Hertie School

Master's Thesis

# PREDICTING VOTING BEHAVIOR FROM SEARCH ENGINE QUERIES

Authored by

**Marina Wyss**

Supervised by

**Dr. Simon Munzert**

April 2020

# Table of Contents

**Tables and Figures**

## Executive Summary

Digital data are increasingly being used to target individuals for both commercial and political means, and to augment traditional forecasting models for everything from elections to the stock market and the spread of disease. Much work has been done on the topic of predicting political preferences and outcomes from sources such as social media or geographically-aggregated Google Trends data. This research builds on prior studies by utilizing the individual-level search query and survey data of 708 Americans from April 2018 through the 2018 U.S. Midterm Elections to predict whether or not an individual voted, and if so, for the Democrats or Republicans.

The entirety of participants' query histories are included in the data, allowing for the analysis of not only the actual language used, but also search behavior metrics (like the time of day) and characteristics such as the sentiment of a search. Additionally, unlike most prior studies that rely exclusively on data from Google Trends, this research includes queries from other engines such as Bing.com and DuckDuckGo.com, which is relevant as search engine user-bases differ in significant ways.

A variety of text analysis methods are employed, such as sentiment analysis, bag-of-words models of relevant political keywords, and a participant's entire search query history in the form of state-of-the-art word embeddings using BERT. Several machine learning models are tested, including neural networks, gradient boosting machines, support vector machines, and (regularized) logistic regression. The results are compared to theoretically-driven socio-demographic baseline models in order to ensure a robust evaluation of model performance.

The results from this study indicate that turnout may be easier to predict than party choice, though neither research question showed extremely impressive results. One configuration of model and dataset – a BERT representation of a participant's complete query history where certain political keywords were present, modeled by a neural network – was able beat the baseline model in terms of predicting turnout status. However, none of the configurations tested were able to match the baseline model when predicting party choice.

While these results may be a positive indication for those concerned about digital privacy, the methods employed do show promise, and with a larger dataset and more computational power it is quite possible that future research could improve upon the predictive capacity of similar search query data. Interpretability remains a challenge, though, as the best-performing models were those that are particularly difficult to explain or confirm on a theoretical basis, something that will need to be actively addressed should such data be used to make predictions on a wider scale in the future.

Code for this project can be found at https://github.com/MarinaWyss/search-engine-thesis

## 1. 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), potentially creating increased 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 micro-targeting and enhanced forms of digital communication can actually 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 recent protests in Hong Kong (Howard & Duffy, 2011; Shao, 2019).

However, many argue that while there are clearly benefits of data-based electioneering for grassroots movements, the technological platforms that enable them are controlled by an elite class, who may not always have the best interests of 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 inherent complexity surrounding these issues 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 the 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, as in 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 broader text analysis methods in an attempt to uncover the intention behind a search, not just the words used. A variety of machine learning models are utilized, with modest success in predicting whether or not someone voted, but low predictive performance when considering which party a voter chose.

## 2. Literature Review

To place this research in context, an 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 turnout status and political party preference is defined, followed by a review of prior studies that have successfully used such data to achieve these means. Finally, the promise of search engine data in particular is discussed.

### 2.1 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 elections, featuring prominently in the 2016 Trump and Clinton campaigns (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 at 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 of microtargeting 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 similarly 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).

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.

### 2.2 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 turn 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). 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 second search term (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).

### 2.3 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 politicians' 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 that it was not possible to create 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 of Twitter data 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).

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, 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).

### 2.4 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.

## 3. Data

### 3.1 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 January 2019 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 whether the responded voted in the 2018 midterm election, and if they voted, for which party.

In addition to the panel survey data, web tracking data (which includes the text of 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).

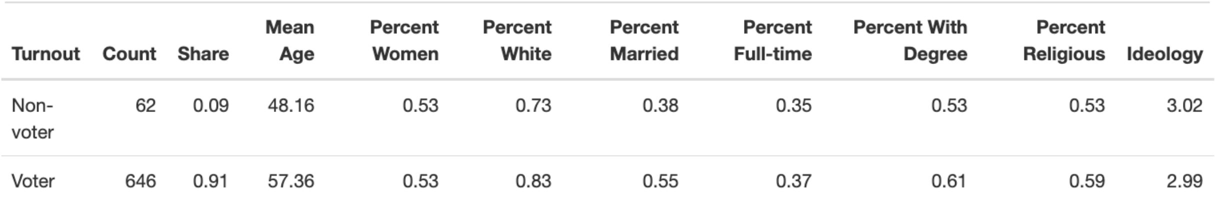
An important note is that the data face one major challenge: Only a subset of the participants had full URL information available, and thus information on the search query text used. This led to a significant decrease in the available sample size. Luckily, those with and without full URL information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

### 3.2 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 the participant 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 turnout 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 does rely on self-reporting, a necessary limitation of the setup.

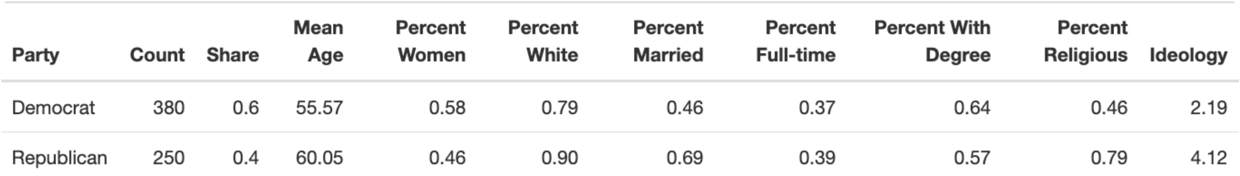
*Table 1: Descriptive Statistics by Turnout*



| Turnout | Count | Share | Mean Age | Percent Women | Percent White | Percent Married | Percent Full-time | Percent With Degree | Percent Religious | Ideology |
|---|---|---|---|---|---|---|---|---|---|---|
| Non-voter | 62 | 0.09 | 48.16 | 0.53 | 0.73 | 0.38 | 0.35 | 0.53 | 0.53 | 3.02 |
| Voter | 646 | 0.91 | 57.36 | 0.53 | 0.83 | 0.55 | 0.37 | 0.61 | 0.59 | 2.99 |

In terms of other demographic differences, it is notable that voters are 9.2 years older than non-voters, on average. 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, while the other characteristics do not vary much between groups.

*Table 2: Descriptive Statistics by Party Choice*



| Party | Count | Share | Mean Age | Percent Women | Percent White | Percent Married | Percent Full-time | Percent With Degree | Percent Religious | Ideology |
|---|---|---|---|---|---|---|---|---|---|---|
| Democrat | 380 | 0.6 | 55.57 | 0.58 | 0.79 | 0.46 | 0.37 | 0.64 | 0.46 | 2.19 |
| Republican | 250 | 0.4 | 60.05 | 0.46 | 0.90 | 0.69 | 0.39 | 0.57 | 0.79 | 4.12 |

Taking a look at only those who reported voting 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. The other 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 ENGINE** One advantage of this research in comparison to prior studies is the ability to leverage query data from search engines 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). 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.

*Table 3: Descriptive Statistics by Search Engine*

| Search Engine | Count | Share | Mean Age | Percent Women | Percent White | Percent Married | Percent Full-time | Percent With Degree | Percent Religious | Ideology |
|---|---|---|---|---|---|---|---|---|---|---|
| Bing | 82 | 0.12 | 53.32 | 0.52 | 0.75 | 0.48 | 0.32 | 0.55 | 0.62 | 3.01 |
| DuckDuckGo | 20 | 0.03 | 65.72 | 0.44 | 0.83 | 0.61 | 0.17 | 0.67 | 0.55 | 3.39 |
| Google | 535 | 0.76 | 55.88 | 0.53 | 0.83 | 0.54 | 0.39 | 0.62 | 0.57 | 2.93 |
| Other | 26 | 0.04 | 66.04 | 0.60 | 0.84 | 0.64 | 0.24 | 0.60 | 0.81 | 3.24 |
| Yahoo | 45 | 0.06 | 61.85 | 0.50 | 0.90 | 0.52 | 0.38 | 0.40 | 0.64 | 3.48 |

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.

**TIME OF DAY** It is reasonable to assume that demographic characteristics may be associated with the time an individual is online or able to make search queries (for example, stay-at-home mothers may be more likely to make search queries on a personal device during the workday). Therefore, the average time of day that a user typically made a query was analyzed. While Democrats and Republicans had a similar mean query time of around 1:17 and 1:08pm, respectively, non-voters had a notably later average time than voters: 2:26pm, versus 1:15pm for voters.

**NUMBER OF QUERIES** Throughout the entirety of this project, only the first sequential search query is considered. For example, if someone input the same query multiple times on the same day, 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 skew (demonstrated by a median of 93.5). This ranged from only 1 search total to 8,139. 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 considerable drop for non-voters, who have less than 200 search queries on average per person in total.

