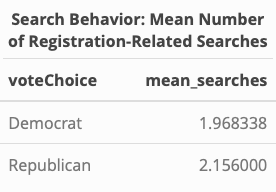
Among voters, keywords associated with Republicans feature more strongly than those associated with the Democrats, though clear and intuitive differences emerge. Republican voters are likely to search for such neutral terms as “flowers” and “cat,” but also search for conservative news outlets like “drudge.” In contrast, Democrats are more associated with the keywords “lesbian,” and “vegan.”

**RELEVANT KEYWORDS** In addition to general search behavior, a major focus of this analysis focuses on the predictive nature of searches for specific, politically-relevant keywords.

*Registration-related keywords* Following the work of Stephens-davidowitz (2013) and Street et al. (2015), voter-registration-related keywords such as “voting” and “register” were analyzed to see if variation exists on voter vs. non-voter and partisan preference bases.[[1]](#footnote-1) Here, notable variation was present:



As is evident in the tables above, approximately 29% of voters made a search for a registration-related term, in comparison to 21% of non-voters. Voters also made over twice as many individual queries containing registration-related keywords than non-voters (averaging 1.99 in contrast to 0.85 per person).



Among voters, the total number of searches was similar for Democrats and Republicans, though a larger proportion of Democrats searched for a registration-related term: 33.5%, compared to 26% of Republicans.

*Candidate Names*

**TOPIC**

Conclusion:

* Frequency of searches and vocabulary matter
* Sentiment doesn’t, at least at the word level

# Methodology

Summary

## Datasets and Model Variations

Summary

## Model Evaluation

### Metrics

Both research questions outlined in this study are binary classification tasks with a notable class imbalance. Therefore, accuracy – despite being the most common metric used for classification tasks – is not a suitable choice on its own. Precision, recall, and F1 scores will be reported to help give a more authentic evaluation of model performance.

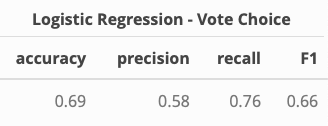
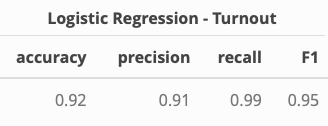
### Baseline Models

In an effort to overcome the shortcomings of many prior studies relying on digital data to predict electoral outcomes, it is important to establish a meaningful baseline model for comparison (Gayo-avello, 2012). Cranmer and Desmarais (2017) argue that rather than a null model, which is incredibly unlikely in social sciences, a baseline model should either reflect the most recently established “state-of-the-literature,” or a benchmark model that does not make use of the theory behind the research question being studied. For example, prior studies have utilized incumbency or traditional polling as a baseline model when studying the predictive power of Google Trends (C. Lui et al., 2011), or simply the majority-label of all users statewide for predicting candidate preference with web-browsing history (Comarela et al., 2018).

**Majority-class Metric**

Given the class imbalance present in both research questions (91-9 for turnout and 60-40 for vote choice), a simple majority-class metric is important to consider when evaluating results. For example, it is necessary that the accuracy exceeds 91% for the question on turnout, given that simply predicting the status of “voter” for all observations would already result in 91% accuracy. Therefore, the majority-class metrics set a baseline bar of 91% accuracy when assessing turnout models, and 60% for the vote choice models.

**Socio-Demographic Logistic Regression Models**

Traditionally, models predicting voting behavior have often relied on socio-demographic characteristics. To evaluate whether search engine behavior can predict voter turnout and vote choice to a meaningful degree, two simple logistic regression models utilizing theoretically-driven socio-demographic characteristics (namely age, race, gender, level of education, religiosity, and marital and employment status) that have been shown to be associated with the outcomes of interest were created (see Appendix D for details). These resulted in predictive accuracy of 92% and an F1 score of 95% for the turnout model, and an accuracy of 69%/F1 score of 66% for the model predicting party choice. 

