Polarization and Sentiment Forecasting Using Various Economic Terms Such as Diversity and Education

Zihao(Gavin) Zou
April 5, 2023

Abstract

This paper examines how various economic terms came to predict people's polarity and sentiment. With this research question, this paper investigates on Twitter dataset taken during the 2020 US election, and the dataset covers over 1.2 million tweets with corresponding attributes from October 15th to November 9th, 2020. More specifically, this paper discusses the following aspects: 1) Examine and skin through the apparent relationships between sentiments and the chosen factors. 2) Investigate the location variables and dig into the geographical predicting factors. 3) Adding more empirical economic terms and research into statistic models to control for variables and refer to significant conclusions. Ultimately, we conclude that economic terms such as user's internet exposure, diversity score, education levels and days to the election can well predict polarization score. In contrast, some other terms do not show economically significant relationships.

1. Introduction

Social media posts have been the target of social research for a long time. Using computer programs and AI databases, researchers can gain more insights from the posts. Plenty of papers have investigated sentiment analysis and related specific topics, as Pagolu figures how sentiment on Twitter "predicts stock market movements" (Pagolu et al., 2016). There are also papers investigating the effectiveness of sentiment analysis with programming models; for example, Agarwal provides a solution of "a binary task of classifying sentiment into positive and negative classes" (Agarwal et al., 2011). However, there isn't a detailed paper that tries to identify key attributes of users and use them to forecast a person's emotions or polarity. Thus, in this paper, I explore how various factors can show a predictive effect on people's sentiments and polarization using the 2020 US Presidential election tweets. The Y variable is the polarization score, while the X variables are state, followers, user join date, country, and tweet posted time. Meanwhile, as we go further down the report, I will add other economic terms for forecasting and prediction analysis on variables such as GDP, diversity score, population, etc.

The polarization score is calculated using the tool pack called TextBlob. TextBlob connects to a "large natural language processing model" that is continuously developed (Barai, 2021). This model takes in a sentence and calculates the polarization score that ranges from -100 to 100, where a negative one hundred shows extremely "assertive, aggressive, or hostile," while one hundred indicates extremely "approachable or positive" (Barai, 2021). More specifically, the program abstracts each word from the sentence and calculates the sentence's sentiment score based on each word's scores. For example, the word "happy" is given a score of 50, so the sentence "I am happy" will result in a polarization score of positive 50 - as the word "I" and "am" have individual scores of 0. This score is believed to identify "people's sentiments and polarization accurately," according to Barai (Barai, 2021). Thus, I am using data generated by this tool pack in this research.

The first five X variables are the most frequently seen and used in academic research. Meanwhile, the variables are also chosen based on my limited choices from data scraped from Twitter. The summary of why I chose those variables is in the summary statistic table and in below. First, the state shows how the location within the United States affects what people discuss, as I am interested in potential geographical patterns in the United States. Second, the country variable shows how international regions will affect topic popularities. Third, join data implies a person's internet experience and exposure, and we observe how that could affect people's polarity. Fourth, followers show how social media influence can affect people's way of speaking on different issues. Last, created data identifies how topics trend over a period, thus allowing me to analyze the overall polarization score on the internet over time. Then, I will also introduce education levels, diversity scores and population for the four regressions targeting user attributes, tweet attributes, education, and economy attributes.

As we shall see in the following report, in the graphing section, I identify significant relationships between the polarization score and variables: locations, number of followers and tweet created time. While for the account registered time, which is an indicator of the user's previous exposure to the internet, has an ambiguous relationship with the polarization score. Those patterns will be explained in more detail in the paper below. The mapping section introduces location significance and explains significant geographical distribution of the polarization score. The regression and machine learning proves the significance of specific variables when we separate variables into specific categories. Finally, we draw a simple conclusion that economic terms such as diversity score, education levels and created day can

well predict polarization score. In contrast, some other terms do not show economically significant relationships.

This report is separated into four parts. The first part is on some basic summary statistic tables and general graphs, which I use to refer to a general relationship between the Y and X variables. The second part details the graphs and maps, where I will present more specific visual and geographical relationships if they exist. The third part is where I scrap from the internet and find potential datasets that could further resolve the concerns of previous parts. Finally, the last part introduces regressions and machine learning on regression trees to explore more accurate relationships and investigates some machine learning predictions. Now, we will first investigate the data section.

2. Data

In this paper, we are using the tweets and the corresponding information of the tweets during the 2020 US election. Meanwhile, before we dive in further, it is important to understand the background of the dataset. There are several reasons for using election tweets for sentiment analysis. First, the US election is not only a national event, but the world is also "watching the race and trying to determine what the outcome will mean for them" (Wu et al., 2020). Because people were most active on social media during that period, this event will give us the best opportunity to do a wide range of sentiment analysis and allow us to obtain more data. Second, the popularity of politics has been at the top of all other topics. Particularly, people tend to get emotional and irrational with political debates. Thus, this political debate background will exaggerate people's polarization, making this analysis easier.

Therefore, I found a Twitter dataset on Kaggle, provided by Manch, a data scientist, available for download. Manch set a time constraint using Twitter API to filter out the election tweets when downloading the files(Hui, 2020). So while the 2020 US election happened on November 3rd, 2020, the data was filtered and collected using dates from October 15th to November 9th, 2020. Manch Hui collected those data using Twitter API to research his interest in data visualization and election results(Hui, 2020), and this dataset is particularly useful for this paper.

After data cleaning, merging, and sorting, there are over 1.2 million tweets and their corresponding information, such as posted date, country, etc. Using those 1.2 million tweets, I generated a polarization score according to the text for each one of the tweets. Therefore, we have the initial Y variable and X variables for the research. Furthermore, in the upcoming sections, I scraped data from Wikipedia to learn about each country's population, diversity score, and each state's education level (Wikipedia, 2023). Using those scraped data, I expand the list of X variables to gain more insights into economic factors predicting people's sentiment and polarity. And now, we will get a glance at the summary statistics tables.

Meanwhile, I am also going to use data on the country's GDP estimates, racial diversity level, and each US state's education level for further predictive analysis. For those data, I scrape them from Wikipedia using the Python library BeautifulSouop. And all those data will conclude the data being used in this paper.

3. Summary Statistics

In the first step, I chose five variables from the original dataset to examine if there are any apparent relationships between sentiments and the chosen factors. Therefore, with those selected five variables and the Y variable polarization score, we are going to look at the summary statistics first. Please note that summary statistic table outputs are attached at the end of the section instead of after each interpretation.

The first variable, "created time," refers to when the tweet is published on Twitter. The reason for choosing it is that we can relate the timing of a post to what's happening in the world, and thus we might be able to find some interesting facts. In this case, timing is even more critical because we use the 2020 US election data. As it moves closer to the time of the election, people's sentiments or polarization is likely to increase. Therefore, it is important to look at this relationship in our analysis; meanwhile, we can look at the data and see if it truly exists. In this summary statistics table, we find that there are about 1.3 million timestamps recorded in 26 days(because we rounded dates by day when cleaning the data). The 31st of Oct in 2020 is the middle day according to the number of tweets published within the 26 days. With more than 1.3 million timestamps and a massive amount of sample size, I believe the data can represent reality well.

The second and the fifth variable are location variables. The reason for choosing them is because the different region has different culture. This culture can hugely affect how people do and say things on Twitter. Thus, considering the effect of the country, we can relate to the sentiments of people around the world and compare if there are any significant differences. Meanwhile, the summary statistics table is straightforward, simply pointing out which country and which state has the most tweets sent out.