*Figure 1: Mean Number of Queries by Partisanship*

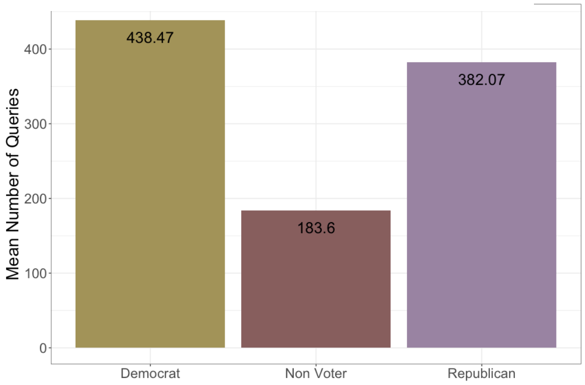
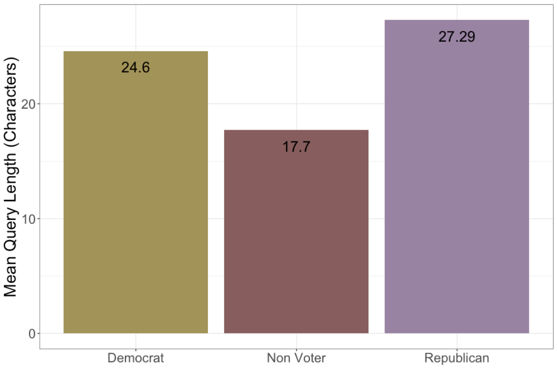


*Figure 2: Mean Query Length by Partisanship*

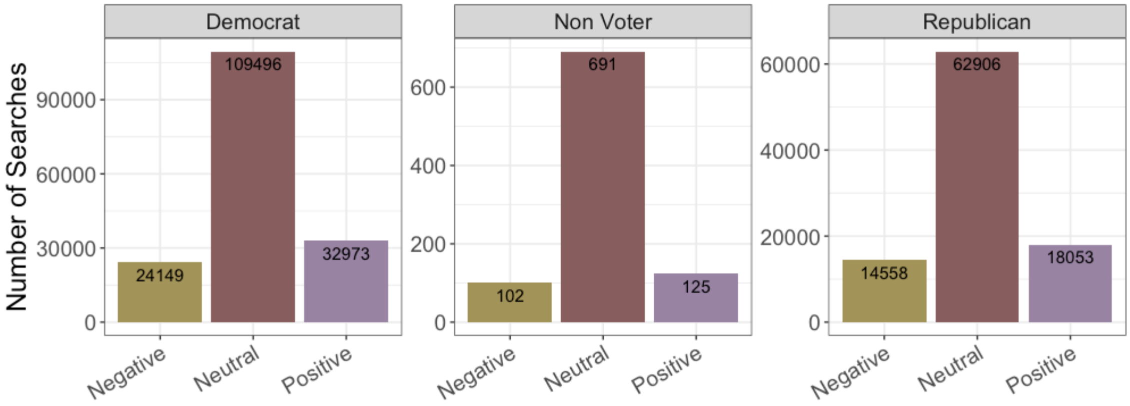


**QUERY LENGTH** Query text length ranged from a low of 1 character to a maximum of 322 characters, with a mean character length of 25.7 (and median of 22.12). There is a significant positive skew, though more minor variation on a partisan basis. Republicans lead Democrats in terms of query length, but not by a considerable amount. Non-voters have the shortest mean query length at just 17.7 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 behavior. 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), which 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."). The later method was employed throughout the study (for details on each method and results, see Appendix D).

Overall, 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.

*Figure 3: Sentiment Distribution by Partisanship*



**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 "american," while the terms "best," and "news" are more associated with Democrats.

*Figure 4: Top 10 Query Terms (Democrats)*

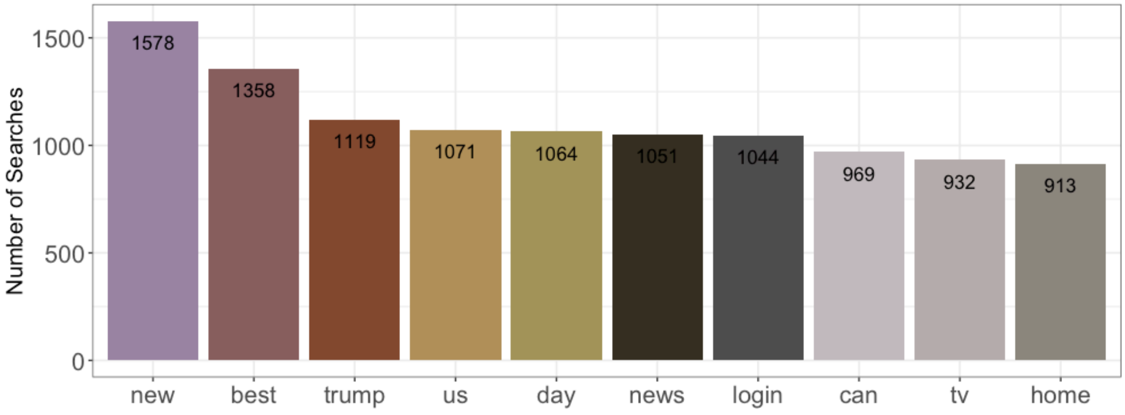
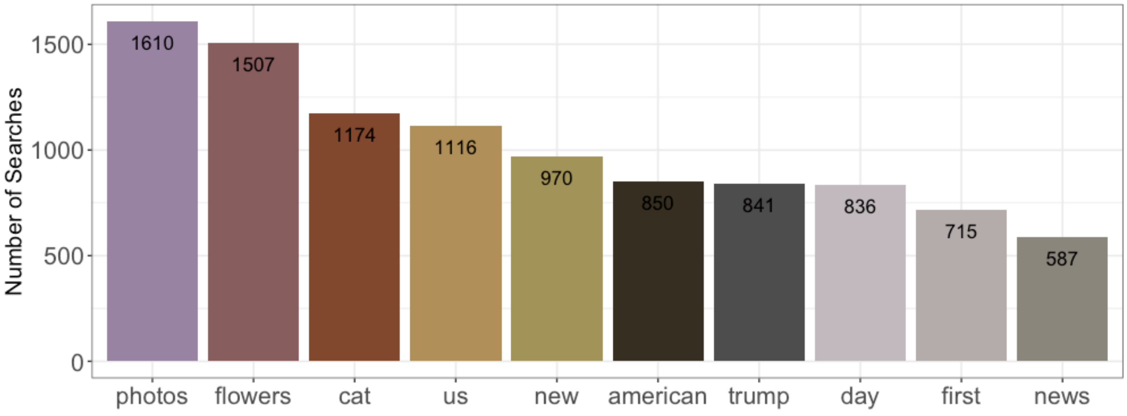


*Figure 5: Top 10 Query Terms (Republicans)*

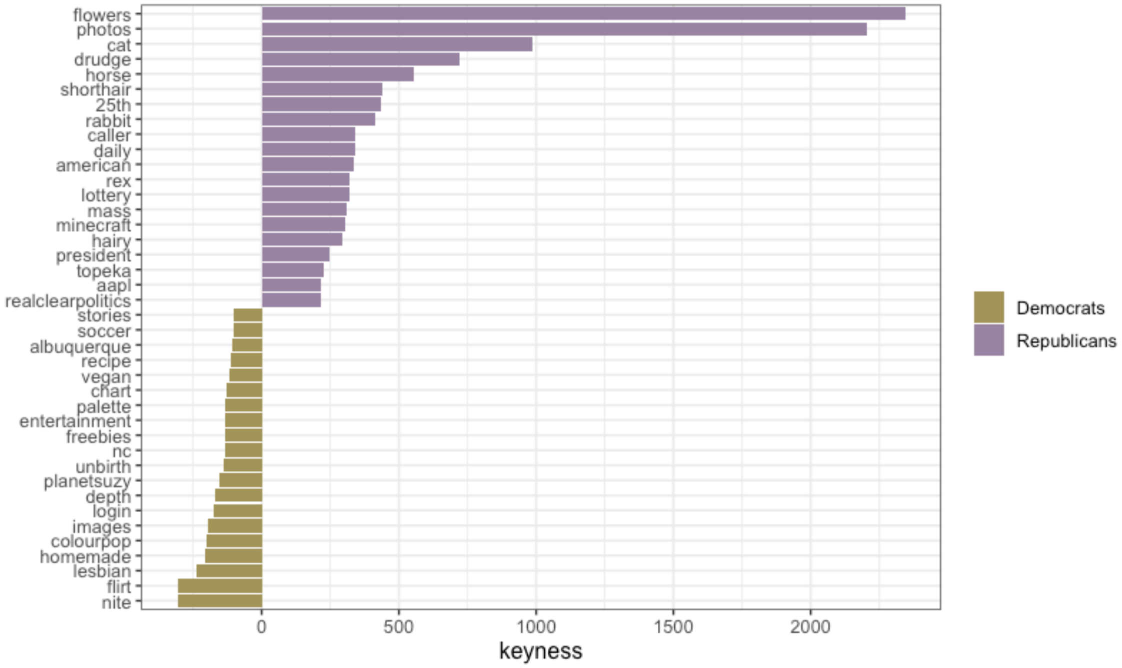


Differences in vocabulary can be further analyzed based on *keyness*, which relies on chi-squared to compare the relative frequency of terms between two documents to identify the terms that are most strongly associated with each group.

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.

When 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.

*Figure 6: Keyness of Vocabulary by Partisanship*



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,” as noted in their top search terms, but also search for conservative news outlets like “drudge,” and “daily” “caller.” 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 is the predictive potential of searches for specific, politically-relevant keywords. For the complete lists of keywords, see Appendix C.

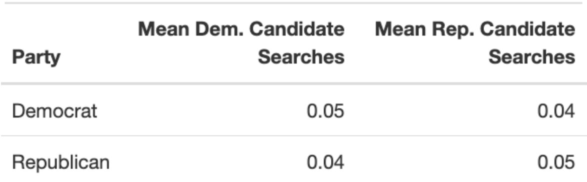
*Registration-related keywords* Following the work of Stephens-davidowitz (2013) and Street et al. (2015), voter-registration-related keywords such as “voting” and “register” were evaluated to see if variation exists on a voter vs. non-voter and partisan preference basis. Here, notable variation was present: Approximately 33% 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 2.49 in contrast to 0.92 per person).

*Table 4: Registration Searches by Turnout*

| Turnout | Mean Number of Searches | Proportion Making Search |
|---|---|---|
| Non-voter | 0.92 | 0.21 |
| Voter | 2.49 | 0.33 |

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: 36.7%, compared to 28% of Republicans.