# Appendix

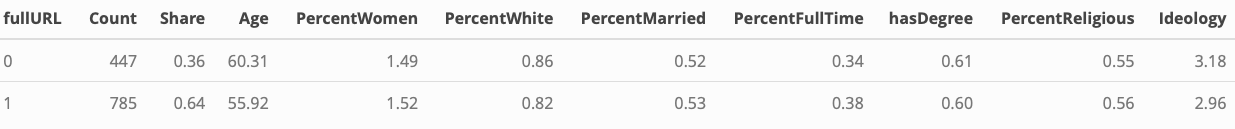
1. **Sample Information**

Respondents were sampled based on YouGov’s methodology that considers both demographic and political targets, and reweighted to more accurately represent the U.S. population based on YouGov’s weights. As YouGov explains, respondents “were weighted according to a sampling frame constructed by stratified sampling from the full 2016 American Community Survey (ACS) 1-year sample with selection within strata by weighted sampling with replacements (using the person weights on the public use file). The sample cases were weighted to the sampling frame using propensity scores. The sample cases and the frame were combined and a logistic regression was estimated for inclusion in the frame. The propensity score function included age, gender, race/ethnicity, years of education, and region. The propensity scores were grouped into deciles of the estimated propensity score in the frame and post-stratified according to these deciles. The weights were then post-stratified on 2016 Presidential vote choice, and a four-way stratification of gender, age (4-categories), race (4- categories), and education (4-categories), to produce the final weight.”

1. **Data Challenges**

For an as-yet unknown reason, some of the participants lacked complete URL information. This means that they were also lacking search query data, and therefore had to be omitted from the analysis. In order to ensure that these participants did not differ in important ways from those included in the project, descriptive sample statistics were taken, and a logistic regression analysis was performed.

Very little difference exists in terms of key socio-demographic indicators. Those without full URLs are approximately 4 years older, slightly more likely to be white, and slightly more conservative than their counterparts with full URL information, but the differences are slight.



The results of the logistic regression were also promising, precisely because they were so poor. Indeed, if the samples do not systematically differ, we would expect demographic variables associated with turnout and vote choice (namely age, race, gender, level of education, religiosity, and marital and employment status) to *not* be able to predict whether or not the participant has full URL information, which is in fact our result. None of the variables were statistically significantly different from their reference category, and accuracy was only 65% (which is right in line with the true proportion of full URLs - 63.71% - indicating the model is no better than a guess).

Odds-ratios Model for Full URLs

===========================

Dependent variable:

---------------------------

fullURL

---------------------------------------------

genderfemale 1.168

(0.840, 1.627)

raceblack 1.276

(0.685, 2.482)

racehispanic 1.774

(0.673, 5.556)

raceasian 3.051

(0.517, 57.940)

raceother 1.061

(0.474, 2.542)

educcollege 1.293

(0.837, 1.987)

educadvanced 1.212

(0.717, 2.050)

marstatmarried 1.034

(0.743, 1.441)

employstudent 3.769

(0.607, 73.686)

employretired 0.929

(0.568, 1.513)

employemployed 1.223

(0.773, 1.923)

religionreligious 1.027

(0.726, 1.448)

agemiddle 0.819

(0.246, 2.364)

ageold 0.645

(0.196, 1.823)

Constant 1.706

(0.548, 6.045)