The third variable is the "user follower count," which refers to the "fans" an account has on Twitter. The reason for choosing this is that we always debate whether popular people tend to be polarized or judgmental. Are those popular people all polarized, or are they less polarized and more open to other opinions? This topic relates to this research because people often get extreme in politics, so studying the effect of fans during this special period couldn't have better timing. In this summary statistics table, we find that, on average, an account has about 17k fans, with a huge standard deviation. Meaning this data is very positively skewed. Meanwhile, we can also tell that almost 75% of people don't have fans bigger than 2k. When dealing with this data, graphs might look messy. I must be more careful about outliers.

The fourth variable is the "join year," which refers to the year the Twitter user signed up for the account. This section can be an excellent predictor of how much internet exposure this person has. An interesting topic is: do people who join the internet longer will be polarized on what they say? Again, as we are looking at the US election data, polarized people are more prominent, and thus it's easier for us to identify the effects. The summary statistics table shows that about 1.5 million people were registered not long ago in 2019. 2019 has been the highest registration in the years. All people in this dataset are registered across the 16 years horizon. However, 16 years is a long time, and registrations might be more linear. Thus, I should look for outliers and be careful when referring to certain relationships.

Finally, for the Y variable polarization score, the summary statistics table shows that the max and min of the score are 100 and -100, respectively. The mean score is 5, meaning people on Twitter during the US 2020 election are slightly positively polarized. For this research, we will carefully investigate the score to identify possible relationships.

	country
count	581979
unique	188
top	United States of America
freq	265698

	join_year
count	1279738
unique	16
top	2019
freq	157094

	state
count	433137
unique	702
top	California
freq	45780

	created_time
count	1279738
mean	2020-10-31 15:42:55.972894464
min	2020-10-15 00:00:00
25%	2020-10-26 00:00:00
50%	2020-11-04 00:00:00
75%	2020-11-07 00:00:00
max	2020-11-09 00:00:00

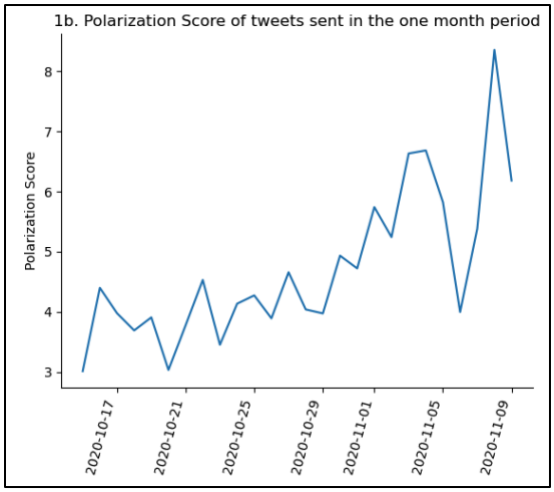
	user_followers_count
count	1279738.000000
mean	17477.614726
std	314040.065344
min	0.000000
25%	75.000000
50%	422.000000
75%	1953.000000
max	82417099.000000

	polar_score
count	1279738.000000
mean	5.529619
std	26.227029
min	-100.000000
25%	0.000000
50%	0.000000
75%	10.000000
max	100.000000

4. Initial Visualization

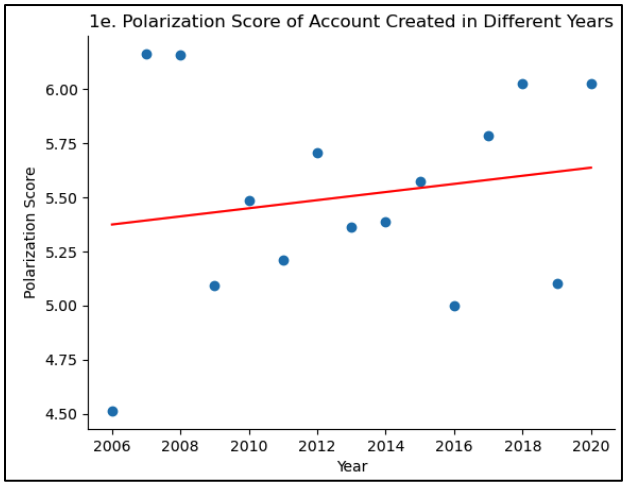
Now that we have initially understood the distribution and summary statistics of all variables, we will now dive into some potential relationships that exist between those variables. In the following graphs, I only illustrate variables that appear to have apparent predictive relationships with the polarization score. For ambiguous relationships, I explore them with more quantitative methods in the regression section. Thus, in the space below, I will introduce graphs for variable tweet created day, account registered year, and state(location variable).

For the variable created day, as introduced above, measures when the tweet was sent on Twitter. In this case, we measure how close it is to the election date. In observing this graph, we see that, when it gets closer to the US election, people are getting more than more polarized across the one-month period. This graph follows our intuition as the debates over the election should become more aggressive over time.

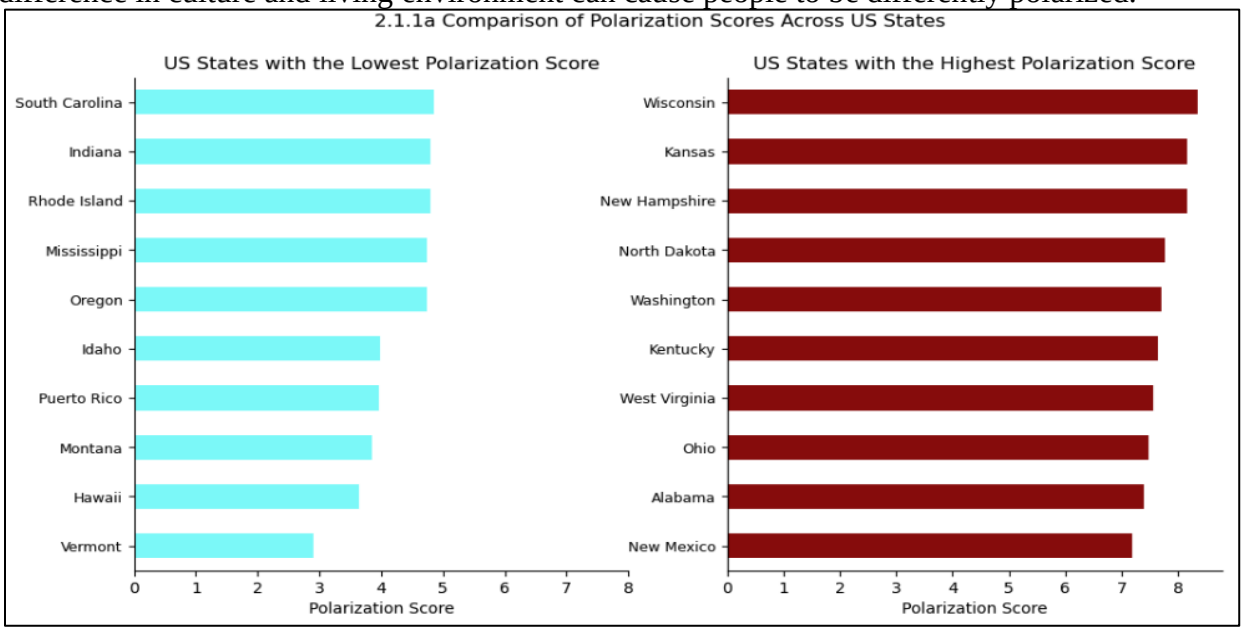


For variable accounts created a year, as introduced above, measures when the account is registered. This variable represents how much internet experience the user has. Meanwhile, for better observations, I also drew a red trendline shown below. I didn't make this a line graph because the dots are spread out, and a line graph won't visually look good. Instead, a scatter plot with a trend line does a better job here. In observing this graph, we see an upward-sloping relationship between the account registered year and polarization score. So, this graph is saying that people who registered later and people who have less internet experience are likely to send

out more polarized tweets. The potential reason could be that people with less internet exposure know less about the internet environment and thus become less careful and more polarized on what they say.



