Introduction:

I started this project out hoping to give an in-depth analysis of internet chatter surrounding the 2018 Brazilian Presidential election. As a Brazilian citizen who is extremely interested in data surrounding online presence and usage, I could not help but wonder if there were particular elements that were predictive of election results. In 2018, the current Brazilian President Jair Bolsonaro was infamously elected after unprecedented level of pro-Bolsonaro political propaganda was spread WhatsApp. That same year Brazil had over 108 million users on WhatsApp[[1]](#footnote-1), the third largest amount out of any of other country in the world.

Given Brazil’s large online population, I decided to investigate if online search trends were at all predictive of the election winner.

Aim:

* To discover what are the predictive variables of winning a presidential election
* Particularly focusing on search terms on Google and their popularity

What is Covered in the Report:

- What are key indicators of voting labor?

- Search term trends before and after elections

- The predictive value of search terms on voting labor?

- The predictive value of geography on voting labor?

Problem Statement and Background:

I intended on using online chatter in order to predict if a certain candidates’ ability to win based on the amount of chatter surrounding their candidacy. However, after submitting my proposal in which I outlined how I would pull data from Twitter and Google, I realized that I did not have the ability to pull on Twitter without getting through a paywall and so quickly dropped the twitter component of the project. After spending a fair amount of time on Google Trends, I came to the conclusion that I could not simply just analyze how much each candidate was searched for as a measure prediction given that I would have too little data. With this in mind, I decided I would shift my project’s gears into seeing if certain keywords were predictive of a certain candidate winning.

Having realized that I would focus my project on Google and Google Trends, I started compiling a variety of different terms I believe Brazilians would be searching for in preparation for a Presidential election. I then planned on using these search terms as a predictive factor in the Brazilian Presidential Election. After asking multiple friends and family, I came up with a list of top 10 terms:

* Economia – economy
* Corrupção – corruption
* Previdência – Retirement Benefits
* Plano Governo – government plan
* Responsability Social – social responsibility
* Desemprego – unemployment
* Bolsa Familia – Brazil’s top welfare program
* Currículo – curriculum
* Pobreza – poverty
* Inflação – inflation

I intended to use 10 items as I planned on seeing overall mentions of each item and comparing them. In doing so, however, I ran into my first major problems with the project. The first was the fact that Google does not give raw numbers of how many times a term was used in a search but gives you relative terms instead. The second is that Google also caps the number of terms you can pull a relative search for at five and as such I would not be able to use all the search terms I had planned to. Issue number one is a recurring one that has multiple implications on the results of this paper and shall be addressed as we go along. However, issue number two was easily solved by putting them all in the search alternating words until I got both a diverse group of words that served to pull insights on different topics that would influence voters as well as terms that would have enough data for the analysis to be significant. I finally ended up with my five search terms which are the basis for the research conducted in this project:

* Bolsa Família
* Desemprego
* Economia
* Plano Governo
* Previdência

After narrowing down my search terms, I decided that I would try to pull data from the two weeks prior to each election round, however in doing so the software I was using *pytrends* would pull each day as an individual query and since the searches are relative, the levels for each search term would no longer be comparable and the predicted model not only flawed but also likely of very low predictive value. As such, I pulled the trends my month in using the *‘all’* timeframe while building my payload in pytrends, and after pulling interest over time, filtered out for only the searches done in the month of the election (October, which was pulled on November 1st), which I would compare against the search done in the month after the election (November, which was pulled on December 1st).

Since Google Trends allows us to pull data starting in 2004, I pulled the relevant keywords searches in four election years: 2006, 2010, 2014 and 2018. After doing so, it was clear to me that my research would require another level to it and as such I decided to not only pull the keyword search on a national level, but also pull them for the most populous state of each of the five Brazilian Regions[[2]](#footnote-2):

* North: Amazonas
* Northeast: Bahia
* Midwest: Mato Grosso Do Sul
* Southeast: São Paulo
* South: Rio Grande do Sul

Finally, when it came to the predictive model for this project, I realized that it would be hard to a winner by party because Brazil has over 30 parties[[3]](#footnote-3) and there are not two leading parties as there are in the United States. Yet, in the elections that I was analyzing, the Labor Party’s candidate made it to the final round ever time and as such I decided to create a Dummy Variable using the winner being from the Labor Party in each state as 1 and the winner being in any other party as 0.