---------------------------------------------

Observations 672

Log Likelihood -427.684

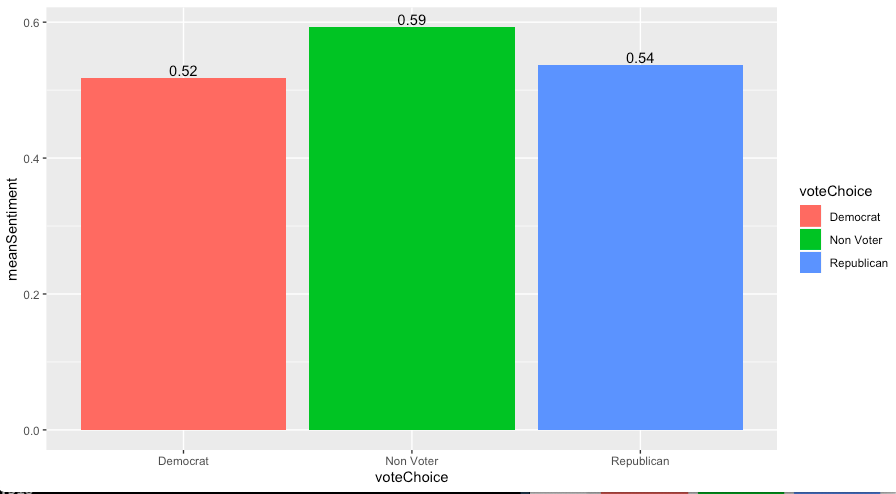
Akaike Inf. Crit. 885.367

=================================

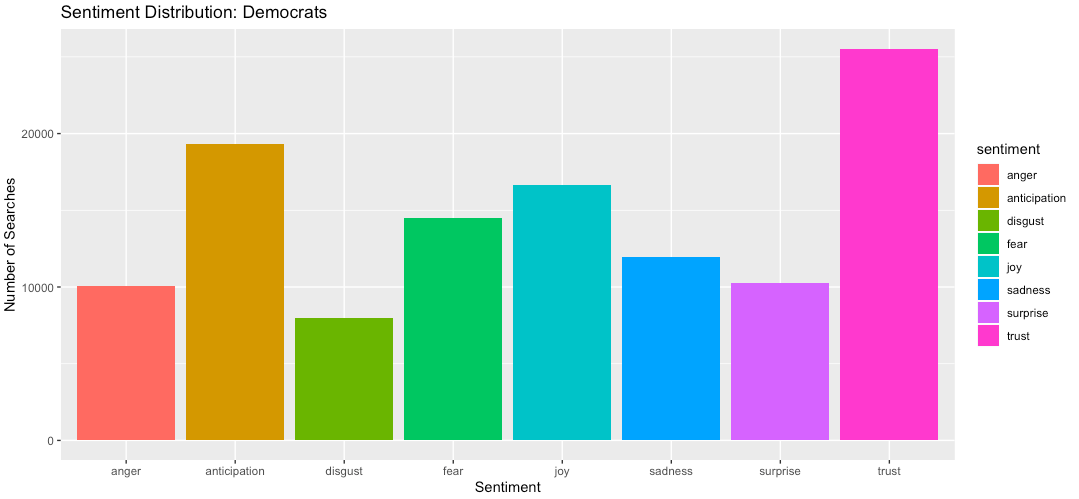
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