For the variable location state, the bar chart below shows some interesting output. First, even the lowest-scoring states have scores above zero, meaning people in the US are at least somewhat positively polarized. Second, I intentionally set both axes into the same scale to better compare; and it turns out that the difference between the highest and lowest scores is not big. Thus, we should not conclude that there is a significant difference in the polarization score across the country. Third, if we are to compare, we find that most high-scoring states are in the east of the US(Wisconsin, Ohio, Kentucky, West Virginia etc.), while lots of low-scoring states are in the south(Montana, Idaho, Oregon etc.). So maybe we can refer to the fact that residents in the east of the US tend to send out more positive tweets than those in the West. Lastly, the potential reason for such distribution can be different cultures and environments. For example, in Canada, Vancouver people tend to be more relaxed and less busy than those who live in Toronto - this difference in culture and living environment can cause people to be differently polarized.



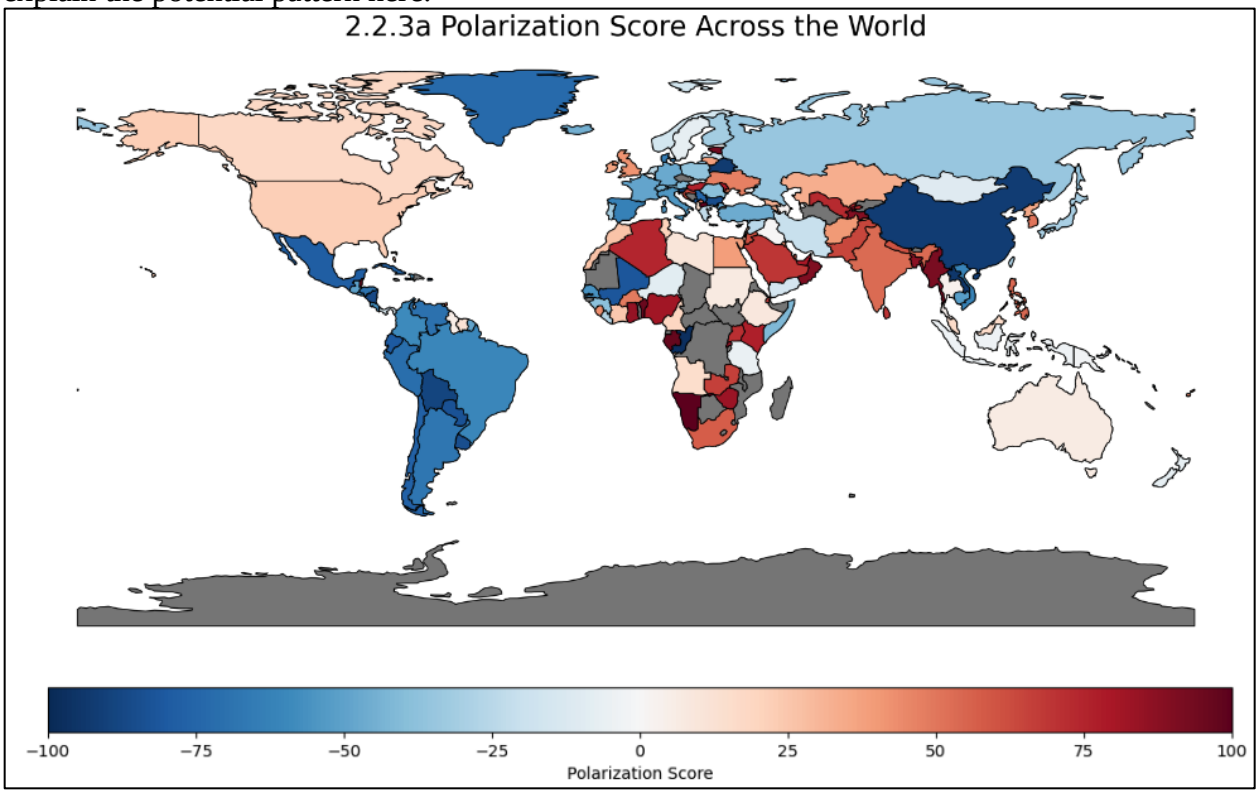
However, for the location variables, we will address the question better in the below mapping section, where I directly show the map for the polarization within the US. I am offering

this graph here so we can have different perspectives to see if the phenomenon we observe is valid. And now, let's turn to the geographic visualization section.

5. Geographical Visualization

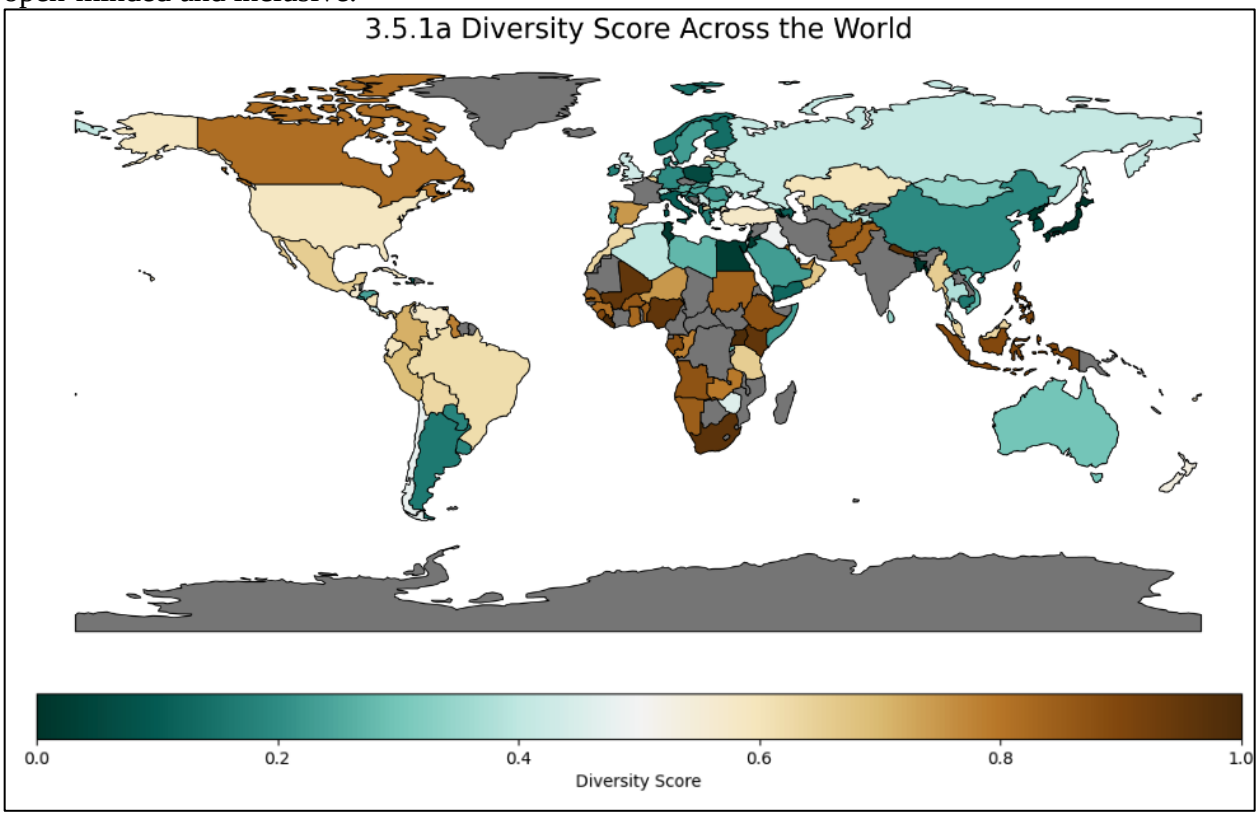
Quantitative variables can be well shown using line graphs and bar charts, but location variables can only be well-presented with maps. Thus, in this section, we investigate some mapping visualizations and summarize some predictive relationships. In the space below, we will first look at the distribution of the polarization score across the world and the US on the map. Then, to explain such a pattern shown by the map, I add diversity score and education levels for comparison.