*Table 5: Candidate Searches by Partisanship*



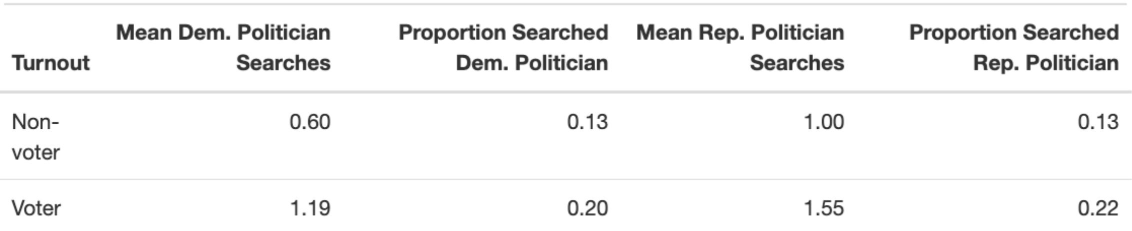
| Party | Mean Dem. Candidate Searches | Mean Rep. Candidate Searches |
|---|---|---|
| Democrat | 0.05 | 0.04 |
| Republican | 0.04 | 0.05 |

*Politicians' Names* According to Lui et al. (2011), the behavior of searching for candidate names during an election year may be predictive of party choice. Interestingly, searching for a candidate can either bode well or poorly for that candidate. They point out that one may search for a candidate they know little about (which may point to an advantage for the candidate that was not searched for). Alternatively, if only one candidate is searched for, this may mean that their opponent has not even gained enough traction to be known at all. Lastly, it is highly likely that individuals will encounter negative information about the candidate as a result of their search, which would also potentially point to a disadvantage at higher search levels.

To evaluate the impact of such searches, the number of queries for the House of Representatives candidates in each participant's state was tallied. Unfortunately, given the small dataset size, there were not enough searches at the candidate name level for a meaningful analysis (only 30 total, over all participants and the entire timeframe). Among the searches that were made, little variation exists.

Therefore, the analysis was expanded to include not only candidates in a given state, but rather whether the participant searched for any Democrat or Republican Congressmember or well-known political figures (such as Barack Obama or Al Gore). This widened the sample to 135 participants who searched for a Democrat at least once, and 147 who searched for a Republican at least once.

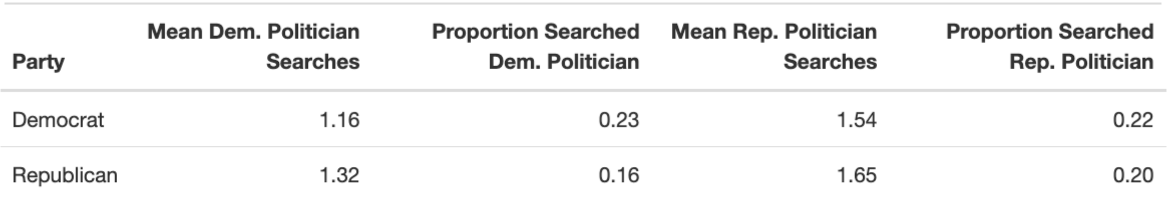
*Table 6: Politician Searches by Turnout*



| Turnout | Mean Dem. Politician Searches | Proportion Searched Dem. Politician | Mean Rep. Politician Searches | Proportion Searched Rep. Politician |
|---|---|---|---|---|
| Non-voter | 0.60 | 0.13 | 1.00 | 0.13 |
| Voter | 1.19 | 0.20 | 1.55 | 0.22 |

Among non-voters, 13% searched for a Democrat and 13% searched for a Republican at least once. There were an average of 0.6 individual searches for a Democrat, and 1 search for a Republican. Voters have higher search volumes than non-voters, with 20% searching for a Democrat at least once, and 22% searching for a Republican at least once. The mean number of searches for a Democrat was 1.19, versus 1.55 for a Republican.

*Table 7: Politician Searches by Partisanship*

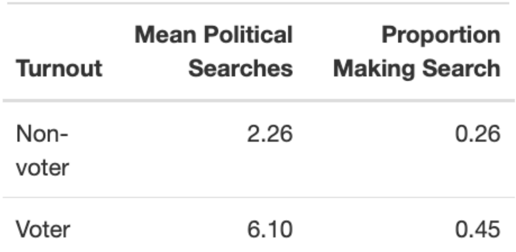


| Party | Mean Dem. Politician Searches | Proportion Searched Dem. Politician | Mean Rep. Politician Searches | Proportion Searched Rep. Politician |
|---|---|---|---|---|
| Democrat | 1.16 | 0.23 | 1.54 | 0.22 |
| Republican | 1.32 | 0.16 | 1.65 | 0.20 |

Turning to those who voted, 23% of Democrats searched for a Democrat at least once, and 22% searched for a Republican. Interestingly, they searched for Republicans more times, on average, with a mean of 1.16 searches for Democrats and 1.55 searches for Republicans. Republican voters also searched for Republican politicians at a higher rate and frequency than Democrat politicians, with 16% of Republicans searching for a Democrat at least once and 20% searching for a Republican, and an average of 1.32 searches for a Democrat compared to 1.65 searches for a Republican.

*General Political Terms* Searches for more generally politically-relevant terms (such as "constituent," or non-partisan offices such as Supreme Court Justices) may be a suitable metric for broader political issue-attention. On this metric there is a notable discrepancy in the mean number of such searches between voters and non-voters, with voters being almost twice as likely as non-voters to make at least one general political search, and average almost three-times as many total. Among voters, little variation exists on this metric, however.

*Table 8: General Political Searches by Turnout*



| Turnout | Mean Political Searches | Proportion Making Search |
|---|---|---|
| Non-voter | 2.26 | 0.26 |
| Voter | 6.10 | 0.45 |

Based on these findings, there does appear to be a credible possibility for search behavior and the actual text searched as perhaps holding predictive power for the two research questions of interest.

## 4. Methodology

### 4.1 Target Variables

The goal of this research is to determine if it is possible to predict 1) whether an individual claimed to have voted and 2) if so, for which party (the Democrats or Republicans), based solely on search query history.

For question 1, regarding turnout, the following survey 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 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 analysis of party choice. 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, resulting in another binary target variable.

### 4.2 Features

**SEARCH BEHAVIOR** As discussed in the literature review, online behavior is known to vary on the basis of socio-demographic characteristics, which is in turn associated with political preferences. Searching for relevant keywords has also shown to have predictive power in some applications. Therefore, participants' search behavior is analyzed in the first dataset.

This includes:

1. The search engine used
2. The average time of day the participant created a query
3. The mean number of queries per day
4. Average query length (in characters)
5. The mean query sentiment

Search query topic was also considered as a feature. Given the short nature of search engine queries, a Gibbs Sampling Dirichlet Mixture Model (GSDMM) as developed by Yin & Wang (2014) and implemented in Python by Ryan Walker (2017) was implemented on the tokenized search query data. GSDMM is an altered LDA method which is designed for short text, given that it assumes that each document is comprised of a single unique topic, in contrast to a mixture of topics in traditional LDA. However, despite several attempts at properly hyperparameter tuning and pre-processing the data, meaningful topics were not able to be uncovered with this unsupervised clustering technique. Given the large number of unique terms in the dataset (over 44,000 once filtered), supervised methods were also not feasible to implement as acquiring labeled data would not have been reasonable to manage.

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), additional features were created based on:

1. Whether or not an individual searched for terms related to registering to vote, and if so, how many times within the period of the study.
2. Whether and how much they searched for the candidate names of those running for the House of Representatives in their state, and if so, the party associated with the candidate (Republican, Democrat, or other).
3. Whether and how much they searched for any partisan political figure, including all members of Congress at that time, recent and current Presidents, Presidential cabinet members, or well-known political figures (including those not serving as of the time period of this study, such as Hillary Clinton); specifically, whether they searched for a Republican or Democrat official.
4. Whether and how much they searched for terms related to politics generally, or non-partisan offices (i.e. "filibuster," or "Sotomayor").

For the complete lists of keywords, see Appendix C.

**QUERY TEXT** The second set of features takes a more open approach to discovering what might be predictive, and conducts analyses on:

1. The top 1000 most-searched for unigrams (out of a total of around 44,000 unique search terms in the dataset).
2. The top 1000 most-searched for bigrams (in an effort to capture names, or phrases such as "White House").
3. A participant's entire search history (unigrams).
4. A participant's entire search history (unigrams), but only those searches which contain a keyword relating to politics, based on the keyword lists used to create the search-behavior features described above.

Each of the query text feature datasets are analyzed individually, as is the search behavior dataset. This results in five total datasets: *Search Behavior, Top 1000 Unigrams, Top 1000 Bigrams, Entire Search Text,* and *Entire Political Search Text.*

### 4.3 Time Frame

The data employed come exclusively from the months before the election (April – early November, 2018). Several models were also run on data from only the week prior to the election. However, in no cases was the predictive power improved, and given the large number of models tested, these results are not reported.

### 4.4 Class Imbalance

Given the class imbalance present both target variables – whether someone voted or not, and if so, for which party – the training datasets were balanced before conducting any analysis. This was done using SMOTE (Synthetic Minority Over-sampling Technique), which essentially uses a *k*-nearest neighbors approach to create synthetic observations of the minority class and simultaneously exclude members of the majority class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

### 4.5 Models Used

Both research questions – whether someone voted or not, and if so, for which party – were tested on all of the following models. The *Search Behavior* dataset was implemented with logistic regression, k-nearest neighbors, gradient boosting (XGBoost), and a neural network. The *Top 1000 Unigrams* and *Top 1000 Bigrams* datasets were explored with regularized logistic regression, a support vector machine, and XGBoost. Lastly, the *Entire Search Text* and *Entire Political Search Text* were modelled using BERT word embeddings and a neural network.

All models rely on supervised learning, and were hyperparameter tuned (as appropriate) using repeated cross-validation before testing on the test set for the final results. Coding was completed in R for all of the models, with the exception of the neural network and BERT embeddings, which was done in Python.

**K-NEAREST NEIGHBORS** The *Search Behavior* dataset was initially evaluated using k-nearest neighbors (KNN). KNN is conceptually straight-forward, and simply aims to classify new observations based on their “similarity” to training observations. It identifies the *k* most-similar observations to the new observation on the basis of a distance measure. It then determines a predicted class for the new observation based on the most-common class of those “nearest neighbors.” KNN is sensitive to feature scale, so the data were scaled before modeling.

**LOGISTIC REGRESSION** Given the binary nature and simple data structure of the *Search Behavior* dataset, a simple logistic regression was implemented.