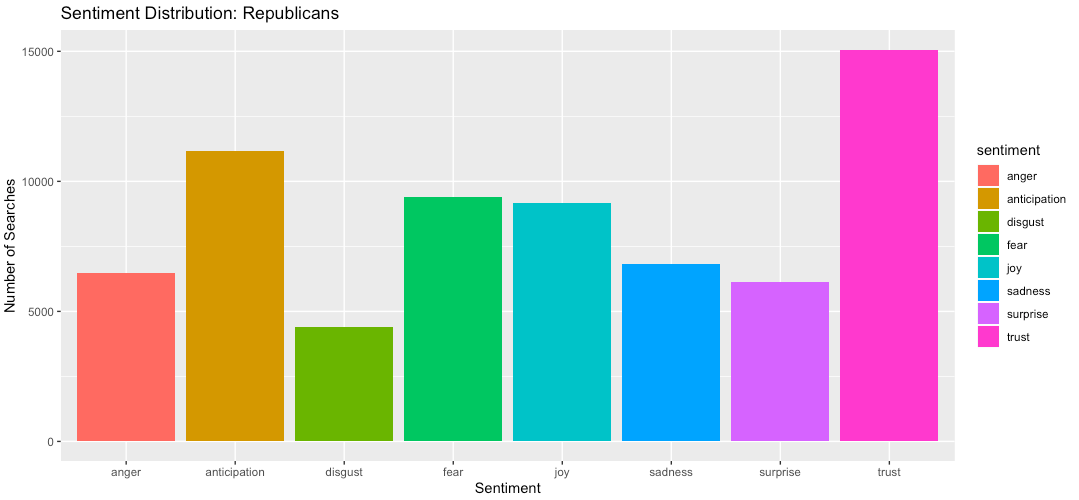
1. **Sentiment Analysis**

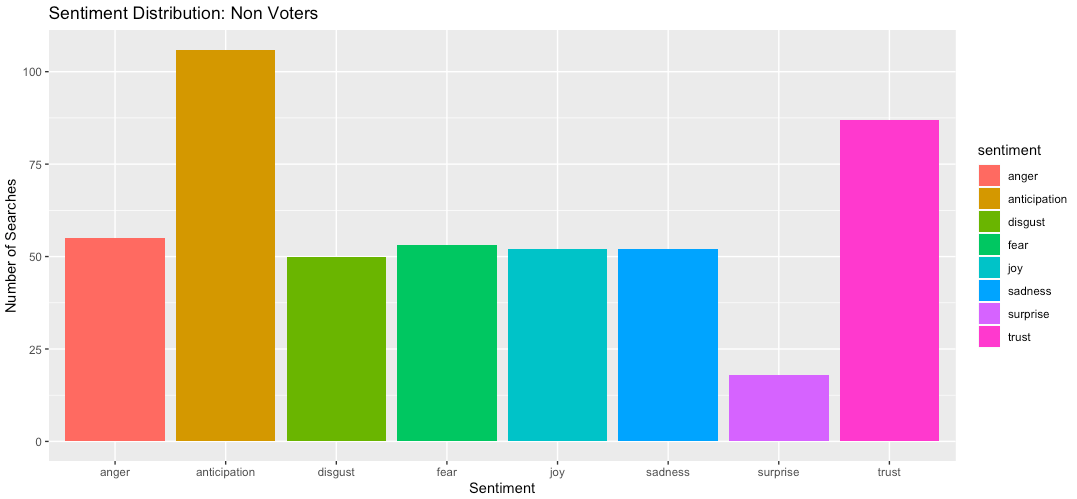
Beyond merely the terms used, the sentiment associated with each word could theoretically be used to gain a fuller understanding of the intention behind a query. To begin this analysis, all users’ queries were analyzed using the Bing sentiment library (B. Lui, n.d.), which simply codes a word as positive or negative. Negative words were then coded 0, and positive words 1, such that higher sentiment scores indicate more positivity. The resulting scores were then averaged for each partisan grouping across all searches. The findings show that only minor variation exists, with non-voters utilizing the most positive query terms on average, followed by Republicans and Democrats at approximately the same.



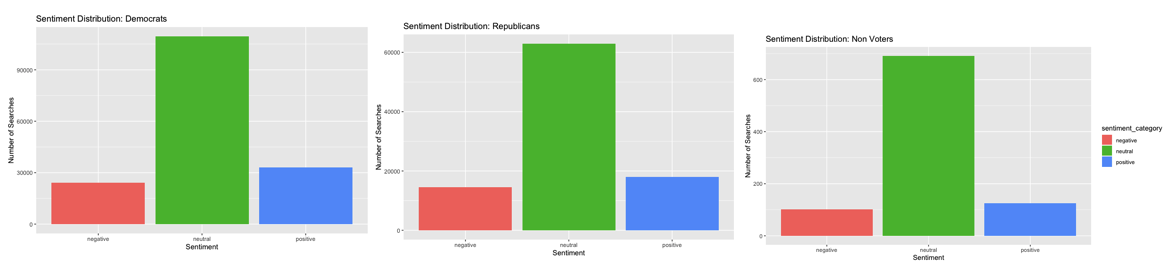
Next, the NRC Emotion Lexicon (Mohammad, 2016), which associates words with eight different emotions (anger, anticipation, disgust, fear, job, sadness, surprise, and trust) was applied to the dataset and analyzed by partisanship. This showed very little variation in emotional distribution, with the exception of non-voters, who used more words associated with anger and disgust, and fewer associated with trust, than their voter-counterparts.



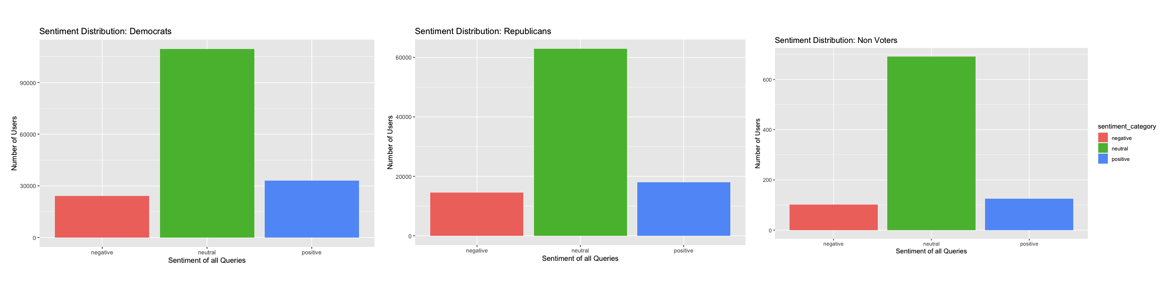




A similar analysis was conducted at the sentence level using the sentiment package, which considers the sentiment more holistically by taking into consideration items such as negation (i.e., today is *not* a good day). This package deemed the majority of the sentence-based queries to be neutral in nature, and also showed very little variation on a partisan basis.