Data:

All the data used in this project, apart from the word-of-mouth research I did to get search terms Brazilians would search for in the eve of each election, came from pytrends, the unofficial API for Google Trends.[[4]](#footnote-4) This Application programming interface (API) allowed me to pull results tracked by Google Trends in python, without having to manually scrape the Google Trends website. While I initially did struggle with pytrends in order to standardize my results, I was able to ensure standardized data through the use of “timeframe=‘all’” when I built my payload.

The unit of observation of this project was the relative number of times that certain keywords were searched in the month before and the month after the election, in Brazil and each of the most populous states of each of the 5 Brazilian regions.

Since Brazil does not have a two-party system in order for me to get the outcome of interest i.e. who would win, I created a dummy variable which accounted for the Labor party winning in that particular region, given that it is the only party that was in the final round of every election since 2006.

For the main part of my project, I used both the states that we are analyzing as well as the keywords as predictive variables in order to see if there was any predictive value in the results. After doing so I ran the model only using the keywords as the predictive variables.

There were a couple of issues in the data. While the missingness matrix did not show any data missing, there were a couple of search terms whose value in a certain state at a certain time was zero. Additionally, since the score we get for each search term is relative (i.e. a 50 doesn’t mean it was searched 50 times but rather has a score of 50 relative to a time in which there was the most searches and that would be 100), if this experiment was to become an actual model of use, there would need to be further manipulation of the data.

Analysis:

As previously mentioned in the data section, all of my data came from the pytrends API, however I did perform a variety of changes to the data in order to results:

Step-by-step of what I did:

1. First, I built out my payload which was the same for all the results I ran except for the geo which I changed each time
2. After getting the data for all the months since the API starts (2004) I narrowed it down solely to the months of interest i.e. the month of the election October (by retrieving data from November 1st) and December (retrieving data from December 1st)
3. After doing so for each geo – i.e. all the regions of interest, I melted each of the result tables I had in order from go from wide to long, so that I could draw graphs for Keyword Popularity in each of the 6 regions.
4. Next, I drew line graphs to show the popularity of each search term in each region, which was the easiest way to visualize the trends. I expected that there would be a sharp decrease between the October searches to the November searches and that was the case for the most part.
5. I then started working on my prediction model. I visualized all the distributions for the variable for the training data in order to see how my data was distributed.
6. After doing so, I preprocessed my data into the pipeline since we are only using information from the training data.
7. Next, I started to run machine learning models in order to get the best model, score and parameters for my data. I gaged its performance and finally I interpreted the importance of each variable in the after starting a permutation.
8. After seeing the first model I just ran the predictions for a model without the States as variables and got a different predictive value

Results:

There are two parts to my results. The first will focus on the graphs from each of the 6 geos (Brazil, Amazonas, Bahia, Mato Grosso do Sul, São Paulo and Rio Grande do Sul) and the keyword search trends throughout the years. My expectations would be that the search terms would be searched more in the month of the election (October) than in the month which follows. The second follows prective modeling and sees the predictive power of my model.

Part 1 Keyword Popularity Visualized

The plot for the whole country of Brazil, with the exception for the “Bolsa Familia” keyword represented by the red line, looks exactly the way I thought it would. All lines start with a sharp decline which indicates that the searches fell from October to November. While the falls may be small for some off the terms, all terms especially “economia” and “desemprego” can be seen to be following the predicted trends.

Next, we have Amazonas, which I had predicted would have some less than pleasing results. The reason being that Amazonas is a state in which internet connectivity is quite low. I was already prepared to have low results as I figured that it would not be too indicative of anything in this particular state. The results in this graph are actually counterintuitive to my hypothesis and yet in the grand scheme of the country graph as we saw previously, this had no major effect in the country’s overall trend. This also indicates that while we do not know how many searches were made in Amazonas, we can speculated they were quite low in context of the whole country.

Bahia for the most part (except for “desemprego”) seems to go along with our hypothesis that there would be drops in the month after the election. My prediction here would be that since “desemprego” also had a weird shape for Amazonas, it is a term that does not really rely on the elections for people to be interested in searching about it. However, as previously stated, given that this variable was searched for according to the hypothesis in the country-level graph I would suggest that the number of searches in Bahia was small compared to that of other states.

While at first glance the keyword popularity for the state of Mato Grosso do Sul seems to be all over the place, it is possible to see that those sharp decreases from October to November are there for almost all years and key terms. The blue line for “previdencia” here is the most deceiving because of the sharp decrease from 2006 to 2010, nevertheless it still follows the expected trends.