For the world polarization distribution, the map shows that polarizations of tweets across the world have huge differences, with gray colour representing missing observations. North America tends to have more positive tweets, while South America and Europe tend to have more negative tweets. While Australian countries were in the middle ground - not too positive nor negative, African countries had mixed polarization scores across the continent with some missing observations(as no tweets are posted from those countries). The potential difference can be for many reasons, such as level of income, education, living environment, culture etc., or it could be the combined effects of all those variables. To dig deeper into the trends and the root reasons for those patterns, I am going to use the diversity score as a new variable to try and explain the potential pattern here.



The world diversity score map looks like the one below, with gray colour representing missing observations. As the map suggests, most Asian countries and European countries are not much diversified, while most North American countries and African countries seem to be very diverse. The reason could be that the culture in Europe and Asia is less open or simply because the immigration policies are stricter than those in North America and Africa.

We can compare this map with the previous map on the polarization score. It's easy to find out that the yellow countries(more diversified) in this map are pink countries(more polarized) in the polarization map. Thus, there might be a relationship such that people are more polarized in more diversified countries. This phenomenon seemed contradictory to our intuition: I thought people in diversified countries should be less polarized because those people are more open-minded and inclusive.

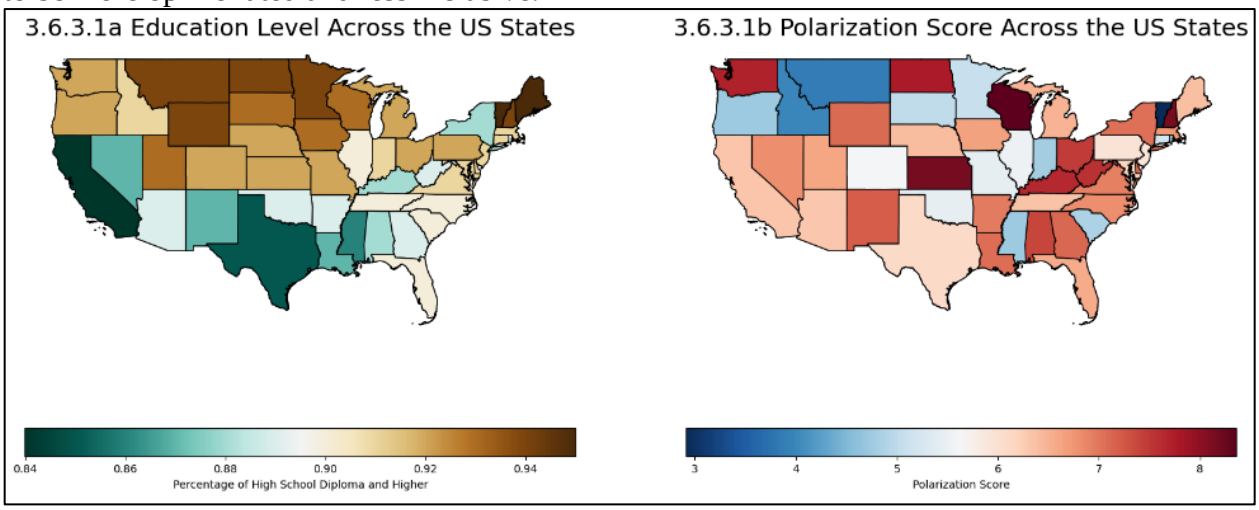


Now we've seen the world map, we will also be looking into the US because we are using the US election data, and by limiting the range to one country, we better control for other effects. Thus, I will draw a map of the US polarization score distribution and try to explain this distribution using people's education levels.

First, we look at the distribution of the US polarization score. Remember from the previous part, we inferred from the previous part that: we find that most high-scoring states are in the east of the US(darker red), while lots of low-scoring states are in the south(lighter red and blue). However, the summarization that "most high-scoring states are in the east of the US, while lots of low-scoring states are in the south" might not be true. By looking at the map, low-scoring states are mostly in the northwest instead of the west, while the eastern states tend to be more polarized overall. Thus, this map would mean that people in the east of the US tend to send out more extreme, less open mind, and more opinionized tweets than those in the northwest of the US.

Second, does an education level explain a part of the pattern? We look at the map of education levels across the US and the polarization map together. As we can see from the two graphs, most green states to the left(low educational level) are red on the right(high polarization score). To see this pattern better, we can look at states in the south; those states are green in the left graph and red in the right graph. We can also look at the north. The blue states in the right graph are yellow in the left graph. Thus, it's likely that a higher education level will imply a

lower polarization score. Economically, this pattern suggests it's likely that education level and polarity are negatively correlated. Practically and realistically, people in less educated areas tend to be more opinionated and less inclusive.



However, by now, we have only observed very broad relationships by maps and line graphs. Even though those relationships give us very good hints on how the final predictions may look, we are not sure how accurate and effective those relationships are. Thus, we need multi-regression models to validate those relationships. Meanwhile, with the help of previous investigations, we are more effective in choosing our regression variables. Thus, we will now look at the regression and machine learning section.

6. Regression and Machine Learning

In this section, we investigate regression models to validate our previous findings and draw final conclusions. With the help of previous findings, I classified all variables into four categories for four regressions. Those models focus on user attributes, tweet attributes, education levels, and country/economy specific, respectively. Lastly, I will conduct some machine learning practices and compare the results.

However, before I dive in further, there are a few common attributes of all the regressions. Thus, I will mention them here at once instead of repetitively explaining them. First, because of the huge sample size and nature of the variables chosen, it’s hard to identify the true relationships between X and Y. Thus, I think about the relationships intuitively, and linear assumptions make sense for all X variables. For example, squared or root of the population won’t make sense in an interpretation, while linear interpretation works the best. Thus, I assume linear relationships for all the X variables. Second, because of the huge sample size, all regression results turn out to be highly statistically significant, including but not limited to the P value, F test, and student T-test. Third, also because of the huge sample size, with more than 1.2 million rows of data, the regression R-squared is going to be very small. Particularly, for example, tweets from Canada can have a polarization score from -100 to 100, but as long as the tweets are sent from Canada, they will have the same X variable population. This effect makes the R-squared even smaller. Thus, based on the huge sample size and the nature of this research, we will not focus too much on R-squared. Finally, the objective function for all regressions is to minimize the squared error term with the following objective function:

$$\min \frac{1}{N} \sum_{i=1}^{N} (\text{Polarization Score}_i - (\beta_0 + \beta_1 X_i))^2$$

With all those information and background on hand, we will now investigate those regressions and regression tree one by one.