**REGULARIZED LOGISTIC REGRESSION** Unlike the *Search Behavior* dataset, the *Top 1000 Unigrams* and *Top 1000 Bigrams* datasets rely on document feature matrices, where each term in the dataset is a column, and each row an observation. If the participant searched for a particular term, that column is populated. This results in a large, sparse matrix. Regularization can be an option in such a scenario, particularly when the number of features outweighs the number of observations. The method employed here is regularized logistic regression using a lasso penalty, which performs automatic feature selection by pushing the coefficients of less important features to zero. This was implemented on the *Top 1000 Unigrams* and *Top 1000 Bigrams* datasets using the `glmnet` package in R, which has the added benefit of automatically scaling features before modeling.

**SUPPORT VECTOR MACHINE** The *Top 1000 Unigrams* and *Top 1000 Bigrams* datasets were also evaluated using a support vector machine (SVM), given that SVMs are particularly well-suited to high-dimensional data. In a classification setting, they operate by seeking to find a hyperplane within the feature space that best separates the classes. They are able to accomplish this with the help of the so-called "kernel trick," whereby the feature space is expanded into multiple dimensions, making it easier to accommodate non-linear data and find the best possible decision boundary.

**XGBOOST** XGBoost has gained significant attention in recent years for its impressive predictive performance on both classification and regression tasks. It is an augmented gradient boosting machine that automatically allows for regularization and high levels of efficiency. Gradient boosting generally refers to building an ensemble of "shallow" trees. Each tree, while not very predictive on its own, serves as the baseline for the next tree to build from. This "boosting" allows the next tree in the sequence to learn from the previous trees' mistakes – resulting in a very powerful model that is suitable for both dense and sparse data. Therefore, the *Search Behavior* data set, as well as the *Top 1000 Unigrams* and *Top 1000 Bigrams* datasets were evaluated using XGBoost.

**NEURAL NETWORK** Finally, a basic feed-forward neural network (multi-layer perceptron) was applied to the *Search Behavior*, *Entire Search Text,* and *Entire Political Search Text* datasets. Neural networks are some of the most powerful models available today, and are designed in such a way as to mimic the structure and function of the human brain to help with finding patterns that may be too complex for other machine learning models.

Neural networks are composed of a series of highly interconnected neurons (individual processing units), which are organized into layers. The input layer receives the data, and the output layer predicts which class a new observation belongs to. In between exist one or more "hidden" layers, which is where most of the computation takes place. Neurons of the input layer are connected to the first hidden layer through channels, each of which is assigned a numeric weight. The inputs are multiplied by these weights, and their sum is sent as input to the neurons in the hidden layer. Each of the neurons in the hidden layer also has a numerical value, called the bias, which is added to the sum. This new value is then passed through an activation function, which determines whether the particular neuron is activated or not. If the neuron is activated, it will send data to the neurons of the next layer, over the channels, and so on. Thus, the data is fed through the network via a process called forward propagation. In the output layer, the neuron with the highest value is activated, and determines the output. During training, the model compares the real target value to the predicted values in the output layer, and adjusts the weights accordingly through back propagation, in an iterative process.

Given that neural networks are highly sensitive to feature scaling, the *Search Behavior* dataset was pre-processed accordingly. The *Entire Search Text* and *Entire Political Search Text* datasets had a different form of pre-processing: Rather than passing a document feature matrix of all possible terms over all searches, word embeddings were used to represent the block of query text for each participant.

Word embeddings are a way to numerically represent text such that words with similar meanings have similar representations. They allow for dense representations as real-value vectors, as opposed to sparse one-hot representations like document feature matrices. Many embedding models come pre-trained on large corpora of text, which can further improve embedding quality and training time.

For this project, the BERT (Bidirectional Encoder Representations from Transformers) model was applied. BERT is an open-source, pre-trained NLP model that was developed in 2019 by researchers at Google AI Language, with the key innovation being its bidirectional nature: Unlike previous context-free models like word2vec or GloVe, which generate a single embedding for each word (leading to a lack of contextual differentiation among words with multiple meanings), BERT relies on Transformer, which is an attention mechanism that learns contextual relations between words (or sub-words) in a text. The most basic Transformer includes an encoder that reads text input and a decoder for prediction. BERT relies only on the encoder, which reads the entire sequence of words at once (as opposed to left-to-right or right-to-left), thereby learning the context of the word based on all of its surroundings (again, bidirectional). These innovations have led to BERT demonstrating consistently superior performance to other similar NLP algorithms (Devlin, Chang, Lee, & Toutanova, 2018). The resulting word embedding vectors were used as features in the neural network setup.

### 4.6 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, and F1 scores are also reported to help give a more holistic evaluation of model performance.

**MAJORITY CLASS** Given the class imbalance present in both research questions, 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. This majority-class metrics set a baseline bar of 91% accuracy when assessing turnout models, and 60% for the vote choice models.

**BASELINE MODELS** A common shortcoming of prior studies relying on digital data to predict electoral outcomes is the lack of 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 be 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 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).

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 that have been shown to be associated with the outcomes of interest (namely age, race, gender, level of education, religiosity, and marital and employment status) were created (see Appendix E for details).

The resulting model on turnout managed to achieve predictive accuracy of 93% and an F1 score of 95% on the test set. The model for party choice achieved accuracy of 74%, and an F1 score of 65% on the test set. Given that these metrics are higher than the majority class, they set the bar for evaluating the search-query models' performance.

## 5. Results

### 5.1 Turnout Model Results

Overall, the results from the first research question regarding whether it is possible to predict if a participant reported to have voted based solely on their search history, was modestly promising. Given the high bar for baseline metrics (majority class of 91%, and a socio-demographic model with accuracy of 93% and an F1 score of 95%), it is not surprising that few models were able to compete, though a few do manage that task.

In terms of accuracy, the top three models are the three that utilized the neural network (*Search Behavior, Entire Search Text,* and *Entire Political Search Text*). These three all beat the majority-class metric, and the *Search Behavior* and *Entire Political Search Text* also narrowly beat-out the socio-demographic logistic regression model.

Turning to F1, no configuration was able to meet or beat the baseline socio-demographic model. However, many still performed quite well. In addition to the neural network models, all of the models utilizing XGBoost excelled in this area.

*Figure 7: Turnout Model Comparison*

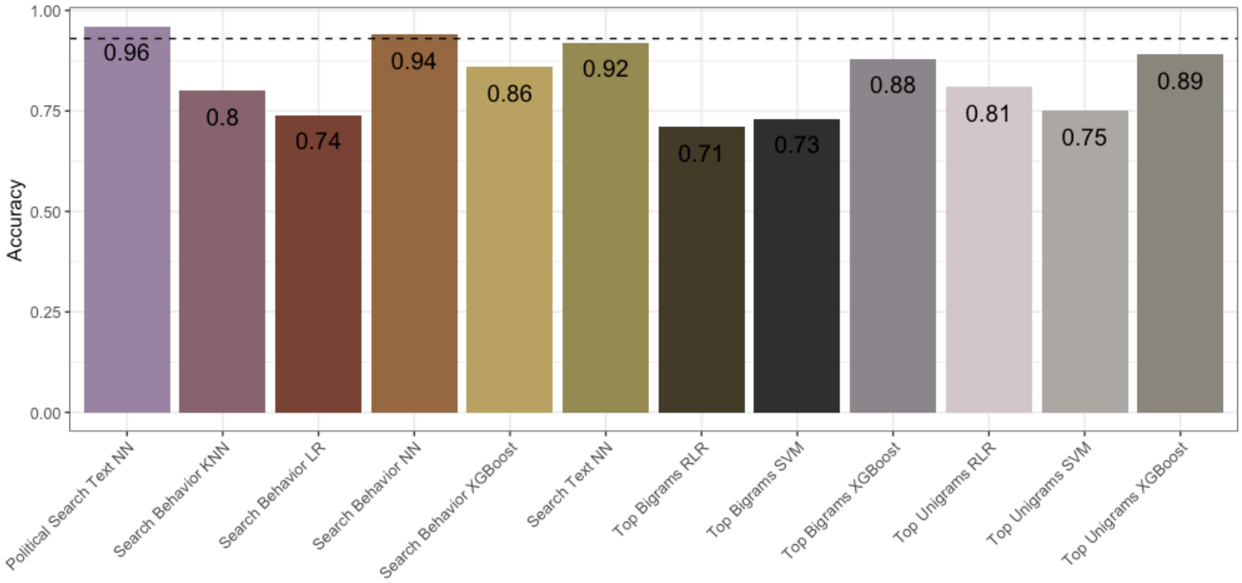
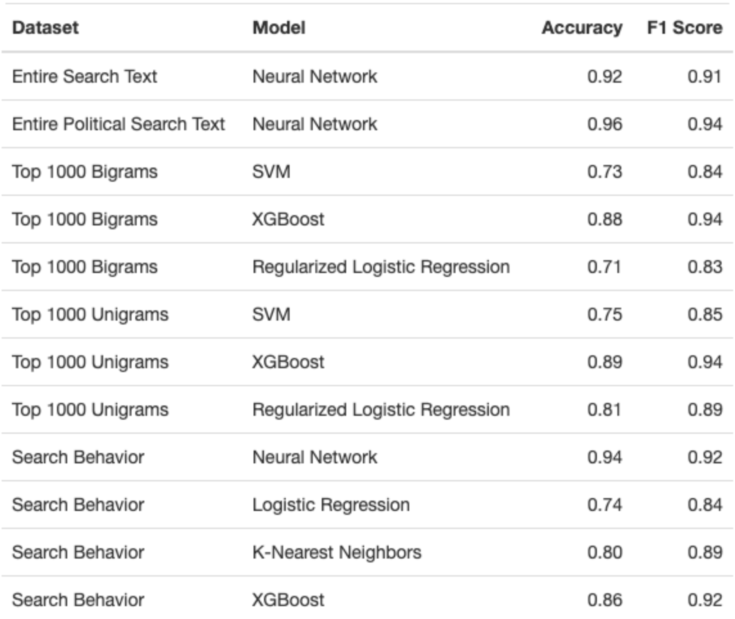


