Project Proposal- Data Driven Policy

September 15, 2022

Kyle Arbide and Asare Buahin

In the age of social media, we have seen how quickly political views and ideas can be spread amongst the masses. For example, the COVID-19 pandemic has shown how something as simple as a mask can be politicized, with its use-or lack thereof-being representative of a person’s politics. For each political issue, politicians have incentive to adjust the narrative of the problem in a way that will garner support in the form of potential voters. For our project, we are interested in examining the effects of policy changes on the types of rhetoric used by politicians on social media platforms, independent speeches, and congressional testimonies. Some of the topics we have considered for our analysis are those related to some of the major policy changes from the last 6 months. This includes abortion (the overturning of Roe v Wade), student loan forgiveness (Biden’s $10,000 forgiveness), climate change, and inflation (inflation reduction act)

We plan to examine how this change in speech is realized amongst popular politicians. The major events listed above are some of the hottest topics in the current political climate, and we hope to measure how rhetoric may have changed as a result of these events. More specifically, we will use Natural Language Processing to aid with analyzing the verbiage from speeches and social media posts by these politicians over time. To fully understand the Twitter would be the best place to grab social media data since it is the most popular site for politicians in the USA to share their politics. We can gather Twitter data through the Social Feed Manager set up by the GWU library. Furthermore, the character limit on Twitter posts will give us a clear and concise idea of the words that are being used. For speech data, The White House has an archive of transcribed speeches from Joe Biden. UVA also has transcribed iconic speeches delivered by past US presidents. [www.congress.gov](about:blank) has data on proceedings and debates of US congress which will be a dataset we can use to highlight commonly contested political opinions over time. We can also compare the buzzwords seen in social media posts and those seen in these congressional debates to see if we can identify any patterns or changes in language used during the implementation of Twitter user policy changes. Finally, to get transcriptions of campaign speeches from politicians up for election in November, we can use a creative techniques discovered by Data Science students at UVA which gather the auto generated captions from speeches posted on Youtube (<https://datascience.virginia.edu/projects/text-analysis-2020-us-presidential-election-campaign-speeches>). Applying their work to midterm election candidates rather than presidential candidates should allow us to gather speeches from a large number of sources efficiently.

We are also considering looking at how the rhetoric of right wing politicians has changed as a result of Trump presidency from 2016-2020. Has this radicalized the speech of a majority of the party members, or just a few? Do candidates following this rhetoric have greater success in the midterm elections, or is it a losing strategy? We will spend more time considering which question we would like to ask.

Reflection

Kyle Arbide

While going through the reading for this assignment, I tried my best to relate the content of the reading to my previous experiences through work and studies with Data Science. Throughout this exercise I was able to connect common concerns from the past to those mentioned in the reading to gauge a better understanding of the benefits and limitations of Big Data. Early within the article “The End of Theory”, I was reminded of the concept of models being a “black box”, specifically when the author mentions Google’s page rank algorithm and how “They don’t know why this page is better than that one”. Despite all the research being put into visible, explainable model and AI, without theory there can be a lot we do not know about problems where large amounts of data are available. The largest pitfall that I have come across through my studies that exists as a result of this “black box” is how some models have been proven after the fact to have large amounts of racial and gender bias, which was only ever addressed after they were put into practice and had the chance to affect the lives of many people.

This issue ties directly into one of the six provocations of big data, which is that claims to objectivity and accuracy are misleading. The part of this sectioned that resonated with me are that it does not matter how mathematically sound a model is, the second a researcher begins to interpret the results bias is introduced into the equation. The book “How to lie with statistics” addresses this for much simpler cases of visualization and averages (from a rather sinical point of view as well), so it easy to believe that it can also be done at a larger scale, with even greater consequences. Whether or not it is addressed, researchers have added incentive to discover a relationship where there may not necessarily be one, even though there is just as much knowledge gained in the discovering that there is no relationship. This section also mentions how again, even though the math is completely sound, data errors can introduce large amounts of bias into a model. These datasets that contain inherent bias within them also make it easy to understand the next provocation, that bigger data is not always better data. My start in data came from analyzing the results of survey studies, so I always found this concept easy to understand when looked at through the lens of survey sampling. You would never generalize the opinions of DC residents through the results of a survey conducted in Foggy Bottom. No matter if the survey has over a million participants, it will consist of a majority of young, college educated individuals who are not representative of the population of the district by any sense. This is a concept I will be sure to keep in mind as we work on our class project, as the paper addresses how this issue relates to using Twitter data in particular, and the unknowns associated with a sample of tweets taken through their API.

Finally, to touch on some of the points from “Where have all the Data gone?”, the rising popularity of these models may not be keeping up with the quality of data used to feed them. Whether this is from conflict of interest (what is mentioned in the paper) or from demand to produce actionable results without taking the time to collect proper, comprehensive data, we cannot always be sure that we are working with the best data there is to offer.