>>> Regression on User-Specific Attributes

For the first regression, I am going to look at user-specific attributes. This user attribute topic is important because everyone is different in some ways. The difference in those factors can largely influence how and what a person says on the internet. Thus, a muti-regression on user attributes is needed to infer some individual-specific relationships to the polarization score(Y variable). Thus, I will choose the following three variables: User join date, fans owned and whether in the US or not, and the regression equation should look like the following:

$$PolarizationScore_i = \beta_0 + \beta_1 \widehat{JoinDate}_i + \beta_2 \widehat{FansOwned}_i + \beta_3 \widehat{InUS}_i + u_i$$

This regression yields the results in the following table.

To interpret the results, first, the intercept does not have any economic meaning. Because it does not make sense to have a user who registered in year 0 when the internet is not yet established, we will focus on the other three variables.

Second, looking into the variable join year, which indicates the user's previous internet exposure, the coefficient is slightly positive at 0.034. This result says that people with less internet exposure(join year is bigger) tend to be more opinionated and less careful about what they say online. Concerning the research question, we would say that: the user join year can well forecast people's polarization level, so that when people have one year less of internet experience, their polarization score tends to increase, on average, by 0.034 points. 0.034 point is not economically significant considering the base range of polarization score is from -100 to 100, but this number is highly statistically significant. Again, this makes sense because people who know the internet well will be cautious about what they say and tend to be less polarized - so they don't become the target of attacks on the internet.

Third, looking into the variable "user follower count," an indicator of the user's social media influence, the coefficient is rounded to 0 with three decimal places. This result says that people's social media influence does not predict an increase or a decreasing trend in the polarization score. Concerning the research question, we would say that: the user followers count can well forecast people's polarization level as the coefficient is highly statistically significant; however, there doesn't seem to be a straightforward linear and numeric relationship between those two variables.

Fourth, looking into the variable "InUS," which is a dummy variable indicating whether the user is in the United States, the coefficient is positive at about 0.83. This result says that people in the US during election time tend to be more opinionated. About the research question, we would say that: using the 2020 US election data, the user location can well forecast people's polarization level so that if a user is in the US, their polarization score tends to be bigger, on average, by 0.83 points than those who are not in the US. 0.83 point is not economically significant, considering the base range of polarization score is from -100 to 100, but this number is highly statistically significant. Again, this makes sense. As it's the US election, people in the US tend to be more polarized than people elsewhere. Plus, having "InUS" as a variable better control the effects for other variables.

To sum up, this model shows that user registered time and whether the user is in the US have predictive effects on the polarization score. The prediction says that people with less internet experience and people in the US tend to be more polarized.

	Dependent variable:polar_score		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Intercept	-53.756***	-50.879***	-63.636***
	(17.576)	(17.626)	(17.657)
join_year	0.030***	0.028***	0.034***
	(0.009)	(0.009)	(0.009)
user_followers_count		-0.000**	-0.000
		(0.000)	(0.000)
InUS			0.830***
			(0.070)
Observations	581,979	581,979	581,979
R^2	0.000	0.000	0.000
Adjusted R^2	0.000	0.000	0.000
Residual Std. Error	26.668 (df=581977)	26.668 (df=581976)	26.664 (df=581975)
F Statistic	11.536*** (df=1; 581977)	8.155*** (df=2; 581976)	51.739*** (df=3; 581975)
Note:			*p<0.1; **p<0.05; ***p<0.01

>>> **Regression on Tweet-Specific Attributes**

For the second regression, I am going to look at tweet-specific attributes. This tweet attribute topic is important because each tweet is different in some ways. The difference in those factors can largely predict or indicate the overall social environment on the internet. Thus, a muti-regression on tweet attribute is needed to infer some tweet-specific relationships relating to the polarization score(Y variable). Thus, I will choose the following three variables: retweet counts, likes and created day, and the regression equation should look like the following:

$$PolarizationScore_i = \beta_0 + \beta_1 \widehat{ReTweet_i} + \beta_2 \widehat{Likes_i} + \beta_3 \widehat{CreatedTime_i} + u_i$$

This regression yields the results in the following table.

To interpret the results, first, the intercept represents the average polarization score for tweets sent on 2020-11-15 with zero likes and zero retweets. This intercept gives us an understanding of the overall internet atmosphere as a baseline. The intercept has a highly statically significant polarization score of 2.735.

Second, looking into the variable retweet count, which indicates the popularity of a tweet, the coefficient is slightly negative at -0.002. This result says that tweets that are more popular tend to be less opinionated and more inclusive. Concerning the research question, we would say that: the retweet counts can well forecast the polarization level of the tweet so that when the number of retweets goes up, the polarization score of the tweet tends to decrease, on average, by 0.002 points. 0.002 point is not economically significant, considering the base range of polarization score is from -100 to 100, but this number is highly statistically significant. This result suggests that popular contents on Twitter are more polarized, which aligns with our previous conclusions.

Third, looking into the variable number of likes, a broader indicator of the popularity of a tweet compared to the variable "retweet number," the coefficient is rounded to 0 with two decimal places. This result is not surprising as its range of measure is, in nature, broader than retweet counts. These zero coefficients say that people have different tastes regard to the content they see on the internet, so the number of likes does not show an increasing or decreasing trend in the polarization score. Concerning the research question, we would say that: the number of likes can well forecast people's polarization level as the coefficient is highly statistically significant; however, there doesn't seem to be a straightforward linear and numeric relationship between those two variables.

Fourth, looking into the variable created day, which indicates how close it is to the US election: the higher the number is, the closer it is to the US election. The coefficient is positive at about 0.17. This result says that the closer it is to the US election, people tend to be opinionated. About the research question, we would say that: using the 2020 US election data, the time to the election can well forecast people's polarization level so that when it gets one day closer to the election, people's polarization score tends to be bigger, on average, by 0.17 points. 0.17 point is not economically significant, considering the base range of polarization score is from -100 to 100, but this number is highly statistically significant. Again, intuitively this makes sense. As it gets closer to the election date, more arguments and debates tend to show up, raising the polarization score.

To sum up, this model shows that the number of retweets and the day to the election have a predictive effect on the polarization score. The prediction says that tweets of more retweets tend to be less polarized, and the time closer to an election event makes people more polarized.

	Dependent variable:polar_score		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Intercept	5.529***	5.528***	2.735***
	(0.023)	(0.023)	(0.058)
retweet_count	0.000	-0.002***	-0.002***
	(0.000)	(0.001)	(0.001)
likes		0.001***	0.001***
		(0.000)	(0.000)
created_day			0.168***
			(0.003)
Observations	1,279,738	1,279,738	1,279,738
R^2	0.000	0.000	0.002
Adjusted R^2	-0.000	0.000	0.002
Residual Std. Error	26.227 (df=1279736)	26.227 (df=1279735)	26.198 (df=1279734)
F Statistic	0.135 (df=1; 1279736)	13.376*** (df=2; 1279735)	933.345*** (df=3; 1279734)
Note:			*p<0.1; **p<0.05; ***p<0.01

>>> Regression on Education Levels

For the third regression, I am going to look at education levels. This education attribute topic is important because levels of education largely affect and determines how a person talks and behaves. The difference in the levels of education can largely predict or indicate the overall polarization scores as well. Thus, a muti-regression on education attributes is needed to infer

some relationships to the polarization score(Y variable). Thus, I will choose the following three variables: high school and higher percentage, bachelor and higher percentage, and advanced degree percentage, and the regression equation should look like the following:

$$PolarizationScore_i = \beta_0 + \beta_1 \widehat{HighSchool}Pct_i + \beta_1 \widehat{Bachelor}Pct_i + \beta_1 \widehat{Advanced}Pct_i + u_i$$

This regression yields the results in the following table.