*Table 9: Turnout Model Results*

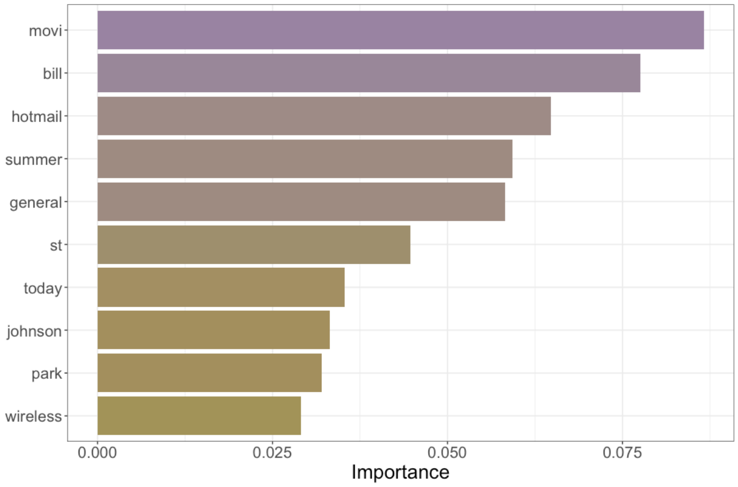


| Dataset | Model | Accuracy | F1 Score |
|---|---|---|---|
| Entire Search Text | Neural Network | 0.92 | 0.91 |
| Entire Political Search Text | Neural Network | 0.96 | 0.94 |
| Top 1000 Bigrams | SVM | 0.73 | 0.84 |
| Top 1000 Bigrams | XGBoost | 0.88 | 0.94 |
| Top 1000 Bigrams | Regularized Logistic Regression | 0.71 | 0.83 |
| Top 1000 Unigrams | SVM | 0.75 | 0.85 |
| Top 1000 Unigrams | XGBoost | 0.89 | 0.94 |
| Top 1000 Unigrams | Regularized Logistic Regression | 0.81 | 0.89 |
| Search Behavior | Neural Network | 0.94 | 0.92 |
| Search Behavior | Logistic Regression | 0.74 | 0.84 |
| Search Behavior | K-Nearest Neighbors | 0.80 | 0.89 |
| Search Behavior | XGBoost | 0.86 | 0.92 |

A key challenge in evaluating these results is that the top performing model – the neural network on the BERT-encoded *Entire Political Search Text* dataset – is not directly interpretable. Not only are neural networks known for being difficult to interpret, common methods would not produce meaningful results on the embeddings. While variable importance models do exist (such as `LIME`, or the `VIP` package in R), the results would be the most-relevant word vectors, which are essentially meaningless: they would return a numerical vector representation of the most relevant words for prediction, but there is not a way to consistently re-translate these vectors back into written language. This is discussed further in the section on Limitations.

We can, however, look more deeply into a model that also performed well, but used the sparse document feature matrix representations of words instead: the *Top 1000 Unigrams* modeled by XGBoost. A variable importance plot (using the `VIP` package in R) shows that, interestingly, it is rather mundane search terms that had the most predictive power in terms of whether or not someone was likely to be a voter.

*Figure 8: Top 1000 Unigrams Variable Importance*



The relative importance of each feature is calculated using a "permutation approach," whereby a feature is randomly permuted to see how that affects the model's error rate. A feature is considered important if the model's error increases relative to when other features are permuted. On the other hand, a feature is unimportant if permuting it makes little difference in the model's predictive ability. In this case, the stems of the search terms "movie," "bill," and "hotmail" seem to be particularly relevant, though establishing why this would be the case theoretically is not straight-forward, and may point to the model overfitting based on "noise."

Digging even further into individual predictions, below, two wordclouds are presented: The first is an example of the top search queries from a non-voter who was incorrectly predicted to be a voter with a high level of confidence. The second is the top searches of a voter who was accurately predicted to be a voter, also with high probability. Here we can see some more intuitive findings: The non-voter has word stems for terms such as America, state, and President, all terms that are logically more likely to be associated with voters. Interestingly, however, the voter who was also predicted with high probability to be a voter does not have many intuitive searches, pointing again to potential predictive power of everyday, a-political search texts.

*Figure 10: Non-Voter, Predicted Voter Wordcloud*

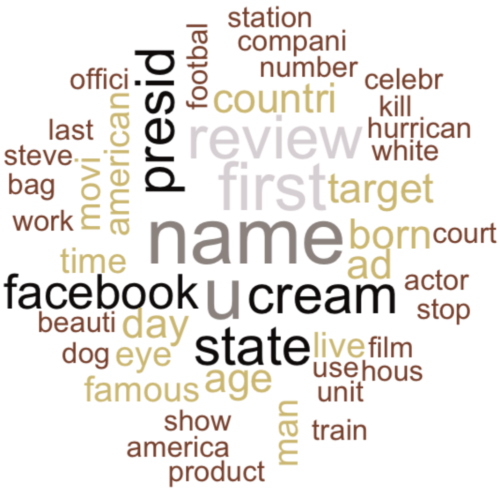
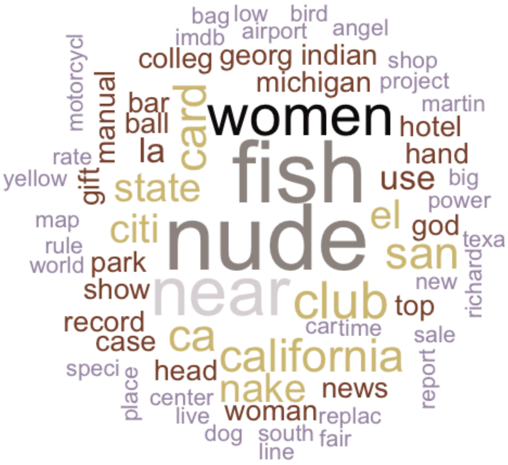


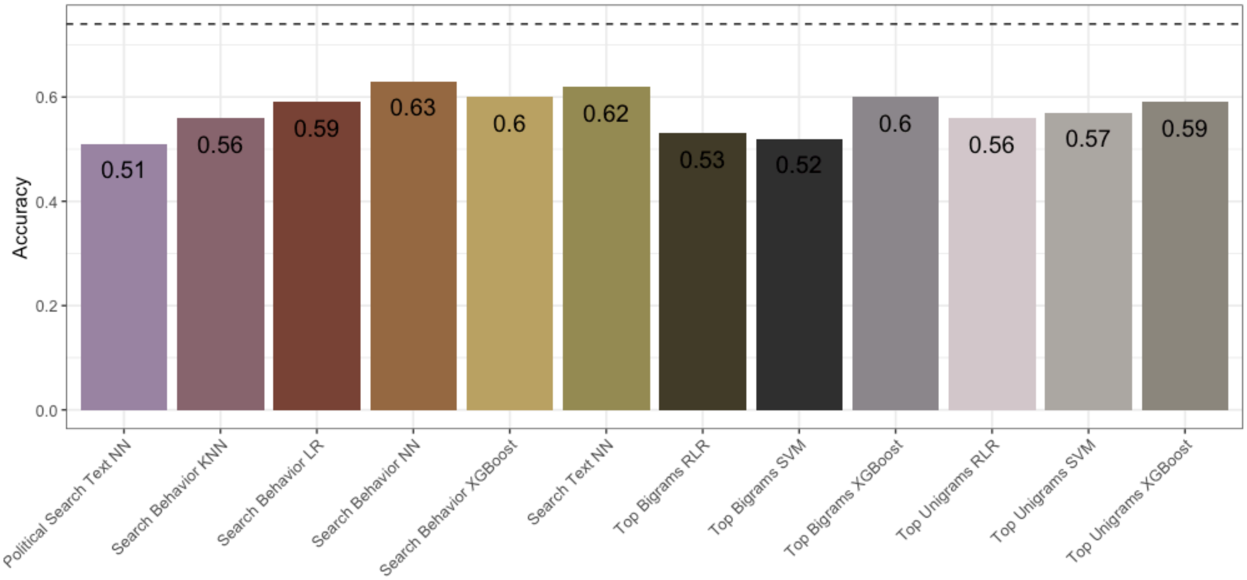
*Figure 9: Voter, Predicted Voter Wordcloud*



### 5.2 Party Choice Model Results

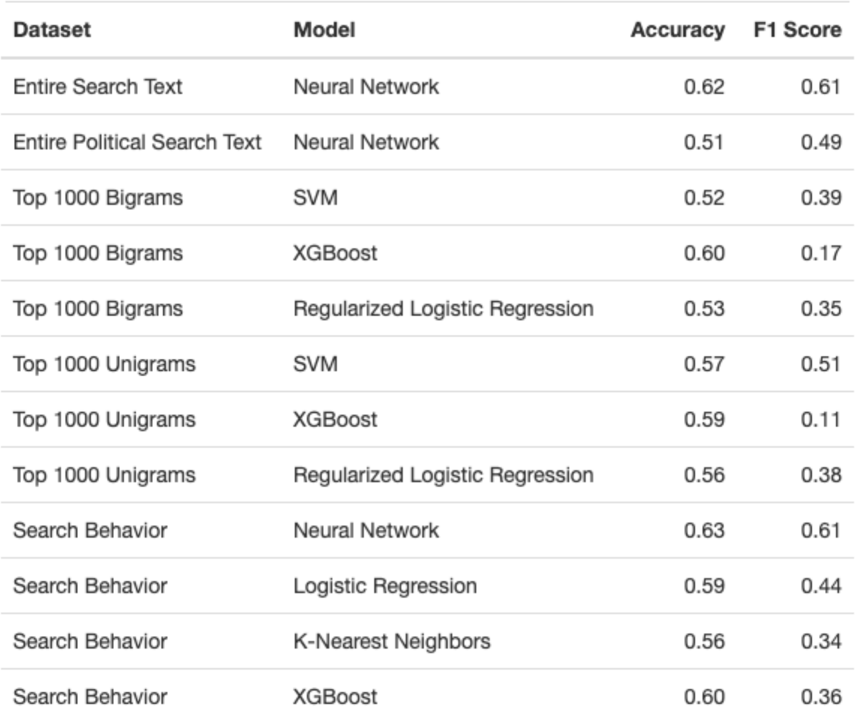
Unlike the first research question, none of the models on party choice reached even close to the level of accuracy of the baseline socio-demographic model. The best configuration – *Search Behavior* utilizing the neural network – achieved an accuracy of 63%, 11% below the baseline model, and only 3% above the majority-class metric.

