Predicting States' Political Alignment with Consumer Trends

Jay Ghosh Max Gannett Dillon Clark Tanner Richard
University of Colorado, Boulder[1] University of Colorado, Boulder[1] University of Colorado, Boulder[1] University of Colorado, Boulder[1,2]
jay.ghosh@colorado.edu max.gannett@colorado.edu dicl6768@colorado.edu tari5044@colorado.edu

[1]Department of Information Science [2]International Affairs Program

Abstract

Brands and corporations have not shied away from taking political stances in recent years (Hydock et al., 2019) Americans are recognizing and resonating with the messages of these brands which causes changes that start to align with political thinking. In applying logistic regression and random forest machine learning models on data from Google Trends and Simmons we have predicted the political alignment of continental U.S. states from 2010-2020, which further illuminates the connections between brands and election behavior phenomena for future discussion in the American political diaspora.

Introduction

The United States has seen a significant increase in the politicization of brands and corporations with each election cycle. Americans are recognizing and resonating with the messages of these brands which causes changes in behavior. In her book, *Authentic™: The Politics of Ambivalence in a Brand Culture*, Sarah Banet-Weiser explains how consumers intentionally seek out brands with similar cultural and social values. While the authenticity of these values is sometimes questionable, consumers tend to purchase products from brands that they can identify with politically. This pattern of consumer behavior provides the theoretical foundation of our project. We seek to analyze how this pattern scales up from an individual to a state-wide level.

The idea that conservatives and liberals act differently in social situations was largely pioneered by the political scientists, John Hibbing, Kevin Smith, and John Alford in their work on the environments of political difference and ideology. In their research, they discovered that conservatives are drawn to consistency and stability while liberals tend to be more willing to experiment and take risks. One way this manifests is in food tastes and other consumption habits (Hibbing et al., 2014). These differences in social behavior impact the general culture for different areas of the United States. We predict that these differences inform political behavior and will help us predict electoral outcomes.

Literature Review

In continuation with past research projects, we hope to evolve a basic concept we coined, computational informational politics. In growing this discipline, we hope to underscore the importance of combining computational and information science techniques for solving political science problems. We hope to expand on the work of machine learning practitioners in this field. Up to this point, there has been little research that combines consumer data with US house election predictions. However, there are numerous other projects that utilize machine learning to make political predictions. In their article, “Election Result Prediction Using Twitter Sentiment Analysis”, Jyoti Ramteke, Samarth Shah, Darshan Godhia, and Aadil Shaikh create a scalable election prediction model using political tweets. They explore the political attitudes embedded in the language of tweets about political candidates in the 2016 U.S. presidential election and train a model to predict how this translates to the greater population. While this study did not go so far as predicting the results of the election, it gets at the idea of abstracting widespread political sentiments from machine learning models. We plan to utilize a similar methodology in the paper.

However, the issue which is in question around political alignment in regards to consumer trends or preferences is a social discipline. There has been extensive cross-disciplinary research in the fields of political science, information science, and data

science. Political Scientist John Hibbing frames the idea of Political beliefs are a part of ourselves and not separate from ourselves and political psychology (Hibbing et al., 2014). If we are predisposed to like certain things or look at the world a certain way that bleeds over into preferences an individual chooses. Establishing a fundamental link that you can begin to discern between a conservative and a liberal from the brands, actions, and consumption habits (Vedantam, 2018). Furthermore, the connection between political strategies by campaigns to use consumer and product data as a frame for winning elections. Affecting political orientation may be altered as the environment, resources and other factors change over time (Bigi, Treen, & Bal 2016). With consumer products and brands being all-encompassing of the cultural dynamic of the electorate of the US, an interdisciplinary approach to the relationship between politics, consumerism, and data science might offer a new avenue of exploration.

We are far from the first to use search data for political prediction. Google Trends specifically has been an effective tool for election results in which traditional polling methods have failed. Mavragani & Tsagarakis (2015) explore the efficacy of using Google Trends data to predict the outcome of a 2015 Greek referendum regarding a bailout package from the central government of the European Union. This election was quite unique due to Greece's economic and political situation in 2015 as well as the one-week time frame between the announcement and vote collection. Traditional poll-taking methods are far less accurate in the context of short timelines like the Greek referendum. Mavragani and Tsagarakis instead looked at search results for "YES" and "NO" across the country during this one-week period. They then aggregate these results over shorter time intervals in order to make predictions about the outcome. The actual referendum results came in at 61.31% of votes "NO" and 38.69% for "YES". Mavragani and Tsagarakis' most accurate prediction was 58.2% for "NO" which was far better than traditional polls which predicted 54.5% or below for the "NO" vote share. These encouraging results showed the power of Google Trends for short-term predictions.

The study from Mavragani & Tsagarakis was far from the first to use google trends to make broad predictions. Their work expands on previous research from Polykalas, Prezerakos, and Konidaris (2013) that showed promising results using a simple predictive

model with Google Trends for election prediction in Spain and Greece. Similarly, this study used search data from the weeks leading up to the respective elections in order to predict the outcomes. This study noted that the closer to the election day the data was collected, the more accurate the results were. Mavragani & Tsagarakis saw accuracy comparable to that of traditional polling methods, all without the inclusion of demographic voter information or policy analysis. This research showed the promise of real-time Google Trends data for predicting political elections and preferences. However, these studies did not look at any search data beyond the actual elections themselves.

Google search data has not only been useful for political predictions but economic ones as well. More specifically, the plethora of search data has been useful in predicting consumers' product attitudes in place of more traditional product testing. Jun, Park, and Yeom (2014) demonstrated the potential of search traffic data for consumer preferences using the Toyota Prius as a case study. Through a detailed analysis of key product attributes relating to this vehicle, the researchers selected key search terms and time periods to create an econometric model that not only illuminates key trends for the Prius but for the whole hybrid car market as well. This study showed that the vast quantity of Google search data is not only useful for high-stakes elections or other political events. The consistent and widespread use of Google's search feature allows for economic insights to be generated over long periods of time. This finding is instrumental in our exploration of long-term search trends for brands and specific products. Despite the extensive research of Google Trends data for prediction, there is a lack of synthesis between economic indicators and election results. Our study seeks to combine these two areas in an American context to better predict US House election outcomes.

The usefulness of Google trends extends into stock price and volume prediction. Hongping et al. (2018) utilized Google trends data as a feature in a back-propagation neural net (BPNN), extending comparison to other models developed in previous studies. They applied a variety of economic indicators as features, including opening and closing prices, trading volume, as well as price highs and lows, as a control model. This model was then compared to one with equal training data, albeit including Google Trends search interest data as an additional feature. Their findings concluded a few use cases in which

Google Trends data was particularly useful as an indicator. From a broad view, the inclusion of Google Trends data as a feature in predicting the S&P and DOW Jones introduced marginal gains in performance, an average accuracy delta of +0.55% for their best performing ISCA-BPNN model *(Hongping et al., 2018)*. However, the inclusion of Google trends as a feature helped to predict the movement direction of a stock considerably more so than without. Similarly, it was especially useful to estimate opening prices, as the search sentiment of a stock is particularly volatile before opening hours, a period in which search trends are quite telling amidst a lack of other economic indicators during the overnight lull. In this study, Lu et al. found this data to be effective in predicting certain aspects of current economic trends remarkably quickly, even offering insight into future trends in the behavior of economic actors. *(Hongping et al., 2018)*.

In this sense, research using Google Trends data as a feature finds it to be most useful as a proxy variable of public interest, benefiting from the immediate availability the platform affords. At the very least, when included in an ensemble of inputs for machine learning models, Google Trends proves to noticeably outperform models that did not include it as a training feature. If search volume data is as reflective of society's interest as it had proven to be in economic forecasting, for the purposes of our study into political leanings, it may reflect public political sentiment as well.

Methods

We aim to use machine learning methods to produce a working model that can help answer our central question: *Can we discern political leanings through non-essential consumerism?* Utilizing this approach gains the validity of a working machine learning model that will predict the relationship and patterns of political meaning for the U.S. 2022 midterm elections. We intend to build and test a variety of feature-heavy machine learning models including Random Forest, Gradient Boosted Trees, as well as Recursive Feature Elimination. Ultimately, we intend to compare those against a classic Logistic

Regression model and other baseline models to determine accuracy and performance. We will use these models to solve both a regression problem, that of predicting the vote share, and also towards classification, whether each state will flip Democrat or Republican, and observe which method works best. Furthermore, applying an Independent Group t-test and the Paired t-test on this model compared to polling models for statistical significance to aid a wider discussion central to the research question. We will first train our model on house election data from 2010 to 2018, and test it on the 2020 data . After we have tuned and assessed our models, we will start to make predictions about future elections such as the 2022 midterms.

There is a very real possibility that our predictions based on brand interest data from Google Trends reflect wider socio-economic factors. Consumer brand interest could be representative of a variety of factors from annual income to company geography. In order to ensure that consumer brand interest is not simply a proxy for some other variable, we will run similar predictive models with more traditional economic features. We plan to run a model using GDP per capita and median income by the state in conjunction with the Google Trends data. Similar or better results from the economic data are a strong indication that the Google Trends data is a cumulation of smaller political indicators. In this event, we will discuss the viability of brand data as a more efficient method for political forecasting when compared to more traditional variables.

To acquire an adequate quantity and timeframe of Google Trends data, we leveraged a third party Google Trends python wrapper known as Pytrends. This library allows researchers to request individual 'payloads' of search data for a given search topic, time, and locale. For each of the selected 191 brands and corporations, a search interest query was requested per each election year from 2010 onwards. However, before sending each query, the string for each corporation was cleaned of unnecessary words and punctuation, such as "Inc," "Motor Company," or "Corporation," as our findings concluded that individuals rarely search for the full name of each company, instead settling for shorthand searches. Google Trends has a function to return similar requests, but included irrelevant searches, such as "Peach Bowl" when querying for "Chick-Fil-A," so this feature was not used. The data retrieved from each request included 50 columns, one for each

state, in which each datapoint reflects average search volume over the specified year. This integer result per state and corporation is normalized according to the state with the highest search volume, which is set to 100. In this way, the dataset reflects variance between states, rather than absolute search volume. Each request was then aggregated into a single dataframe per each year, shaped as 191 rows and 50 columns, not including indices. To incorporate these results into the machine learning training and testing sets, further data transformations were performed accordingly.

Results

We started by using a logistic regression classification model on the Simmons dataset, training on the 2016 dataset and testing on the 2018 dataset. This had a training accuracy of 0.833 and a testing accuracy of 0.54166. In comparison to running this with just Google Trends, we found a training accuracy of 1.00 and a testing accuracy of 0.9375. This indicates to us that the Google Trends vastly outperformed the Simmons dataset in predicting political alignment of states.

We then expanded our range of years to 2010, 2012, 2014, 2016, 2018, and 2020 as the Simmons dataset has limitations on the years we can use and Google Trends has no such limitations. Additionally, we expanded the list of companies we were scraping from Google Trends. In running a logistic regression classification model on this expanded set, we trained on the years 2010-18 and tested it on 2020. This had a training accuracy of 0.988 and a testing accuracy of 0.8.

To increase the speed and accuracy of our classifications we implemented a gradient boosted trees classification algorithm (XGB) and ran it on the same expanded dataset with the same train/test split. This had a training accuracy of 1.00 and a testing accuracy of 0.86.

To validate our findings we ran a cross validation with the XGB classification model, iteratively swapping out one year to be the testing set and the other years serving as our

training sets. The following shows our testing accuracy per testing year: {2010: 0.84, 2012: 0.84, 2014: 0.92, 2016: 0.9, 2018: 0.88, 2020: 0.9}. This shows us that there is a relationship between the Google Search Trends and the states' political alignments and that the model is not just picking up on some noise in the data.

After performing the cross validation, we moved on towards creating regression models to predict the actual vote share that we could expect in each of the states with the same dataset, using the XGB regressor model. After tuning and training this model, we found a training mean absolute error of 0.001 and a testing mean absolute error of 0.06. In converting the model's predictions to classifications we found a training accuracy of 0.85 and a testing accuracy of 0.88.

Finally, to check if our models were not just picking up on latent economic information, we ran an XGB regressor model using just two key economic variables: median income per state and GDP per capita per state. This had a training mean absolute error of 0.08 and a testing mean absolute error of 0.12. In converting the model's predictions to classifications, we found a training accuracy of 0.74 and a testing accuracy of 0.71. Additionally, we created a model combining both the economic and Google Trends variables for regression and that model had a near identical performance to the model that used just Google Trends. This shows us that using Google Trends to predict political alignment in states is better and captures more information than just classical economic variables. Below we can see an ablation table displaying different performance metrics across the regression models using just economic variables, the Google Trends variables, and both sets of features.

	Mean Absolute Error	Mean Squared Error	Classification Accuracy
Econ Train	0.079227049	0.01111136	0.744
Econ Test	0.116001006	0.024310883	0.7059

	Mean Absolute Error	Mean Squared Error	Classification Accuracy
Both Train	0.002257748	1.65E-05	1.00
Both Test	0.061140151	0.007881735	0.8824
GTrends Train	0.001473396	3.61E-06	0.848
Gtrends Test	0.059921494	0.007765341	0.88

Discussion

While the results demonstrated a strong correlation between consumer-brand and state political alignment, many questions remain about the possible significance of the data itself. There are numerous other factors that play a significant part of political election forecasting and prediction. Additionally, many of them are effective indicators of political sentiments such as macroeconomic conditions, average income, rural v.s. urban demographic spread, and media framing. Above, we validated our results by running a model with data representative of economic conditions like GDP per capita and population density. We then paired this data with a few key pieces of Google Trends data to achieve a high performing model. However, further research needs to be conducted to fully understand how both Google Trends and traditional electoral indicators provide better results than simply using the normal indicators alone. This raises questions regarding the validity of our results and methods. However, we do not believe it to be detrimental to our research. There is inherent complexity in the socioeconomic conditions of the United States such as rural vs. urban population spread or varying cultural values by region. We argue that our Google Trends-based predictions capture these factors at some high level. However, we admit that we do not fully understand exactly how this occurs. Further social-science based research is necessary to determine the relationship between

patterns captured by Google's search feature and the underlying factors of the American condition.

Additionally, it is important to note that the results demonstrated an alignment of political affiliations rather than the winners and losers of electoral outcomes. A different style of research project is required to make actual candidate predictions using Google Trends. However, we caution that there are systems and electoral infrastructures that contribute to less than representative outcomes in the US that would make this research difficult. That being said, we still assert that the use of Google Trends for political purposes is still a powerful and efficient tool. Strip away the complexity of the U.S. election system and one can see the aggregate outcome of political alignment through this user search data. It is essential to consider the impact of our results for the future of political predictions, whether it be small-scale local elections or even presidential ones.

Reflexivity

Our research has a large and consequential scope. We aimed to make election predictions and observe the relationship between consumerism trends and political alignment. We are by no means experts in these fields, however, we did our best to use ethical practices in data and information science to ensure we performed good research and made sound claims. That being said, our project sought to show a clear relationship between brands, consumers, and political alignment. Research like this has the potential to change political forecasting as a whole. For years, this field relied on traditional polling methods that are often cumbersome and time-intensive. The relationships we have worked to uncover in this project may offer a new direction for political sentiment estimation. While our research did yield promising results, further predictions based on our methodology should certainly experiment with different variables, methods, and features.

Furthermore, as researchers we come from different backgrounds that inevitably reflect our experiences when approaching our research question. Our team of four is composed of all information science majors, one international affairs, one minor in

business, and two minors in political science. However, we are of different racial, and religious backgrounds that also play a part in our research. There is utmost care and acknowledgment that we do not attend to general this research and conclusion to be used for predicting elections for purposes of winning.

Considering the data, much of the consumer data that is out there is locked behind expensive paywalls that we lack funding to access. Solving these data collection issues before running models and analysis was key to having successful research. We mitigated this by focusing our research on accessible and reliable data, Simmons and Google trends. Another challenge is the ethical implications of political modeling and consumer behavior using machine learning. We have made claims based on only a project that took only one semester which can be problematic. However, we wish to include a wider discussion of the implications of the findings between consumer trends and political alignment. We do not aim to provide a framework for profit or political campaigns to utilize.

Lastly, we recognize that while we found a compelling methodology for political prediction, we are still largely unsure of the deeper social and economic factors behind it. We have discussed possible trends and correlations that relate consumer brand interest to wider social phenomenon. However, we are certainly not experts in social science and we do not pretend to fully understand exactly what is driving the success of our model. We emphasize the need for further study in order to get a big picture view of how our machine learning predictions were so successful.

Bibliography

Bigi, A., Treen, E. and Bal, A. (2016). How customer and product orientations shape political brands. *Journal of Product & Brand Management*, 25(4), 365-372. https://doi.org/10.1108/JPBM-07-2015-0935

Hibbing, J. R., Smith, K. B., and Alford, J. R. (2014). Differences in negativity bias underlie variations in political ideology. *Behavioral and Brain Sciences*, 37(3), 297-307. http://dx.doi.org/10.1017/S0140525X13001192

Hydock, C., Paharia, N., and Weber, T.J. (2019). The Consumer Response to Corporate Political Advocacy: a Review and Future Directions. *Customer Needs and Solutions*, *6*, 76–83. https://doi.org/10.1007/s40547-019-00098-x

Hongping H., Li T., Shuhua Z., and Haiyan W. (2018) Predicting the direction of stock markets using optimized neural networks with Google Trends. *Neurocomputing*, *Volume 285*, *2018*, 188-195. https://doi.org/10.1016/j.neucom.2018.01.038.

Mavragani, A. & Tsagarakis, K. (2015). YES or NO: Predicting the 2015 GReferendum results using Google Trends. *Technological Forecasting and Social Change*, 109, 1-5. https://doi.org/10.1016/j.techfore.2016.04.028

Polykalas, S. E., Prezerakos, G.E., and Konidaris, A. (2013). A General Purpose Model for Future Prediction Based on Web Search Data: Predicting Greek and Spanish Election. *27th International Conference on Advanced Information Networking and Applications Workshops*, 213-218. 10.1109/ISSPIT.2013.6781856

Seung-Pyo Jun, Do-Hyung Park, Jaeho Yeom. (2014). The possibility of using search traffic information to explore consumer product attitudes and forecast consumer preference. *Technological Forecasting and Social Change*, 86, 237-253. https://doi.org/10.1016/j.techfore.2013.10.021

Vedantam, S. (Host). (2018). Hidden Brain [Red Brain, Blue Brain]. *NPR*. https://gimletmedia.com/shows/reply-all/brho4v/64-on-the-inside