To interpret this, first, the intercept does not have any economic meaning. Because it does not make sense to have a country that has a percentage of high school education equal to 0- at least someone in the country is somewhat educated. So, I will focus on the other three variables.

Second, looking into the variable High School and Higher Percentage, which indicates the user's basic education level, the coefficient is slightly negative at -0.014. This result says that countries with more educated people tend to be less opinionated. And thus, to refer that more educated people tend to be more open and inclusive in what they say. Concerning the research question, we would say that: the percentage of people having high school diploma can well forecast a country's polarization level so that when the percentage of high school diploma increase by one percentage point, the country's polarization score tends to decrease, on average, by 0.014 points. 0.014 point is not economically significant considering the base range of polarization score is from -100 to 100, but this number is highly statistically significant.

Third, looking into the variable Bachelor and Higher Percentage, which indicates the user's intermediate education level, the coefficient is slightly negative at -0.108. This result says that countries with more educated people tend to be less opinionated. And thus, to refer that more educated people tend to be more open and inclusive in what they say. Concerning the research question, we would say that: the percentage of people having high school diploma can well forecast a country's polarization level so that when the percentage of bachelor's degree increase by one percentage point, the country's polarization score tends to decrease, on average, by 0.108 points. 0.108 point is not economically significant considering the base range of polarization score is from -100 to 100, but this number is highly statistically significant. Notice that compared to the percentage of bachelor's degrees, the percentage of bachelor diplomas has a larger negative effect on the Y variable. Thus, people who get higher education tend to be more careful and less judgmental about what they say. This result makes sense because people are trained to think critically and be more inclusive in the higher education environment.

Fourth, looking into the variable High School and Higher Percentage, which indicates the user's advanced education level, the coefficient is slightly positive at 0.137. This result says that countries with more educated people tend to be more opinionated. And thus, to refer that more educated people tend to be less open and inclusive in what they say. Concerning the research question, we would say that: the percentage of people having advanced degrees can well forecast a country's polarization level so that when the percentage of advanced degrees increases by one percentage point, the country's polarization score tends to increase, on average, by 0.137 points. 0.137 point is not economically significant considering the base range of polarization score is from -100 to 100, but this number is highly statistically significant. Notice that this conclusion contradicts what we inferred from high school and bachelor's degree diplomas. This could imply that people with an advanced degree tend to be more confident and surer of what they say - thus, they tend to be less open and more polarized.

To sum up, this model shows that education levels have a predictive effect on the polarization score. The prediction says that people with high school and bachelor's degrees tend to be less polarized, and people with more advanced degrees tend to be more polarized.

	Dependent variable:polar_score		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Intercept	7.412***	7.323***	9.533***
	(1.546)	(1.558)	(1.699)
High School and Higher Percentage	-0.012	-0.009	-0.014
	(0.017)	(0.018)	(0.018)
Bachelor and Higher Percentage		-0.004	-0.108***
		(0.008)	(0.033)
Advanced Degree Percentage			0.137***
			(0.042)
Observations	265,684	265,684	265,684
R^2	0.000	0.000	0.000
Adjusted R^2	-0.000	-0.000	0.000
Residual Std. Error	28.407 (df=265682)	28.407 (df=265681)	28.407 (df=265680)
F Statistic	0.441 (df=1; 265682)	0.327 (df=2; 265681)	3.752** (df=3; 265680)
Note:			*p<0.1; **p<0.05; ***p<0.01

>>> **Regression on Country/Economy Attributes**

For the last regression, I am going to look at country/economy attributes. These economic attributes are important because economic conditions sometimes determine not only the environment that people live in but also determine how people think and talk. The difference in those factors can largely predict or indicate the overall people's polarization. Thus, a muti-regression on the general economic variables is needed to infer some economic-specific relationships relating to the polarization score(Y variable). Thus, I will choose the following three variables: population, GDP estimate, and racial diversity; and the regression equation should look like this:

$$PolarizationScore_i = \beta_0 + \beta_1 \widehat{Population}_i + \beta_2 \widehat{GDPpcPPP}_i + \beta_3 \widehat{DiversityScore}_i + u_i$$

This regression yields the results in the following table.

To interpret this, first, the intercept does not have any economic meaning. Because it does not make sense to have a country with a population of 0 - a country should at least have one person. So, I will focus on the other three variables.

Second, looking into the population variable, which indicates if a country is crowded and complex, the coefficient is rounded to 0 with three decimal places. This result says that a country's population does not predict an increasing or a decreasing trend in the polarization score. Concerning the research question, we would say that: the population can well forecast people's polarization level as the coefficient is highly statistically significant; however, there doesn't seem to be a straightforward linear and numeric relationship between those two variables.

Third, looking into the variable GDP estimate, which indicates the country's income level and development level, the coefficient is rounded to 0 with three decimal places. This result says that a country's income and GDP do not predict an increasing or a decreasing trend in the polarization score. Concerning the research question, we would say that: income can well forecast people's polarization level as the coefficient is highly statistically significant; however,

there doesn't seem to be a straightforward linear and numeric relationship between those two variables.

Fourth, looking into the variable diversity score, which indicates the country's diversity level, the coefficient is slightly positive at 4.34. This result says that countries with more diver environments tend to be more opinionated. Concerning the research question, we would say that: the diversity score can well forecast a country's polarization level, so when the diversity score increases by one point, the polarization score tends to increase, on average, by 4.34 points. 4.34 points is economically significant, and this number is also highly statistically significant. This interpretation aligns with previous conclusions. The potential qualitative reasoning behind this could be the cultural conflict. Because in more diversified countries, people have different backgrounds and cultures. In some cases, especially for political elections, those cultures might lead to different ways of thinking - thus, different conclusions. When there is disagreement among some cultures, opinionated debates tend to grow. People argue to prove their way of thinking is of the right logistics. When people get more opinionated and less welcoming, they score a higher polarization score. To contrast, in those who live in less diverse countries, people have the same culture and the same way of thinking. Their living habits and conventions are almost alike. Thus, disagreements, arguments, fights, and debates are less likely to happen in less diversified countries, so those people will score lower on their polarization level.

To sum up, this model shows that diversity levels have a predictive effect on the polarization score. The prediction says that people living in more diversified countries tend to be more polarized.

	Dependent variable:polar_score		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
Intercept	5.568***	5.904***	3.310***
	(0.061)	(0.129)	(0.197)
Population	-0.000***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)
GDP Estimate		-0.000***	0.000***
		(0.000)	(0.000)
Racial Diversity			4.389***
			(0.251)
Observations	222,014	222,014	222,014
R^2	0.000	0.000	0.001
Adjusted R^2	0.000	0.000	0.001
Residual Std. Error	24.781 (df=222012)	24.780 (df=222011)	24.763 (df=222010)
F Statistic	18.186*** (df=1; 222012)	13.475*** (df=2; 222011)	110.540*** (df=3; 222010)
Note:			*p<0.1; **p<0.05; ***p<0.01

>>> IV Regression for Causation Analysis

At this point, a different topic is looking into potential casual relationships by running an IV regression model. Even though this paper talks about predictive effects, it would be fun to look at IV regression results. Those results can further validate our conclusions above about the predictive effects.