*Figure 11: Party Choice Model Comparison*



Similar findings are present when evaluating on the basis of F1 score. While two models, the *Search Behavior* and *Entire Search Text* using neural networks, get closer to reaching the baseline of 65%, several models perform abysmally – as low as 11% for *Top 1000 Unigrams* using XGBoost. An interesting note is that this same configuration performed quite well in terms of predicting turnout.

*Table 10: Party Choice Model Results*

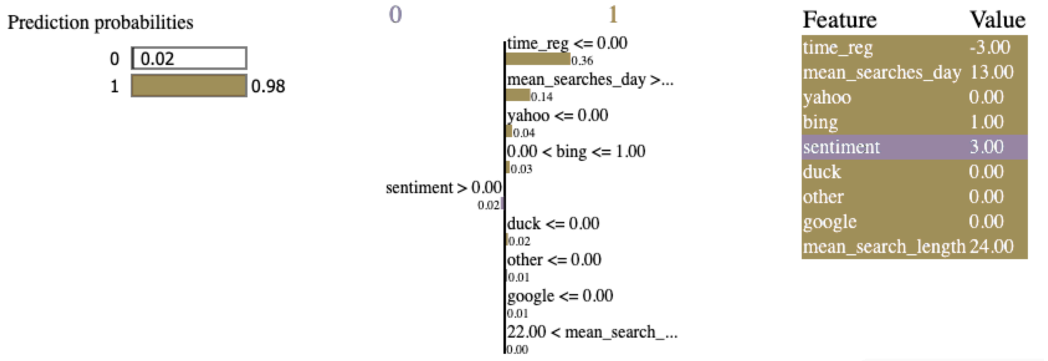


| Dataset | Model | Accuracy | F1 Score |
|---|---|---|---|
| Entire Search Text | Neural Network | 0.62 | 0.61 |
| Entire Political Search Text | Neural Network | 0.51 | 0.49 |
| Top 1000 Bigrams | SVM | 0.52 | 0.39 |
| Top 1000 Bigrams | XGBoost | 0.60 | 0.17 |
| Top 1000 Bigrams | Regularized Logistic Regression | 0.53 | 0.35 |
| Top 1000 Unigrams | SVM | 0.57 | 0.51 |
| Top 1000 Unigrams | XGBoost | 0.59 | 0.11 |
| Top 1000 Unigrams | Regularized Logistic Regression | 0.56 | 0.38 |
| Search Behavior | Neural Network | 0.63 | 0.61 |
| Search Behavior | Logistic Regression | 0.59 | 0.44 |
| Search Behavior | K-Nearest Neighbors | 0.56 | 0.34 |
| Search Behavior | XGBoost | 0.60 | 0.36 |

Overall, the best performing model was *Search Behavior* utilizing the neural network. As described above, decoding the predictions made by neural networks can be challenging. However, the `LIME` package allows for model-agnostic analyses of individual predictions, which may help to shed some light on which features were more relevant during the modeling process. `LIME` (Locally Interpretable Model-agnostic Explanations) assumes that even complex models are linear on a local scale near an observation of interest. It works by fitting a simpler local model (such as a linear model with strong regularization or a decision tree, weighted by proximity of the sampled observations to the instance of interest) around the instance to be explained. Then, small permutations are applied to the original instance to see how this affects the model results on a local level (Ribeiro, Singh, & Guestrin, 2016).

Below is an example of a Democrat participant who was predicted to be a Republican with high probability, based on the *Search Behavior* dataset as modelled by the neural network. `LIME` presents the top 10 most relevant features, their original values, and their influence on the prediction. In this case, being a Bing user, searching quite often per day (13 queries on average) with longer search queries (24 characters), and searching earlier in the day (3 hours earlier than the average search time of 1:15pm) contributed to the Republican prediction. However, having a very positive average sentiment score had a small pull towards Democrat.

*Figure 12: LIME Analysis of Vote Choice Prediction*



In other observations, time of day, how often a participant searched, mean search length, and the search engine used were consistently noted as highly predictive. However, few strict patterns emerged (for example, there was no logical cut-off that was applied consistently as to the number of searches that led to a Republican prediction). This may be due to the nature of classification with neural nets, as the decision patterns may not be as straight-forward as with other models.

## 6. Conclusion

The goal of this research was to uncover whether it is possible to predict if an individual voted, and if so, for the Republicans or Democrats, based solely on their search engine query history. Several machine learning methods were employed in pursuit of this goal, using both feature-engineered metrics on search behavior as well as the actual search query texts. This study builds upon prior research by looking at individual-level data, making use of entire search histories rather than tracking search volumes for particular keywords in a geographic area, incorporating search behavior characteristics such as the time of day and search sentiment, and utilizing data from many different search engines.

On the question of predicting turnout, neural network models using the *Search Behavior* and *Entire Political Search Text* datasets were able to successfully beat a baseline socio-demographic model in terms of accuracy by a small margin, and nearly match F1 scores. However, these models are clear examples of “black box” implementations, with little room for interpretability. When analyzing models that also performed well but did not quite match the baseline metrics, the most important features tended to be mundane keywords, providing little room for theoretical explanation. It seems plausible, therefore, that the higher levels of predictive capability achieved were essentially a matter of chance.

The models on party choice performed even less impressively, with no configuration of dataset and model meeting the baseline level of accuracy set by the socio-demographic model. The best performance was achieved with the *Search Behavior* dataset as modeled by a neural network, but still this fell short of the baseline accuracy by 11%.

There are several areas where this research could have been improved. In particular, a key limitation was the small sample size of only 708 individuals. This sample could also have suffered from selection bias as those who choose to have their online data tracked may differ systematically from the general population. Additionally, panel conditioning effects may have impacted the dependent variables. In general, is quite possible that the dependent variables, particularly the one regarding turnout, were misreported by the participants. This is likely given that 91% claimed to have voted in the 2018 midterm elections, while turnout that year was only 53% (US Census Bureau, 2019).

There were also limitations with the features. In particular, while the keywords used for the *Search Behavior* dataset were collected through a systematic review of Wikipedia, Google Trends, and general political glossaries, their selection was highly subjective. Additionally, no room was made to accommodate for misspellings or nicknames, which could be quite important for the features on searching for individual names in particular.

It is also quite possible that potentially-predictive features were excluded for computational reasons. For example, the *Top 1000 Unigrams* and *Top 1000 Bigrams* were cut-off at 1,000 searches in an effort to balance capturing the most frequently-used, and therefore perhaps important queries, while still being able to run models on a large, sparse document feature matrix.

This challenge was intended to be addressed through the use of word embeddings on a participant’s entire query text, but this process significantly limited the interpretability of the results. Indeed, the most successful models for both research questions utilized embeddings and a neural network. While it is possible to use methods such as `LIME` to identify which features were most important for the neural network in creating predictions, this would simply report the most influential numerical vectors representing the word embeddings. Unfortunately, is not straight-forward to simply “switch back” this vector to the original word. This is because the words themselves form a discrete set of points in the embedding space, and so the model output does not equal the precise location of a particular word. Some potential options would be to use a neural network or autoencoder to “predict” the vector back to its original word, or find the distance between each output vector and the entire corpus vocabulary to pick the word that minimizes that distance. However, none of these solutions were implemented here.

A potential avenue for future research would be to incorporate human labelling to augment the available features. For example, Wojcik and colleagues showed that flu-tracking using search queries can be improved with human labelling of relevant queries (Wojcik et al., 2020). Though labor-intensive, this could help to overcome challenges related to nuance when identifying relevant searches, or potentially creating topics, which was not achievable given the short lengths of the query texts.

While this study did not demonstrate search engine queries as being particularly predictive for voter status and party preference, it is important to note that technological change in the areas of machine learning and deep learning is rapidly progressing. It is quite possible that given a larger sample size and more computational power, future research could achieve highly accurate results. However, given these results and the fact that search query data on the individual level is not typically available at this time, it seems unlikely that search query data will significantly add to the toolbox of those seeking to predict political preferences for the purposes of targeting or forecasting.

## Bibliography

Argamon, S., Koppel, M., Fine, J., Shimoni, A. R., & Science, C. (2003). Gender , Genre , and Writing Style in Formal Written Texts.

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2019). Predicting Voting Behavior Using Digital Trace Data, 1–22. https://doi.org/10.1177/0894439319882896

Brownback, A., & Novotny, A. (2016). Social Desirability Bias and Polling Errors in the 2016 Presidential.

Burger, J. D., & Henderson, J. C. (2006). An Exploration of Observable Features Related to Blogger Age.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, P. W. (2002). SMOTE: Synthetic Minority Over-sampling Technique Nitesh. *Journal of Artificial Intelligence Research 16*, *16*.

Choi, H., & Varian, H. (2011). Predicting the Present with Google Trends.

Chung, J., & Mustafaraj, E. (2010). Can Collective Sentiment Expressed on Twitter Predict Political Elections ?, 1770–1771.

Comarela, G., Barford, P., Christenson, D., & Crovella, M. (2018). Assessing Candidate Preference through Web Browsing History, 158–167.

Conti, G., Point, W., York, N., & Sobiesk, E. (2007). An Honest Man Has Nothing to Fear : User Perceptions on Web-based Information Disclosure.

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling? *Political Analysis*, *25*(2), 145–166. https://doi.org/10.1017/pan.2017.3

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (Mlm).

Dimock, B. Y. M. (2019). An update on our research into trust, facts and democracy.

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, 1–12.

Gayo-avello, D. (2011). Limits of Electoral Predictions Using Twitter, 490–493.

Gayo-avello, D. (2012). A meta-analysis of state-of-the-art electoral prediction from Twitter data.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data, *457*(February 2009). https://doi.org/10.1038/nature07634

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search, *107*(41), 17486–17490. https://doi.org/10.1073/pnas.1005962107

Gurumurthy, A., & Bharthur, D. (2018). DEMOCRACY AND THE ALGORITHMIC TURN, 39–50.

Howard, P. N., & Duffy, A. (2011). What Was the Role of Social Media During the Arab Spring ?, 1–30.

Hu, J., Zeng, H., Li, H., Niu, C., & Chen, Z. (2007). Demographic Prediction Based on User ' s Browsing Behavior, 151–160.