Lastly, the same application was applied to each user’s entire search history as though it were one query, which resulted in similarly unremarkable findings.



Because of the lack of variation, sentiment was deemed to have a low chance of predictive capability, and was therefore not pursued further.

1. **Baseline Socio-Demographic Models**

In order to meaningfully evaluate the performance of the machine learning models based on search engine data in this paper, two logistic regression models relying on socio-demographic variables were created to serve as a baseline comparison. Both relied on the same set of independent variables (age, race, gender, level of education, religiosity, and marital and employment status). For model 1, the dependent variable was turnout (1 – voted, 0 – did not vote), and for model 2, the dependent variable was whether or not they voted for the Republican candidate (0 – Democrat, 1 – Republican).

Model 1 showed that two variables were statistically significantly different from their reference categories: race and marital status. Specifically, Hispanic and Asian Americans had much lower odds of voting than their white counterparts, and married individuals having much higher odds (97.2%) of voting than non-married people.

Model 1: Odds-ratios Model for Turnout

===========================

Dependent variable:

---------------------------

turnout

---------------------------------------------

genderfemale 1.108

(0.544, 2.246)

raceblack 1.245

(0.386, 5.646)

racehispanic 0.362\*

(0.116, 1.287)

raceasian 0.164\*

(0.025, 1.460)

raceother 1.557

(0.278, 29.537)

educcollege 1.861

(0.756, 4.375)

educadvanced 1.381

(0.453, 4.341)

marstatmarried 1.972\*

(0.953, 4.190)

employstudent 4.306

(0.468, 105.742)

employretired 2.646

(0.848, 9.254)

employemployed 1.478

(0.630, 3.382)

religionreligious 1.096

(0.525, 2.220)

agemiddle 1.093

(0.199, 4.638)

ageold 2.393

(0.430, 10.261)

Constant 1.751

(0.351, 11.033)

---------------------------------------------

Observations 442

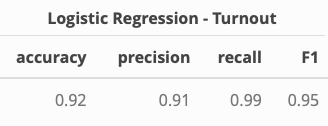
Log Likelihood -121.310

Akaike Inf. Crit. 272.621

================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

In terms of predictive capability, this simple model achieved accuracy of 92%, precision of 91%, recall of 99%, and an F1 score of 95%.



Model 2, which aimed to predict party choice, was not quite as successful in terms of classification, though many more variables showed a statistically significant relationship relative to their reference categories. In particular, women had about 50% lower odds of voting Republican than men, black Americans had about 94% lower odds of voting Republican than white Americans, and the higher educated categories were also less likely to vote Republican. Married individuals had 103% higher odds of voting Republican than unmarried people, and religious individuals had 462% higher odds of voting Republican than the non-religious. Interestingly, young people (between 18-30) had higher odds of voting Republican than both middle aged (30 – 49) and old people (50+) in this sample.

Model 2: Odds-ratio Model Party Choice

=============================================

Dependent variable:

---------------------------

voteChoice

---------------------------------------------

genderfemale 0.499\*\*\*

(0.308, 0.802)

raceblack 0.062\*\*\*

(0.009, 0.230)

racehispanic 1.099

(0.252, 4.479)

raceasian 3.083

(0.644, 15.117)

raceother 0.975

(0.241, 3.843)

educcollege 0.468\*\*

(0.224, 0.955)

educadvanced 0.284\*\*\*

(0.120, 0.651)

marstatmarried 2.032\*\*\*

(1.257, 3.313)

employstudent 13.201

(0.425, 551.043)

employretired 0.769

(0.372, 1.584)

employemployed 1.255

(0.642, 2.470)

religionreligious 5.621\*\*\*

(3.330, 9.788)

agemiddle 63.641\*

(2.580, 6,066.221)

ageold 86.059\*\*

(3.470, 8,457.365)

Constant 0.006\*\*

(0.0001, 0.168)

---------------------------------------------

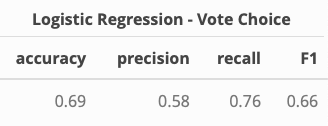
Observations 393

Log Likelihood -214.025

Akaike Inf. Crit. 458.050

=============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01



This model was able to achieve accuracy of 69%, precision of 58%, recall of 76%, and an F1 score of 66%.

1. See the Data Sets section for the complete keyword list. [↑](#footnote-ref-1)