The regression will focus again on how education influences the polarization score. However, unlike the model above, we will introduce an instrumental variable, the GDP estimate. The reason for choosing GDP is that: income is highly correlated with the level of education levels - richer people tend to have better access to better education. Controlling income can be very effective in referring casual relationships for education levels. Thus, to conclude, we will focus on X variables: high school diploma and above, and an instrumental variable, GDP estimates.

The result of the regression is attached below. The below output implies a positive relationship exists between high school population proportions and the polarization score. However, the T-stat is close to zero, and P-value is very big. Meanwhile, the F-test is also very small. All these testing numbers together imply the relationships this regression shows are not statistically significant. Thus, we should not refer meaningful conclusions from this result. And we should not infer this positive relationship truly exists. Moreover, because this model is an IV regression model, we can refer to more information. With the testing results and coefficients, we fail to infer any causal effects. More specifically, after controlling for people's income, we cannot say that the education people received made them more or less polarized.

Finally, from the results we have up till now, including the IV regression and the OLS regressions, we can only conclude that education levels alone do have a predictive effect on people's polarization score. However, this effect is not casual. Thus, again, in this paper we will only focus on predictive effects. And in the machine learning below, I will focus on predictions rather than causations.

Table 1: IV-2SLS Estimation Summary

Dep. Variable	High School and Higher Percentage			
R-squared	-1.945e+08			
Adj. R-squared	-1.945e+08			
F-statistic	0.0005			
P-value (F-stat)	0.9820			
Distribution	chi2(1)			
Estimator	IV-2SLS			
No. Observations	129551			
Date	Wed, Apr 05 2023			
Time	16:20:27			
Cov. Estimator	unadjusted			
	Parameter	**Std. Err.**	**T-stat**	**P-Value**
const	-99.659	4457.8	-0.0224	0.9822
polar_score	14.449	640.64	0.0226	0.9820

>>> Machine Learning and Regression Tree

In the following section, we will focus on machine learning. Specifically, we are going to draw a regression tree to show the potential distribution and predictions of the data. A regression tree works by splitting the data into subsets that only contain similar target values. Then, it continues to split data into sub-categories for further prediction and data fitting. Finally, each node in the tree represents a prediction for the Y variable, which is typically the average value of the Y variable. Now, we will look at the variables and parameters chosen before diving into the model.

For this regression tree, I will choose the variables: diversity score, created day, and user join year. The reason for these three variables is that: those three variables showed both economically and statistically significant relationships to the polarization score. Thus, it makes

sense to look further into those three variables and how they make combined predictions on the polarization score.

One parameter is alpha shown in the minimization equation. Alpha controls the allowance for error in the regression tree model as a penalty factor. With a larger alpha, the tree will be smaller. In this paper, we will not adjust the alpha as it's complicated and hard to adjust.

For the regularization parameter, in this case, the max depth is set to 3. Max depth is a regularization parameter that determines the maximum number of splits a tree can have. If we are to decrease the number on the max depth, we are losing precision and may not predict the numbers well. If we are to increase the number on the max depth, it would further increase the MSE, resulting in a less precise model - which could also over-fit the data. Thus, a max depth of 3 is chosen in this case. With the above information, the objective function of a single split is shown below:

$$\min_{tree \subset T} \sum (\hat{f}(x) - y)^2 + \alpha |\text{terminal nodes in tree}|$$

Thus, taking into consideration of the variables chosen, and in conclusion, the overall regression tree formula is shown below. Particularly, this function tries to solve and get the minimum MSE at each split according to the three variables we have chosen.

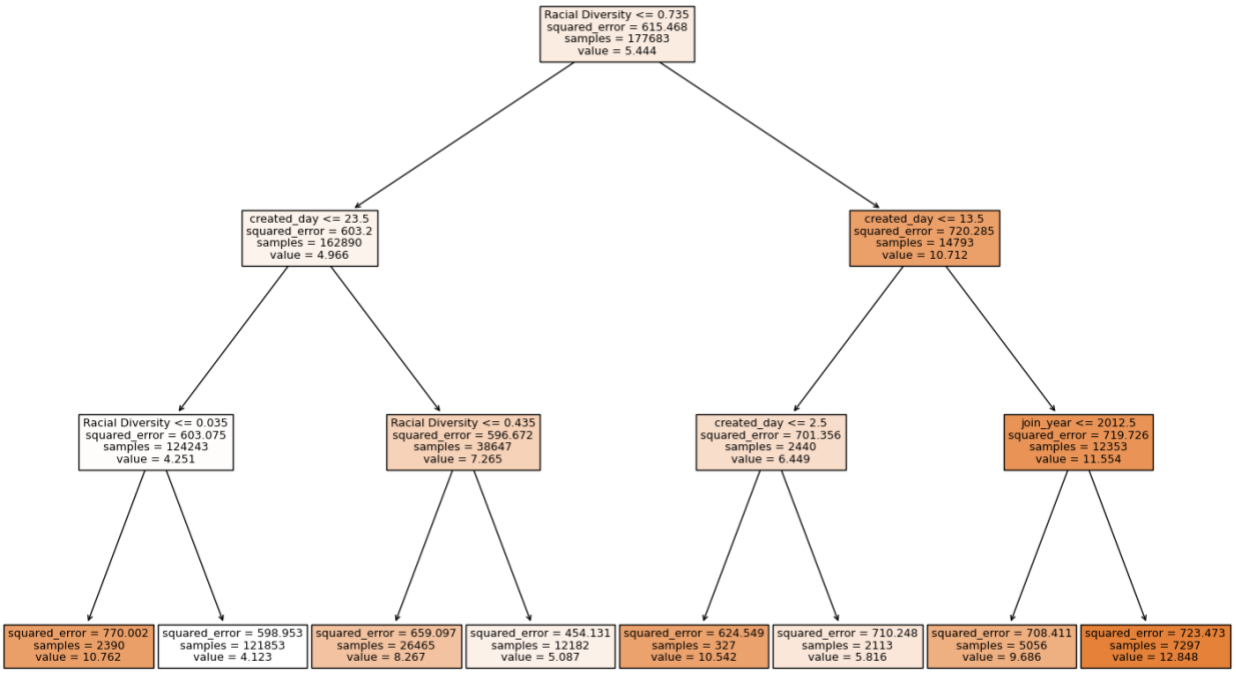
$$\min_{j,s} \left[\sum_{i:RacialDiversity_{i,j} \le s, x_i \in R1} (y_i - \hat{y}_{R1})^2 + \sum_{i:RacialDiversity_{i,j} > s, x_i \in R2} (y_i - \hat{y}_{R2})^2 \right], \left[\sum_{i:CreatedDay_{i,j} \le s, x_i \in R1} (y_i - \hat{y}_{R1})^2 + \sum_{i:CreatedDay_{i,j} > s, x_i \in R2} (y_i - \hat{y}_{R2})^2 \right],$$
$$\left[\sum_{i:JoinYear_{i,j} \le s, x_i \in R1} (y_i - \hat{y}_{R1})^2 + \sum_{i:JoinYear_{i,j} > s, x_i \in R2} (y_i - \hat{y}_{R2})^2 \right]$$

Using this objective minimization objective function, we can form a regression tree that is shown below. In the regression tree, from observations, we can tell that the model splits the data in the following sequence: racial diversity, created day, then user join year. Firstly, the data was divided by diversity score at a threshold of 73.5% with a total of 0.22 million samples. This first split gives a prediction of a polarization score of 5.4. Then, the samples are split by the tweet created day at a threshold of 23.5 and 13.5. The split gives predictions of polarization scores of 4.99 and 10.5, respectively. Then, the model goes into the third split and the predictions for those splits. The model stops at our max depth, which is three splits.