Next, we have the State of São Paulo, the most populous state in Brazil and the one responsible for the biggest amount of the country’s GDP. Apart from “previdencia” and “Bolsa Familia” in 2010, the lines in this graph seem to follow our initial hypothesis and is quite similar to the searches on a national level. This confirms the aforementioned claim that states with a higher population weight more overall on the country level data (as they logically should).

Finally, we have the state of Rio Grande do Sul which exhibits a lot of the same trends as the state of São Paulo, however in different magnitudes. It is possible to see here that “previdencia” in 2010 also exhibits counter intuitive trends and since this has been the case for this variable and year more than once, it would be interesting to investigate what could have caused that.

Modeling and Predictive Power of Model

Going into modeling I expected a few hiccups given that my dataset was quite small and so my train data. Before going into the models and seeing which one worked best, I did a distribution visualization of the and could clearly see that the variables were all over the place.

Next I turned to modeling having voted for the Labor party as the dummy variable we would predict. To my surprise got a very high best score for my modeling of <b> 0.9833333333333332 </b>. I then proceeded to find out that the <b> Random Forest Classifier </b> was the best parameter in which to run my model and the ROC AUC score that shows the performance of the model was a <b> 0.9444444444444445 </b>. These are significantly high for a model and began to get suspicious of them. The accuracy score for the model was <b> 0.75 </b> which seemed to be more in line with model. However, after running the variable importance table and subsequently the plot, it became clear why the model was running so well.

As it is clear from the graph above the inclusion of the states in the model, took away variable importance from the Keywords, which clearly seem to have no correlation whatsoever with the Labor votes at least in a model including the states. Bahia voted for labor during all the elections which is clear from the Reduction in AUC ROC graph which it would be the most importance variable for prediction.

However, I decided that this was not enough, and my exploration was not done. I then decided to run the model without any of the states as variables. In doing so, I saw that the best score for my model was a significantly more humbling of <b> 0.6733333333333332 </b> and that the best model for it was the <b> KNeighborsClassifier. </b> My ROC AUC score also significantly dropped having a low score of <b> 0.5698051948051949 </b> and an accuracy score of <b> 0.6111111111111112 </b>.

Discussion:

As it is possible to see, my initial model was quite predictive. Even though it did not measure the relationship I intended, I can now firmly say that I can name variables which hold significant predictive value to voting labor: the different states. While we can see that the country as a whole, the state of Bahia and the state of Amazonas are highly likely to vote labor, we can also see that São Paulo and Rio Grande do Sul were unlikely to vote labor, which considering the states’ ideologies makes complete sense and makes sense why the model would be so predictive.

However, when we remove the state variables from the model, we can see that the best scoring model scores a mere 0.673, which is significantly lower than the first model. Still, it is quite interesting to see that there was still a high probability than a coin toss for the model’s predictive value.

Ultimately, I believe that this project had a lot of potential pitfalls, from the struggles of using \*pytrends,\* to the way google ranks search trends and more. I believe the project to have been a success because I was able to run two models and have the second one which was focused on key terms give some level of predictability. Beyond that I am firmly able to say that a voter’s location is significantly more indicative of if they will vote labor or not than their search trends.

If given more time, I would rerun the second model and tweak it to see if I could find any relationships that existed between key terms and voting labor. I would also attribute voting labor to winner or not, in order to see what would be worth exploring as a source of predicting election results. I would also have pulled significantly more data, from more terms, standardized them and analyzed them over a longer period of time.

WhatsApp Revenue and Usage Statistics (2021). (2021, November 11). Retrieved from https://www.businessofapps.com/data/whatsapp-statistics/

População brasileira chega a 213,3 milhões de habitantes, estima IBGE. (n.d.). Retrieved from https://www.gov.br/pt-br/noticias/financas-impostos-e-gestao-publica/2021/08/populacao-brasileira-chega-a-213-3-milhoes-de-habitantes-estima-ibge

Pytrends. (n.d.). Retrieved from https://pypi.org/project/pytrends/#interest-over-time

1. https://www.businessofapps.com/data/whatsapp-statistics/ [↑](#footnote-ref-1)
2. https://www.gov.br/pt-br/noticias/financas-impostos-e-gestao-publica/2021/08/populacao-brasileira-chega-a-213-3-milhoes-de-habitantes-estima-ibge [↑](#footnote-ref-2)
3. https://www.tse.jus.br/partidos/partidos-registrados-no-tse [↑](#footnote-ref-3)
4. https://pypi.org/project/pytrends/#interest-over-time [↑](#footnote-ref-4)