Jamieson, K. H. (2018). *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know*.

Jones, R., & Tomkins, A. (2007). " I Know What You Did Last Summer " — Query Logs and User Privacy.

Koppel, M., Argamon, S., & Gan, R. (2000). Automatically Categorizing Written Texts by Author Gender.

Kruschinski, S. (2017). Restrictions on data-driven political micro- targeting in Germany, *6*(4), 1–23. https://doi.org/10.14763/2017.4.780

Lui, B. (n.d.). Opinion Mining, Sentiment Analysis, Opinion Extraction. Retrieved January 4, 2020, from https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Lui, C., Metaxas, P. T., & Mustafaraj, E. (2011). On the predictability of the U . S . elections through search volume activity.

Mohammad, S. (2016). NRC Emotion Lexicon. Retrieved January 4, 2020, from http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Nowson, S., & Oberlander, J. (2005). The Identity of Bloggers : Openness and gender in personal weblogs.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls : Linking Text Sentiment to Public Opinion Time Series.

Pennacchiotti, M., & Popescu, A. (2010). to Twitter User Classification, 281–288.

Pew Research Center. (2016). The Parties on the Eve of the 2016 Presidential Election: Two Coalitions, Moving Further Apart.

Polykalas, S. E., Prezerakos, G. N., & Konidaris, A. (2013). An Algorithm based on Google Trends ' data for future prediction . Case study : German Elections, 69–73.

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2009). Classifying Latent User Attributes in Twitter.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rogers, T., & Aida, M. (2014). Vote Self-Prediction Hardly Predicts Who Will Vote , and Is ( Misleadingly ) Unbiased. https://doi.org/10.1177/1532673X13496453

Schoen, H., Jungherr, A., & Ju, P. (2012). Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions : A Response to "' Predicting Elections With Twitter : What 140 Characters Reveal About Political Sentiment ,"' *30*(2), 229–234. https://doi.org/10.1177/0894439311404119

Shao, G. (2019). Social media has become a battleground in Hong Kong ' s protests. *CNBC*, pp. 1–7.

Stephens-davidowitz, S. (2012). THE EFFECTS OF RACIAL ANIMUS ON A BLACK PRESIDENTIAL CANDIDATE : USING GOOGLE SEARCH DATA TO UNCOVER WHAT TRADITIONAL SURVEYS MISS ∗.

Stephens-davidowitz, S. (2013). WHO WILL VOTE ? ASK GOOGLE ∗.

Stephens-davidowitz, S. (2017). *Everybody Lies*. Dey Street Books.

Street, A., Murray, T. A., Blitzer, J. B., & Patel, R. S. (2015). Estimating voter registration deadline effects with web search data. *Political Analysis*, *23*(2), 225–241. https://doi.org/10.1093/pan/mpv002

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment, 178–185.

Ur, B., Leon, P. G., Cranor, L. F., Shay, R., & Wang, Y. (2012). Smart , Useful , Scary , Creepy : Perceptions of Online Behavioral Advertising.

US Census Bureau. (2019). Behind the 2018 U.S. Midterm Election Turnout. Retrieved January 10, 2020, from https://www.census.gov/library/stories/2019/04/behind-2018-united-states-midterm-election-turnout.html

Walker, R. (2017). GSDMM: Short text clustering. Retrieved February 1, 2020, from https://github.com/rwalk/gsdmm

Weber, I., & Castillo, C. (2010). The Demographics of Web Search Categories and Subject Descriptors.

Wojcik, S., Bijral, A., Johnston, R., Lavista, J. M., King, G., Kennedy, R., … Lazer, D. (2020). Survey Data and Human Computation for Improved Flu Tracking, 1–10.

Wu, L., & Brynjolfsson, E. (2015). *The Future of Prediction : How Google Searches Foreshadow Housing Prices and Sales The Future of Prediction How Google Searches Foreshadow Housing Prices and Sales*.

Yasseri, T. (2016). Wikipedia traffic data and electoral prediction : towards theoretically informed models, 1–15.

Yin, J., & Wang, J. (2014). A Dirichlet multinomial mixture model-based approach for short text clustering. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–242. https://doi.org/10.1145/2623330.2623715

## Appendix

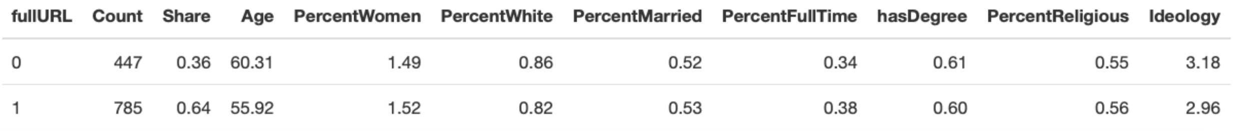
### A. Sample Information

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 weighting scheme. 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."

### B. Data Challenges

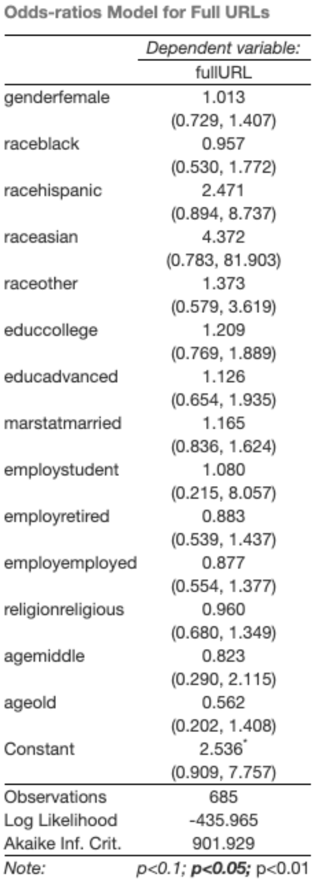
For an as-yet unknown reason, some of the participants lacked complete URL information. This means that they were also lacking search query data, and therefore had to be omitted from the analysis. In order to ensure that these participants did not differ in important ways from those included in the project,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

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



| fullURL | Count | Share | Age | PercentWomen | PercentWhite | PercentMarried | PercentFullTime | hasDegree | PercentReligious | Ideology |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 447 | 0.36 | 60.31 | 1.49 | 0.86 | 0.52 | 0.34 | 0.61 | 0.55 | 3.18 |
| 1 | 785 | 0.64 | 55.92 | 1.52 | 0.82 | 0.53 | 0.38 | 0.60 | 0.56 | 2.96 |

The results of the logistic regression were also promising, precisely because they were actually rather 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 | 1.013 |
| | (0.729, 1.407) |
| raceblack | 0.957 |
| | (0.530, 1.772) |
| racehispanic | 2.471 |
| | (0.894, 8.737) |
| raceasian | 4.372 |
| | (0.783, 81.903) |
| raceother | 1.373 |
| | (0.579, 3.619) |
| educcollege | 1.209 |
| | (0.769, 1.889) |
| educadvanced | 1.126 |
| | (0.654, 1.935) |
| marstatmarried | 1.165 |
| | (0.836, 1.624) |
| employstudent | 1.080 |
| | (0.215, 8.057) |
| employretired | 0.883 |
| | (0.539, 1.437) |
| employemployed | 0.877 |
| | (0.554, 1.377) |
| religionreligious | 0.960 |
| | (0.680, 1.349) |
| agemiddle | 0.823 |
| | (0.290, 2.115) |
| ageold | 0.562 |
| | (0.202, 1.408) |
| Constant | 2.536* |
| | (0.909, 7.757) |
| Observations | 685 |
| Log Likelihood | -435.965 |
| Akaike Inf. Crit. | 901.929 |
| *Note:* | *p<0.1;* ***p<0.05;*** p<0.01 |

## C. Keyword Lists

The following keyword lists were created to engineer the *Search Behavior* dataset features. Each was informed by a combination of Wikipedia, Google Trends, and general political glossaries.

*Registration-related:* absentee, ballot, election, #electionday, midterm, polling, primary, registration, #registertovote, vote, voter, voting.

*General political terms:* alito, amendment, bader, ballot, bipartisan, campaign, candidate, caucus, communism, communist, congress, congressional, congressman, congresswoman, constituency, constituent, constitution, constitutional, convention, delegate, democracy, democrat, democratic, electoral, electorate, federal, filibuster, gerrymander, gerrymandering, gop, Gorsuch, government, governor, gubernatorial, inauguration, incumbent, kagan, kavanaugh, legislation, libertarian, libertarianism, lobbyist, mayor, nomination, nominee, political, politics, poll, president, ratified, referendum, representative, representatives, republican, senate, senator, socialism, socialist, sotomayor, veto.

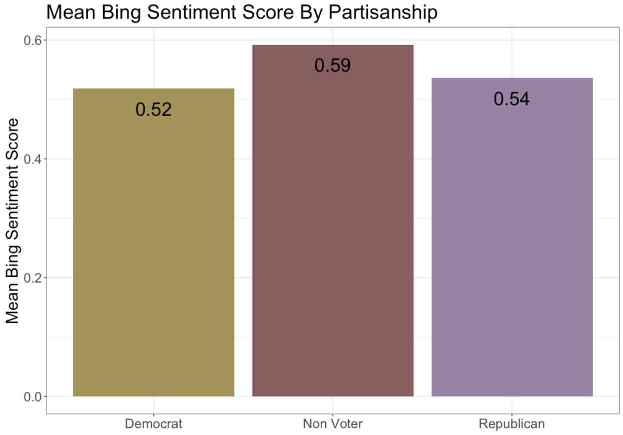
*Candidate info:* All candidate names for the US House of Representatives election in 2018 and which state they ran in.

*Politician names:* All members of Congress in 2018, the Presidential cabinet, as well as well-known figures like Al Gore and Arnold Schwarzenegger.