In terms of error of predictions, the MSE is 24.7, which is acceptable considering that the Y ranges from -100 to 100. However, this error is not ideal. An ideal MSE would be smaller than 10%, which in this case is smaller than 20(100-(-100) = 200, 200*0.1 = 20). The reason for a bigger MSE could be multiple reasons. First, it might be the sample size that is too big. Because we have too much variation within the sample and more than 0.2 million observations in this case, a bigger-than-normal MSE is reasonable. Second, as we introduced above, one X can have multiple corresponding Y(people in Canada share the same number for population number, but people in Canada have different polarization scores, so one population number corresponds to multiple Y). Thus, this effect will also make the MSE larger. A solution to this problem is to gather data on states instead of countries to narrow down the focus - but it will be very hard to find population and GPD data for all states in all countries in the world.

Meanwhile, the reddish the box is, the higher the polarization score predictions. Thus, we can see that the split to the right has a reddish box. This box would indicate that higher racial diversity(first split), closer to election day(higher created day in the second split), and fewer previous internet exposure(higher join year in the third split) will predict the highest polarization score. This conclusion matches what we had in the regression models above, where we also find the same predictive trend those variables have on the polarization score.

However, we can find more information on the regression tree that is unavailable on the normal regression results. In the above results, the OLS coefficient for those X variables is considerably small, but the differences between splits on the regression tree are big. For example, the OLS slope of "created day" is 0.168 points; the second split, also for "created day," has a difference of about 5 points. This result means that after splitting and grouping those data, we can easier and more accurately identify the predictive trends. The potential intuition behind this phenomenon is that we have too many observations lying at low racial diversity scores. Thus, in an OLS regression, the slope will be drawn by the mass of observations at a low diversity score so that the OLS slope will be smaller. Instead, in a regression tree, when we group by MSE, the mass observations are grouped together, so we can better tell the relationships along the change of X variables.



>>> **Short Regression and Machine Learning Summarization**

We've seen four models of different factors and variables and a regression tree that validated the results of the OLS models. Each of these models has shown predictive effects of X variables on the polarization score. Thus, it's important to do a quick summarization here.

Overall, we say that diversity levels, education levels, user's internet exposure, and the time to the election can well predict people's sentiments and polarization, while other terms do not show economically significant relationships.

7. Conclusion

In this paper, we've investigated the several most important variables and how they affect people's polarity by conducting sentiment analysis. And again, as mentioned before, this paper is to work on an overall sentiment and polarization analysis during the 2020 election by researching

different topics of variables, such as user-specific, education levels, economic-specific and tweet specific. Thus, we would say that economic terms such as diversity score, education levels and created day can well predict polarization score, while some other terms do not show economically significant relationships. Specifically, we predict that: people with higher education, more internet exposure, people who are more diverse and more inclusive, and people who are away from major global events tend to be less opinionated and less polarized. While a short sentence cannot fully describe all significant relationships, preferably, we should be looking into those relationships and seeing the messages one by one. Only in this way can we have meaningful implementations of economic concepts and interpretations of this sentiment analysis.

From the initial visualizations, such as scatter plots and line graphs, we first glanced at the potential predictive effects of the X variables. Then, we summarized that Asian countries are more polarized, people with less internet exposure tend to be more polarized, and people get more polarized when it gets closer to the US election.

From the above graphs and maps, although some graphs contradict each other, we can still infer several important conclusions. The conclusions are: when it gets closer to the US election, people are getting more than more polarized; while the polarization score difference is not significant across the US, but comparatively, most high-scoring states are in the east of the US, and lots of low scoring states are in the south; different levels of education could cause the reason for that in different states; North American and African countries tend to be more opinionated than countries in other continents, this could be the effect of a more diversified environment; and that there are no significant relationships between fans owned and polarization score.

From the regression results above, each of the four models has shown predictive effects of X variables on the polarization score. And overall, we say that diversity levels, education levels, user's internet exposure, and the time of the election can well predict people's sentiments and polarization. At the same time, other terms do not show economically significant relationships. Specifically, people get more polarized when it's closer to election; people who are more educated are less polarized; and people with less internet exposure tend to be more polarized. Meanwhile, to step a bit forward and looking deeper, we tried to refer some casual relationships by running an IV regression, but we failed to prove any casual effects. Thus, we can also validate from this perspective that the relationships that we are observing are indeed predictions rather than causations.

Still, even though we believe the conclusions on the four X variables are well-supported, we must be aware of the potential existence of other lurking variables. We only controlled three variables for each regression, and this procedure helps with managing potential lurking effects. However, there might be other variables we may have yet to notice. Further research into sentiment analysis may help us identify the factors we should better control in a regression. Meanwhile, even though TextBlob is believed to accurately identify the polarization score on people's speeches, AI and natural language processing is still a developing field(Barai, 2021). Textblob model, in some extreme cases, can fail to identify human's sentiments accurately(Barai, 2021). Therefore, we should be more careful when processing those data. Finally, as we cannot access Twitter API at this time, we could not investigate the effects of the US election. However, soon, we expect Twitter to open their API systems with more analysis tools and better API services(Barnes, 2023). Then, we can gather new data and form further analysis on the effects the US election event.

In this report and dataset, we only observed the relationships to certain important X variables. In the future, we could explore how those polarization scores reflect people's emotions or living standards; and find some predictions between the five variables and people's emotional standards.

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). *Sentiment analysis of Twitter data*. ACL Anthology. Retrieved March 31, 2023, from https://aclanthology.org/W11-0705/

Barai, M. K. (2021, October 26). *Sentiment analysis with textblob and vader*. Analytics Vidhya. Retrieved March 31, 2023, from https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/

Barnes, J. (2023, February 6). *Twitter ends its free API: Here's who will be affected*. Forbes. Retrieved April 5, 2023, from https://www.forbes.com/sites/jenaebarnes/2023/02/03/twitter-ends-its-free-api-heres-who-will-be-affected/?sh=749cbb286266

Hui, M. (2020, November 9). *US election 2020 tweets*. Kaggle. Retrieved March 31, 2023, from https://www.kaggle.com/datasets/manchunhui/us-election-2020-tweets

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016). *Sentiment analysis of Twitter data for predicting stock movements*. Institute of Electrical and Electronics Engineers. Retrieved March 31, 2023, from https://ieeexplore.ieee.org/abstract/document/7955659

Siebeneck, T., & Wang, C. (2023). *GDP of US States*. GDP by State 2023. Retrieved April 5, 2023, from https://worldpopulationreview.com/state-rankings/gdp-by-state

Wikipedia, C. (2023, April 4). *List of countries and dependencies by population*. Wikipedia. Retrieved April 3, 2023, from https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

Wikipedia, C. (2023, February 16). *List of U.S. states and territories by educational attainment*. Wikipedia. Retrieved April 3, 2023, from https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_educational_attainment

Wikipedia, C. (2023, March 9). *List of countries by GDP (PPP) per capita*. Wikipedia. Retrieved April 3, 2023, from https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita

Wu, J., Muccari, R., & Zhou, N. (2020). *World watches the U.S. vote: Why the election matters everywhere*. NBCNews.com. Retrieved March 31, 2023, from https://www.nbcnews.com/specials/world-watches-us-vote-trump-biden-election/