### D. Sentiment Analysis

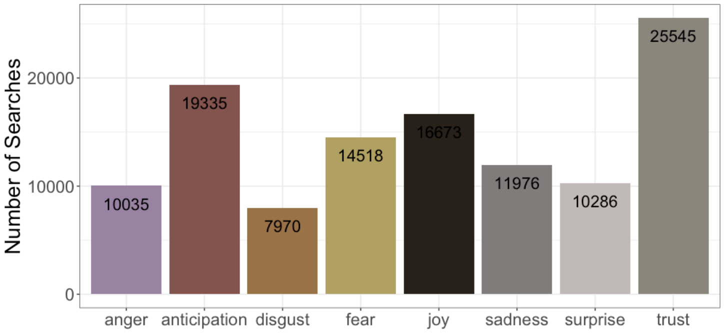
There are several methods available to conduct sentiment-analyses on the word and sentence level. Below are the results of the experiments employed.

1. All users' queries were analyzed using the `Bing` sentiment library (B. Lui, n.d.), which simply codes a word as positive (1) or negative (0), meaning 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 slightly below.

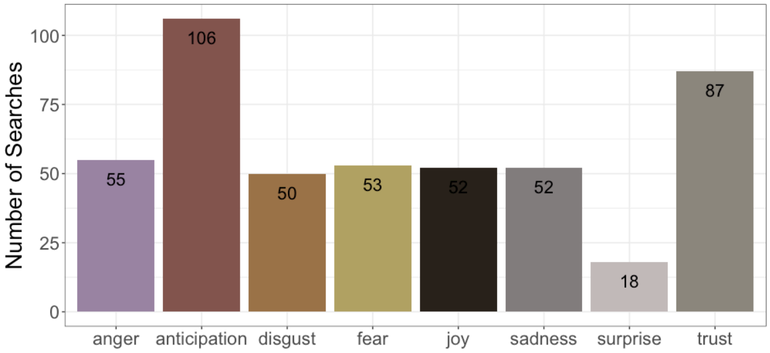


2. Next, the NRC Emotion Lexicon (Mohammad, 2016), which associates words with eight different emotions (anger, anticipation, disgust, fear, job, sadness, surprise, and trust) was applied to the dataset and analyzed by partisanship. This showed slight differences between groups: Democrats were more likely to use terms related to joy, while Republicans were more likely to use words associated with fear. Non-voters used terms with an anticipatory sentiment quite a bit more often, and surprise quite a bit less.

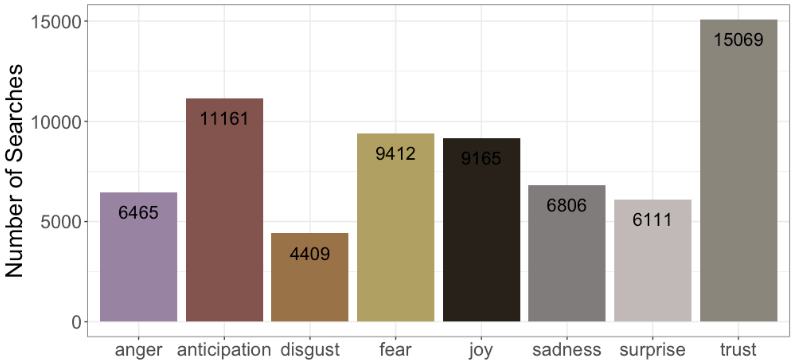
**Democrats:**



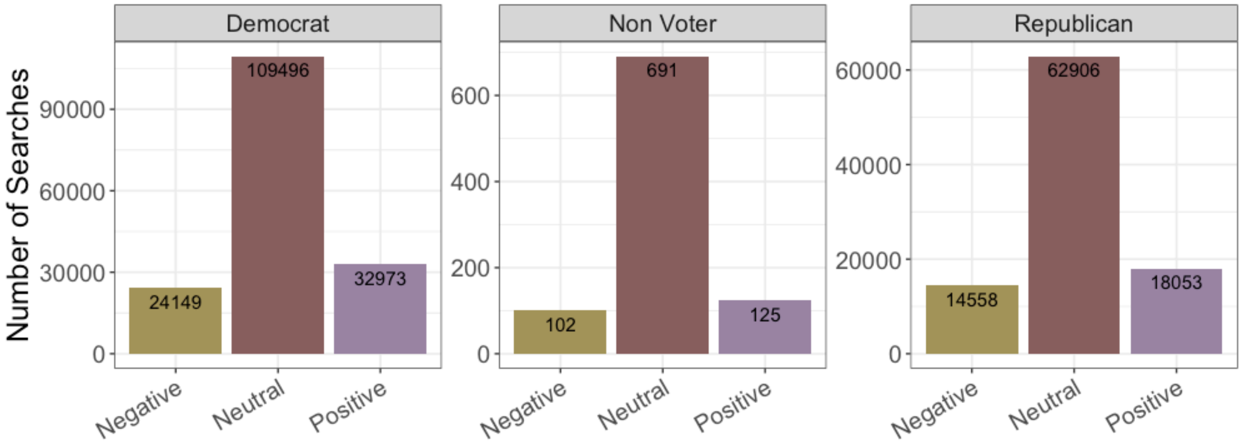
**Non-voters:**



**Republicans:**



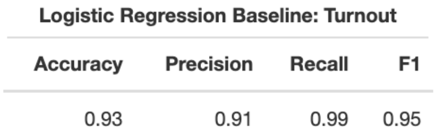
3. Finally, a similar analysis was conducted at the sentence level using the `sentimentr` package, which considers the sentiment (positive, negative, or neutral) more holistically by taking into consideration items such as negation (i.e., today is *not* a good day). This package deemed the majority of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



### E. 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).

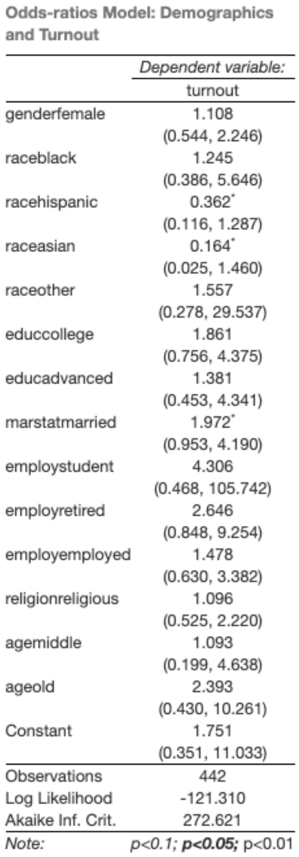
Model 1 showed that two variables were statistically significantly different from their reference categories: race and marital status. Specifically, Hispanic and Asian Americans had much lower odds of voting than their white counterparts, and married individuals having much higher odds (97.2%) of voting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



**Logistic Regression Baseline: Turnout**

| Accuracy | Precision | Recall | F1 |
|---|---|---|---|
| 0.93 | 0.91 | 0.99 | 0.95 |

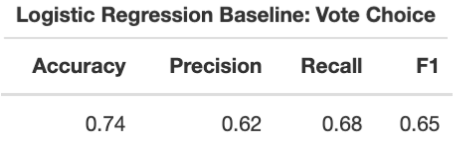
In terms of predictive capability, this simple model achieved accuracy of 93%, precision of 91%, recall of 99%, and an F1 score of 95%.



**Odds-ratios Model: Demographics and 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 |

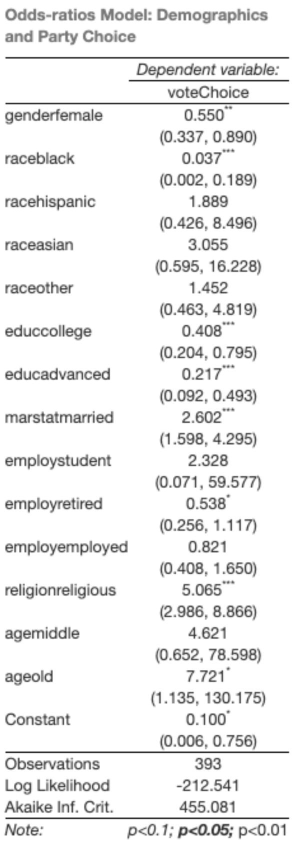
Model 2, which aimed to predict party choice, was not quite as successful in terms of classification, though many more variables showed a statistically significant relationship relative to their reference categories. In particular, women had about 50% lower odds of voting Republican than men, black Americans had about 93% lower odds of voting Republican than white Americans, and the higher educated categories were also less likely to vote Republican. Married individuals had 160% higher odds of voting Republican than unmarried people, and religious individuals had 407% higher odds of voting Republican than the non-religious. Young people (between 18-30) had much lower odds of voting Republican than both middle aged (30 – 49) and old participants (50+).



**Logistic Regression Baseline: Vote Choice**

| Accuracy | Precision | Recall | F1 |
|---|---|---|---|
| 0.74 | 0.62 | 0.68 | 0.65 |

This model was able to achieve accuracy of 74%, precision of 62%, recall of 68%, and an F1 score of 65%.



**Odds-ratios Model: Demographics and Party Choice**

| | *Dependent variable:* |
|---|---|
| | voteChoice |
| genderfemale | 0.550** |
| | (0.337, 0.890) |
| raceblack | 0.037*** |
| | (0.002, 0.189) |
| racehispanic | 1.889 |
| | (0.426, 8.496) |
| raceasian | 3.055 |
| | (0.595, 16.228) |
| raceother | 1.452 |
| | (0.463, 4.819) |
| educcollege | 0.408*** |
| | (0.204, 0.795) |
| educadvanced | 0.217*** |
| | (0.092, 0.493) |
| marstatmarried | 2.602*** |
| | (1.598, 4.295) |
| employstudent | 2.328 |
| | (0.071, 59.577) |
| employretired | 0.538* |
| | (0.256, 1.117) |
| employemployed | 0.821 |
| | (0.408, 1.650) |
| religionreligious | 5.065*** |
| | (2.986, 8.866) |
| agemiddle | 4.621 |
| | (0.652, 78.598) |
| ageold | 7.721* |
| | (1.135, 130.175) |
| Constant | 0.100* |
| | (0.006, 0.756) |
| Observations | 393 |
| Log Likelihood | -212.541 |
| Akaike Inf. Crit. | 455.081 |
| *Note:* | *p<0.1;* ***p<0.05;*** p<0.01 |

## 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 21, 2020 is identical to the version I submitted to the Examination Office on April 17, 2020.

DATE: April 21, 2020

NAME: Marina Wyss

SIGNATURE: