Polarization and a Divided America

AN ANALYSIS OF HOW WE TALK ABOUT POLITICS USING TWITTER AND TWINT

Paul Gagliardi & James Duchesneau | Data Science Tools | 6/7/2021

Table of Contents

GitHub

This project's code and more info is featured on our Github here:

https://github.com/pogags/dataToolsFinalProj

Please look at this for more information

Introduction

As Americans, it can seem that we are more divided than we have ever been when it comes to politics. Anecdotally, the civil unrest of 2020 and January 6th riots at the capital point towards a public that is more polarized and more partisan, and to some extent angrier. In this study, we look to see if we can catch any objective clues that would allow us to show numerical proof that we are truly living in more divided times.

RESEARCH QUESTION

Have we become more polarized? Why is this important?

Knowing whether a constituent is in disarray is important for political leaders and concerned citizens alike. Only by examining our political and cultural realities and having an awareness of how our fellow citizens view issues will we be able to confront societal differences and mitigate political altercations.

For this study one of the best-known political battlegrounds, Twitter, will help us answer the question: "are we more polarized?"

TOPICS AND DATA

We examined tweets from November across five years (2016-2020) and five topics to examine how language has changed. The five topics we used were chosen because of their political significance as well as their tendency to be hotly debated and catalysts for polarization. The topics are:

- Abortion
- Education
- Gun Control
- Healthcare
- Immigration

Using "Twint" our team pulled ~500 tweet from each timeframe for each subject. Tweets were pulled in the time frame of 10/01/20xx to 11/08/20xx, however no tweets from October were used as the tweets were pulled in newest to oldest order and we were able to get the number of tweets we wanted quickly. The tweets were our only dataset, and they had a raw structure of the following variables.

- id (Tweet id)
- created_at (DateTime Creation MST)
- username (username of tweeter)
- tweet (text of the tweet)

We added two more variables to this after converting the CSV they were generated in to a pandas dataframe.

- Year (year of tweet)
- cleanTweet (tweet with @s and links removed)

Twint was open sourced and certainly not Twitter licensed, so getting the package to run correctly took a bit of work. It was thanks to the help of the Twint online community and forums on Github that we were able to overcome these challenges. However, once it was running correctly and able to grab the tweets we wanted, it was incredibly useful. We highly recommend this package for tweet pulls if it continues to be supported in the future. As part of the benefits of using Twint, we had **no missing values when pulling the data.**

METRICS

For this project, we examined two metrics and utilized a third visual representation strategy to try and pull together a picture of how the language over these topics have changed.

- Sentiment Analysis with NLTK:
    - The NLTK package comes prepared with a "Sentiment Analyzer" that will examine a sentence, and return a Positive, Neutral, and Negative sentiment amount.
    - The Positive, Neutral, and Negative scores must all add up to 1, and are scored on a 0 to 1 scale. This package is made specifically for short sentences and thoughts, such as Tweets.
    - The sentiment analysis displayed here will be an average across the tweets for that group.
    - More info: http://www.nltk.org/howto/sentiment.html

- Extreme Word Count:
    - A raw count of the "Extreme Words" used in tweets
    - Multiple occurrences of the same "Extreme Word" yield only count for one extreme word, but two or more different "Extreme Word" will contribute to more being counted.
    - Extreme words are as follows: 'fantastic', 'filthy', 'gorgeous', 'astounding', 'shocking', 'tiny', 'spotless', 'gaunt', 'shocked', 'terrible', 'delicious', 'hideous', 'huge', 'hate', 'ancient', 'brilliant', 'vivacious', 'esctatic', 'exhausted', 'awful', 'terrible', 'horrible', 'filthy', 'wonderflul', 'terrifying', 'furious', 'tiny', 'hateful', 'evil', 'shitty', 'fuck', 'shit', 'asshole', 'ass', 'cruel'

- Word Clouds
    - To take a snapshot of how conversations around these issues may have changed over the years, we created Word Clouds for each year and topic combo. For this paper, we will take a subjective and non-numeric approach in our analysis of these.

OTHER TOOLS

- Twint:
    - Twint is an open-source Python Package on Github. Due to the limitations of the Twitter API calls (for free you can only look back 7 days), an additional tool was needed. It searches for and returns tweets using the twitter search function native to their website.
    - More info: https://github.com/twintproject/twint
- Other Packages:
    - os, matplotlib, seaborn, numpy, pandas, re, pprint

OTHER LITERATURE

As far as similar projects, there does not appear to be any open access "polarization analysis" that has been done using "extreme words" as well as "sentiment analysis", and certainly not any we could find centered around specifically twitter.

However, we were able to find some similar studies on quantifying controversy around topics and looking to see if controversy affects conversation around issues, and they offer some interesting conclusions.

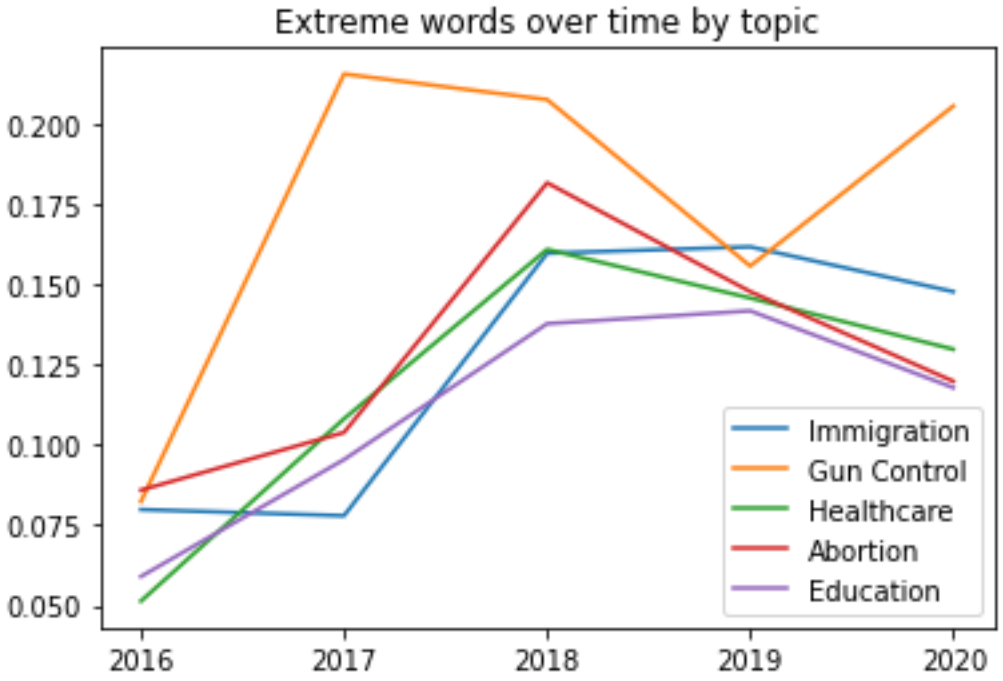
"When, Why, and How Controversy Causes Conversation", published in 2013 by Jonah A. Berger used field data, this study found that low levels of controversy generate more conversation, but at moderate or higher levels of controversy people are less likely to discuss the topic. Interestingly enough however, if the person you discussing the issue with is not your "friend" but rather a stranger, you are more likely to discuss controversial topics with them than a friend. This might explain why some people are more opinionated on Twitter than they would be over something like Instagram, as Twitter is king of the anonymous interactions.

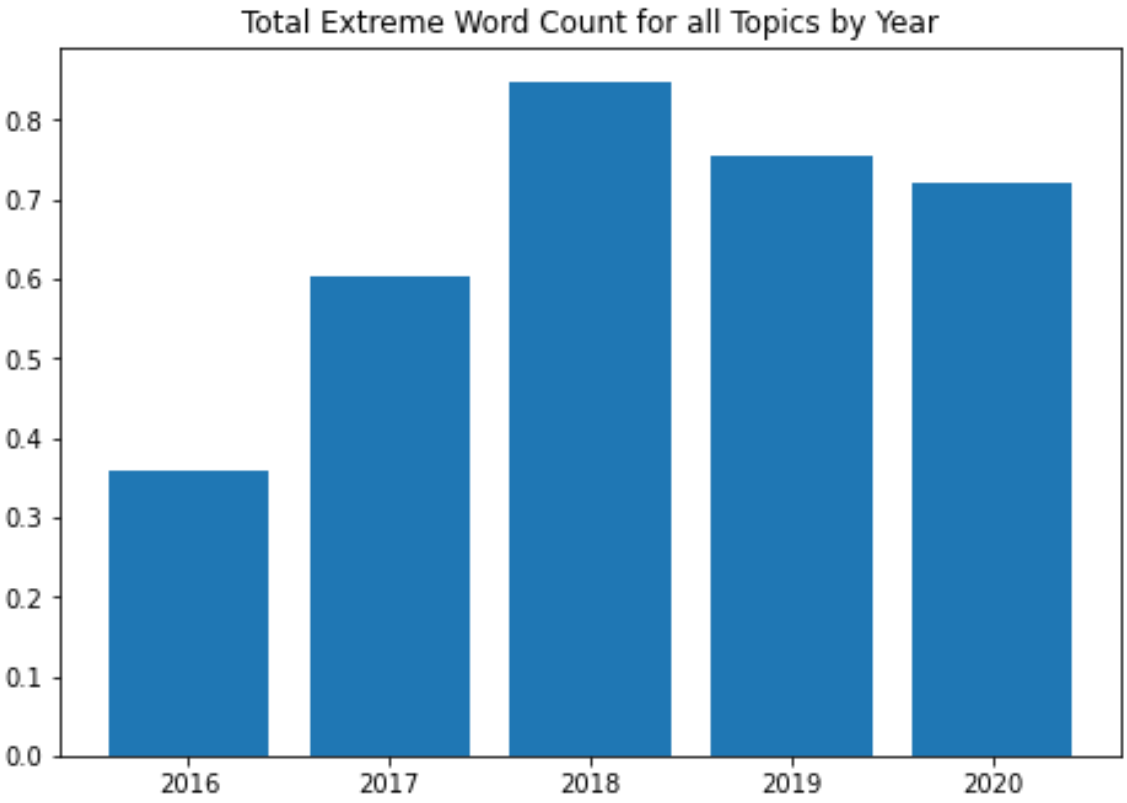
https://repository.upenn.edu/cgi/viewcontent.cgi?article=1342&context=marketing_papers

"Quantifying Controversy on Social Media" is an Aalto University study from 2017, and it deals heavily with social media. The investigation into hashtags provides an interesting look at how trends and hashtags an indicator of a controversial topic. The paper also, similar to us, focuses on social media. The math surrounding the assignment of controversy values to topics is quite formidable and impressive, highlighting the level of complexity that surrounds human interaction. This paper also warns of political "bubbles" and "echo chambers", and encourages social media users to try to interact with social media content outside of their bubble.

https://arxiv.org/pdf/1507.05224.pdf

All Topics


Extreme words over time by topic
0.200
0.175
0.150
0.125
0.100
0.075
0.050
2016
2017
2018
2019
2020
Immigration
Gun Control
Healthcare
Abortion
Education


Total Extreme Word Count for all Topics by Year
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2016
2017
2018
2019
2020

EXTREME WORD COUNT – ALL TOPICS

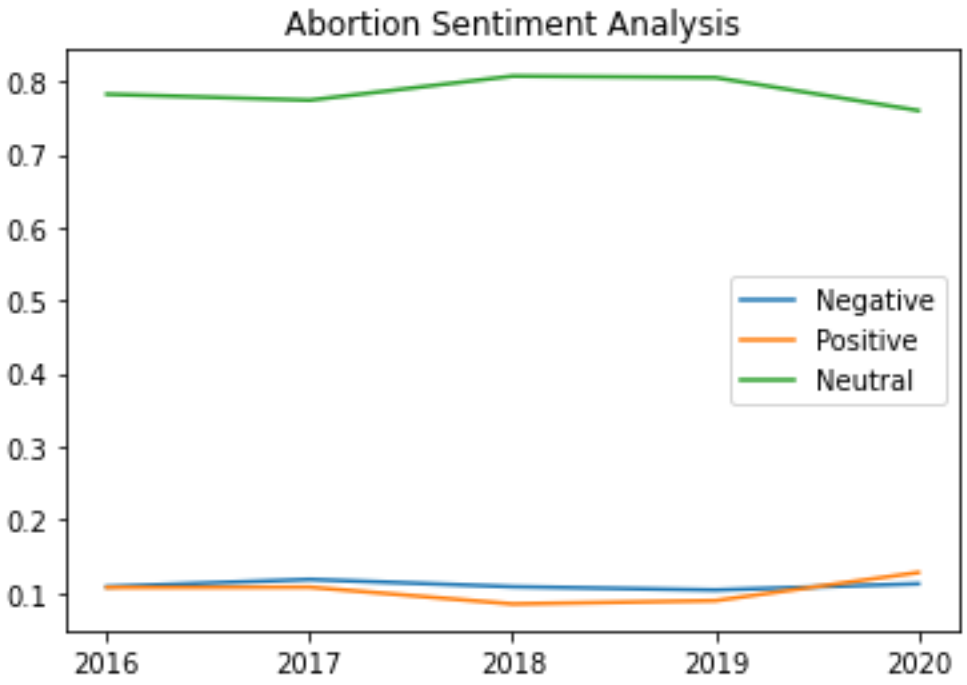
In general, the extreme word count for all of the topics increases significantly since 2016 (~.37% of tweets pulled had extreme words). In 2018, the highest percentage of extreme words per tweet year, had extreme words about 83% of the time on average for the tweets pulled.

Conclusions

With such a large swing in the number of extreme words being used, we can conclude from this data set that **more extreme language has been used over time**, indicating we have indeed become more polarized

Abortion

SENTIMENT ANALYSIS

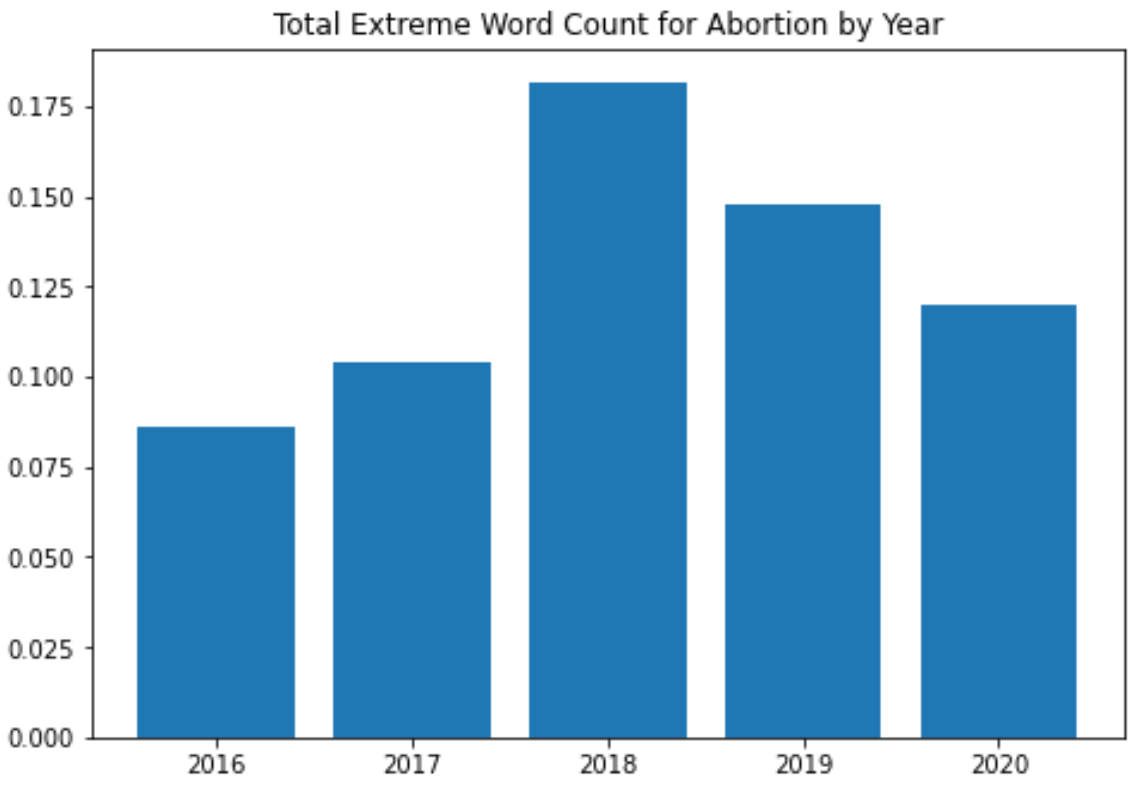


- Sentiment around the topic of abortion became **less neutral**, however this was not a steady drop
- Sentiment also trended steadily **more negative,** but only grew from 0.108 to 0.112
- There was a large **spike in Positive sentiment in 2020**, however no real trends

Conclusions

It is possible that around elections, this issue becomes more important, as neutrality dipped around 2020 and was higher generally in non-election years. Based on this data, we **cannot draw any solid conclusions** about whether we are more polarized around abortion.

EXTREME WORD COUNT - ABORTION

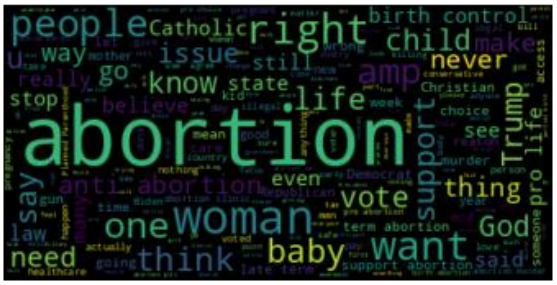


Conclusions

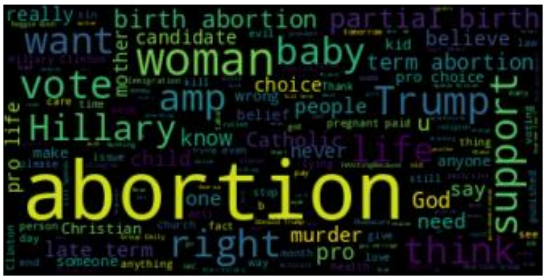
Similar to the all topics trend, and the trend seen for other individual topics, the extreme word count percentage increases after 2016 and the highest year is 2018. It will be curious to see if going forward from 2020 if the trend continues to go down like the bar plot is showing of if it evens out.

WORD CLOUD ANALYSIS

All Years



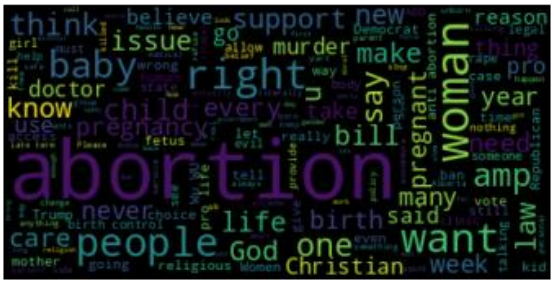
2016



2017

2018

2019



2020

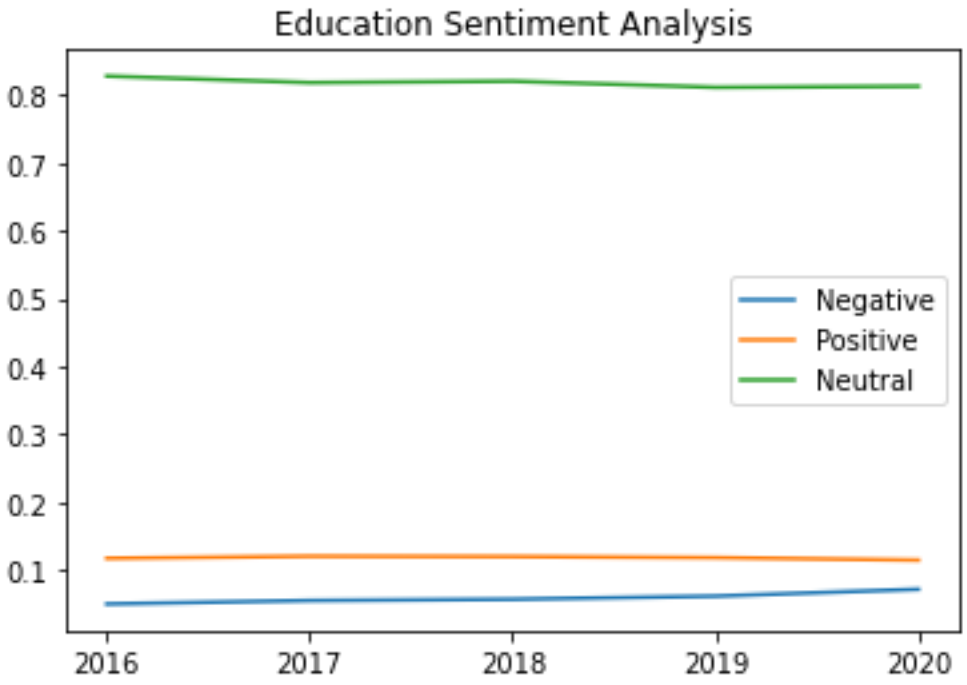
Hillary was featured heavily in 2016, but after the election that year she expectedly does not make another appearance. **Trump** meanwhile makes an appearance in all sections, and **Biden** pops up in 2020.

This is still a conversation centered around **woman** and women's **rights**, and both those words make appearances on all charts as. **Christian, Catholic, and Religious** are also featured, showing that much of the debate on this is framed around religious concerns.

Some additional words that seemed noteworthy were **murder**, likely in the context of calling abortion murder and **birth control** which is another women's right's concern.

Education

SENTIMENT ANALYSIS

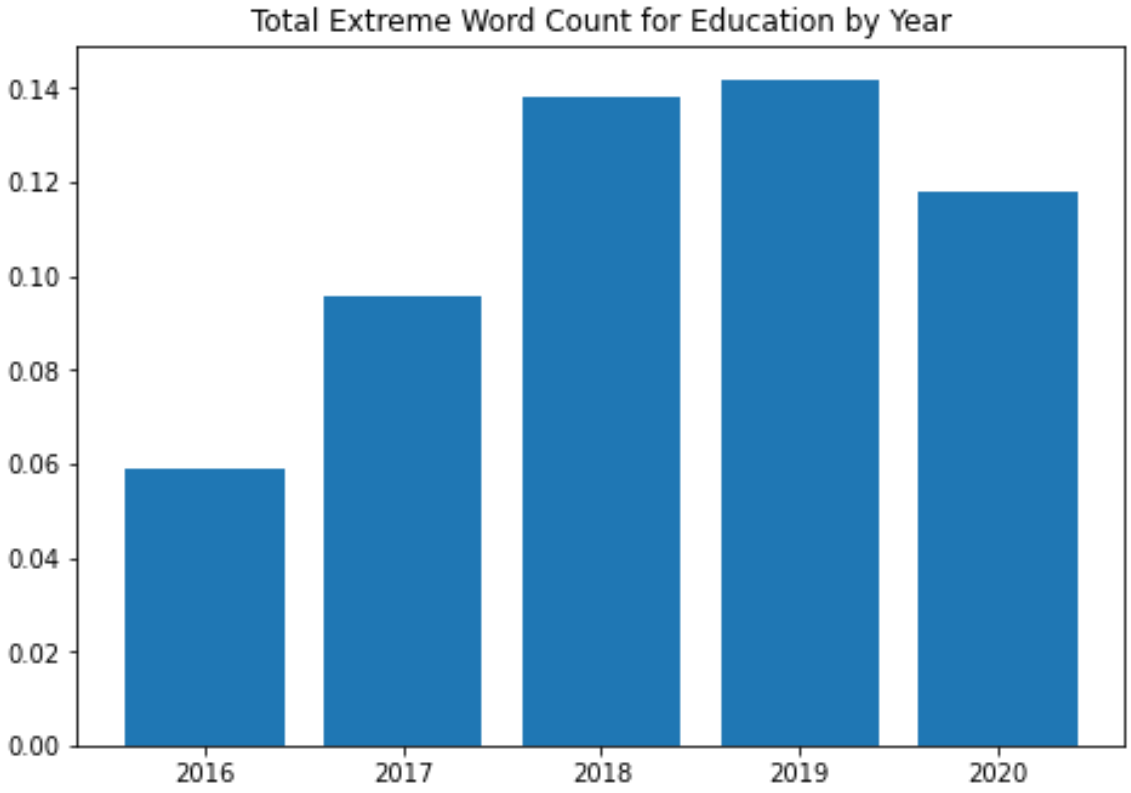


- Sentiment became **more negative** every year, and changed from 0.051 to .072 in the 5-year span, more than a **42% increase**
- **Positive sentiment** decreased slightly, but did not have much change

Conclusions

Education, the topic with the highest average positive sentiment, trended very negatively in this study. We can conclude with some confidence from this data that **people have grown more divided on education.**

EXTREME WORD COUNT - EDUCATION



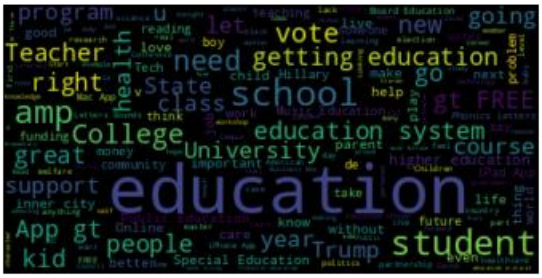
Conclusions

For education, a slightly different trend is observed where the percentage of extreme words increases after 2016 but instead of 2018 being the highest year, 2019 is. Also observed, the years 2018-2020 are all of similar percentage. It would be interesting to see going forward, if this trend for Education evens out or continues to decrease.

WORD CLOUD ANALYSIS

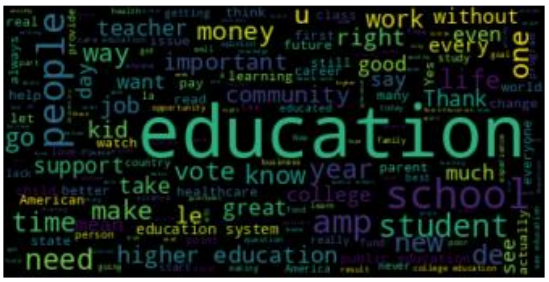
All Years

2016

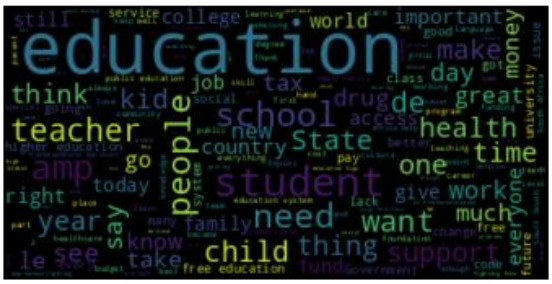


2017

2018



2019



2020

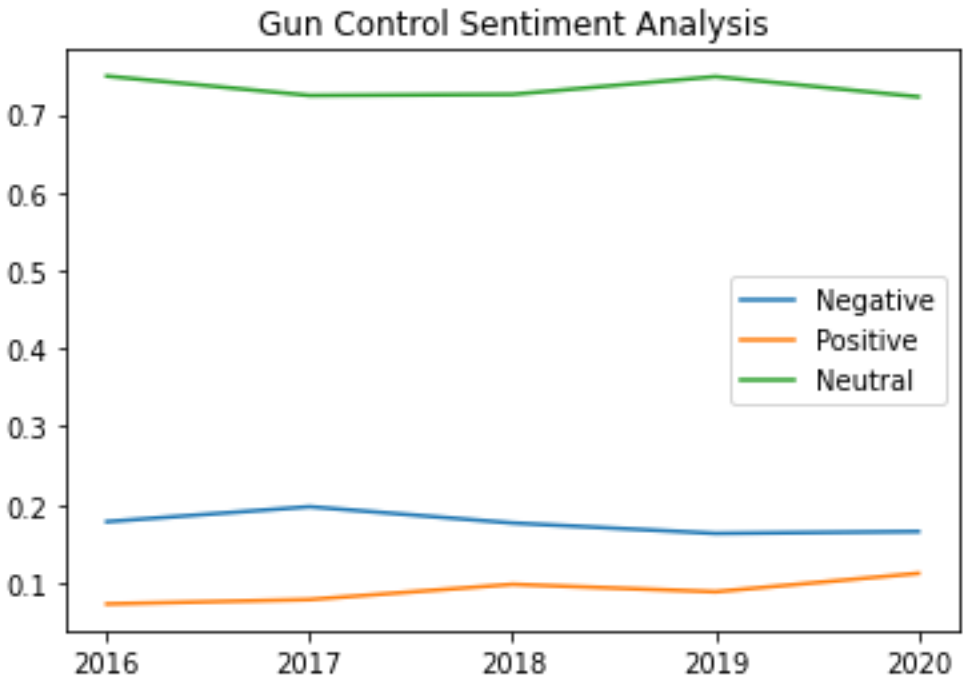
Higher education and **college** seem to be consistent points of contention when it comes to this topic. The cost of a higher education has long been a political issue in the US, and the debate around this has heated up in recent years. Additionally you will see **pay**, **free** and **money**, though pay and free are featured in the "all years" cloud and money is not.

Trump did not get as much love this topic as he did in others, featuring prominently only for the 2016 and 2020 clouds (election years). Meanwhile, **teachers, kids, student**, and **child** are abundant, and featured in the "all years cloud". The connection of this topic to American children could be why some are so passionate about it.

Some other interesting words were **education system** indicating people were talking about the project of American Education as a whole, **Secretary Education** in 2020 (likely centered around the election), and **love** which is difficult to contextualize.

Gun Control

SENTIMENT ANALYSIS



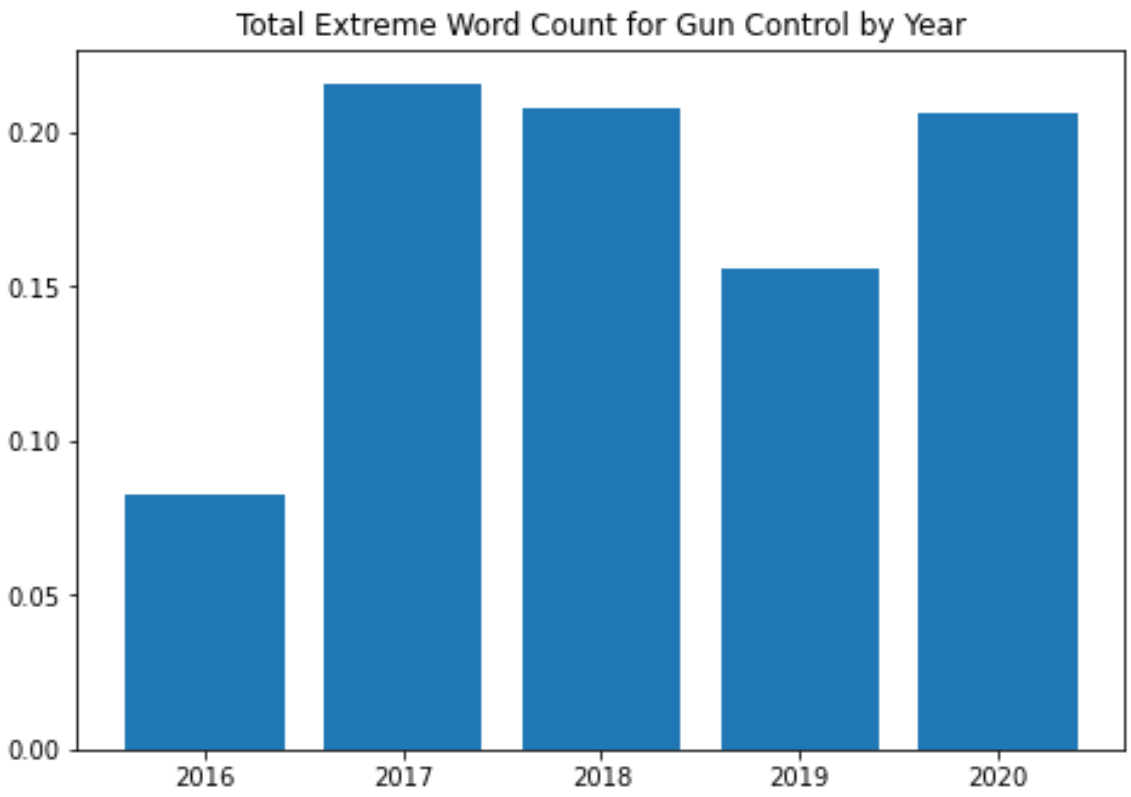
- **Negative sentiment** surprisingly **dropped steadily** for gun control after peaking in 2017. In 2016 negative sentiment was rated at 0.178, and by 2020 it was 0.166
- Equally as surprising, **positive scores increased quite drastically**, from 0.073 to 0.112, a **53% increase**

Conclusions

What do we make of a positive score increase? Well in this context, it can mean people are talking more positively about the idea of gun control. It could also mean people are talking more positively about guns, or aspects of gun control.

However, a 53% increase is significant. Based off this data, the best conclusion we can draw is that guns and gun control are getting talked about with more positive words, and likely less aggressively, so we could say with some confidence that America is **less divided on this topic, or at least more in step with positive sentiment on the topic.**

EXTREME WORD COUNT – GUN CONTROL



Conclusions

Gun Control had one of the most interesting trends overall with an increase after 2016 but continuing to stay at the raised level from 2017-2020. Gun Control is one of the most controversial topics in government and politics, so it is not surprising that it has a continuously high percentage of extreme words. I would expect this trend to continue into the future, especially with it seeming like there is a mass shooting more and more often.

WORD CLOUD ANALYSIS

All Years

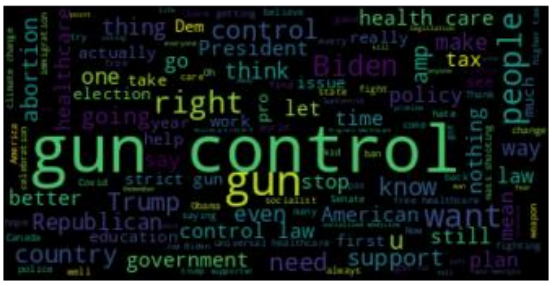
2016

2017

2018

2019

2020

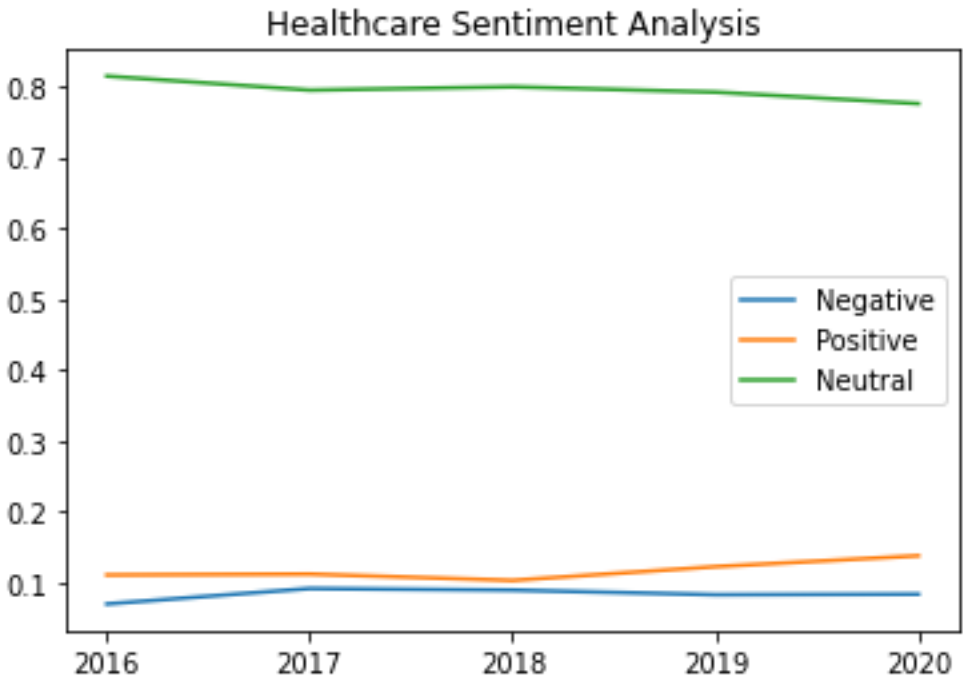


Trump is featured in all clouds for this topic, and he is joined by **Bloomberg** in 2019 (a Democratic candidate with focus on gun control who ran in 2020) and **Biden** in both in "all years" and 2020. **Democrat and Republican** both feature heavily, indicating that this is indeed a partisan issue among the two parties. The **NRA** only make a couple appearances (2016 and 2017).

Need Gun, featured on the "all years" cloud, shows that many people either feel that they need a gun, or are claiming that other people do. **Control Legislation** and **Control Law** show that many people are talking about structural changes, positive or negative, having to do with gun control. In 2017, **Texas Church** and **Texas** are featured, reminiscent of the Sutherland Springs church shootings that occurred there on November 5th, 3 days before these tweets were pulled; often gun control debates rage right after mass shootings.

Healthcare

SENTIMENT ANALYSIS



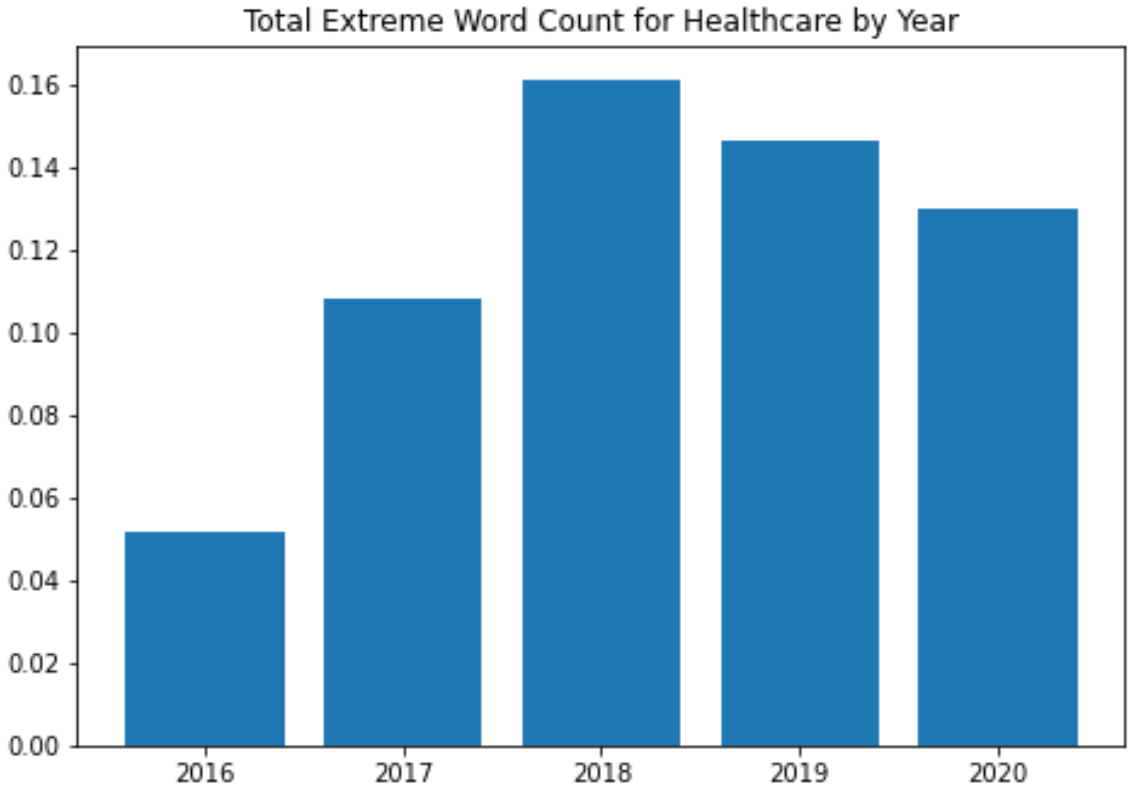
- Sentiment around the topic of healthcare became **less neutral** by the end of the 5 year period, dropping from a neutral score of 0.814 in 2016 to 0.776.
- Sentiment also trended steadily **more negative,** moving from 0.07 to .084
- **Positivity also grew overall** though not as steadily, from .111 to .138

Conclusions

We seem to have become more divided on the topic of healthcare. Neutrality overall had a large drop, and in conjunction with that more people are speaking negatively/positively about healthcare. Examples of negative sentiments in this case could be "my healthcare is bad" or "everyone whines about healthcare". Positive sentiment could be "we need good healthcare" or "I am good without healthcare". However, all four of these examples show people expressing opinions and displaying a less neutral stance.

Therefore, we can conclude from this data that people are **more divided, or at least more opinionated**, on healthcare.

EXTREME WORD COUNT - HEALTHCARE



Conclusions

Healthcare has a similar barplot to the all topics barplot with it increasing after 2016 and having 2018 as the highest year. With the attempt to repeal Obamacare in 2017, it does not surprise me that the trend of extreme words in the tweets pulled kept increasing in 2017 and continued forward. The plot does trend down after 2018 but unless everyone is happy with healthcare I don't see this trend decreasing much more as years go on.

WORD CLOUD ANALYSIS

All Years

2016

2017

2018

2019

2020

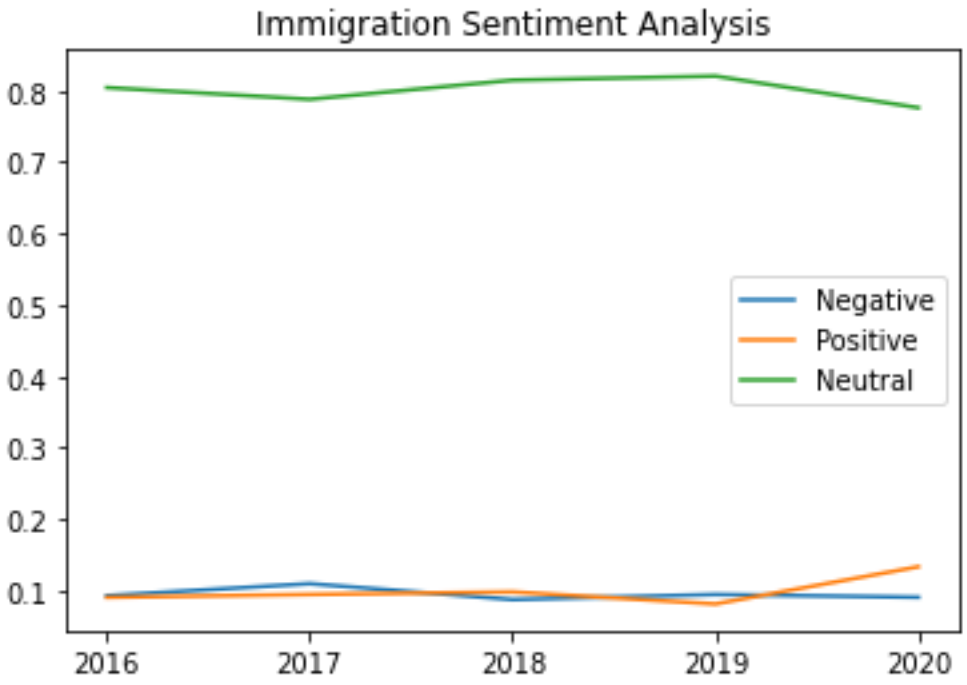
Immediately with the topic of healthcare, it is obvious that **abortion** is also featured heavily. Many consider **abortion** to be a healthcare issue. **Woman** is also featured. **Trump** makes it onto the "all years" cloud, but doesn't feature as heavily in clouds besides 2016 and 2020. **Obamacare** also makes a 2016 appearance but does not return as with **Hillary. Biden** shows up in 2020.

As with education, it seems **money** and **access** to healthcare is a core topic here. **Tax, free, money, pay,** and **right** (as in right to healthcare most likely) all feature in most clouds. **Universal healthcare** sees some mentions in 2019 and 2020, when it was being discussed as a policy idea for Bernie Sanders.

Some other interesting words include **baby** which could be linked to the high cost of having children, and **murder** which is certainly a health issue. **Healthcare workers** had a tumultuous 2020, and they seem to get some love in the 2020 cloud as well.

Immigration

SENTIMENT ANALYSIS

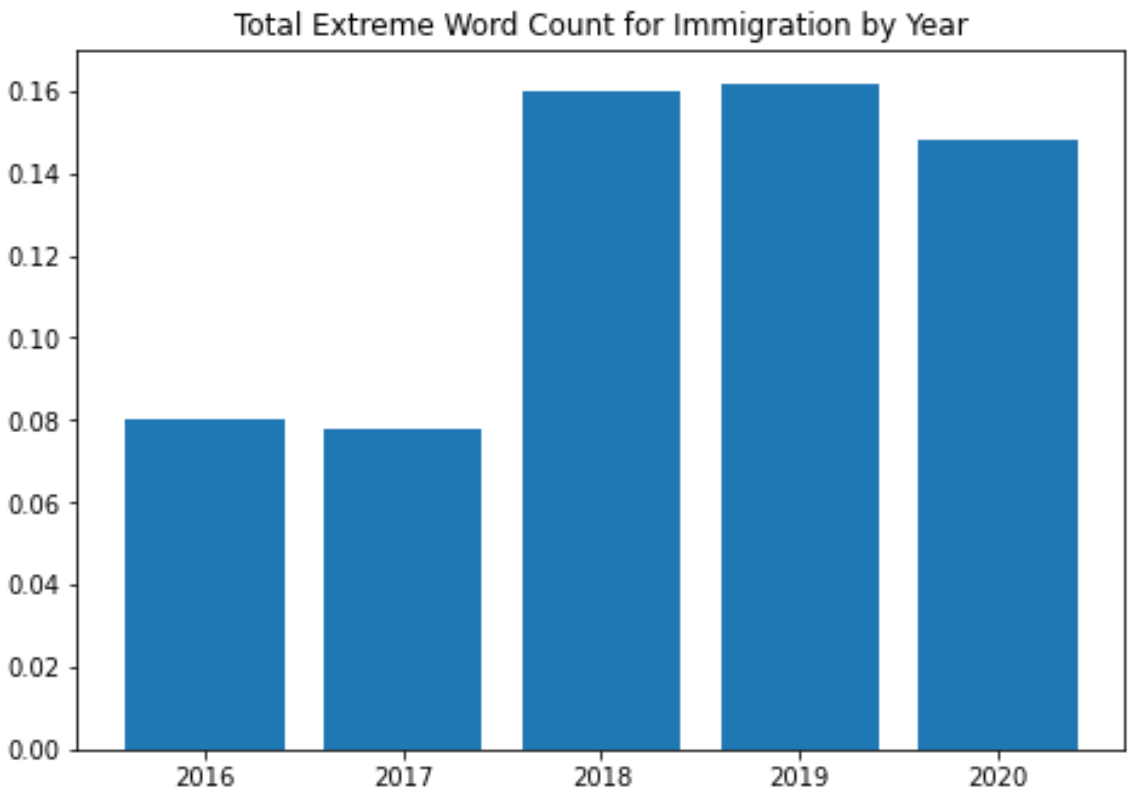


- **Neutrality overall trended down,** however it had its two highest values in 2018 and 2019. It dropped from 0.804 in 2016 to 0.776 in 2020
- **Sentiment little to no negative change,** moving from 0.092 to .091
- **Positivity was very steady until 2020.** In that year it grew from .081 to 0.133

Conclusions

We cannot draw any solid conclusions from this data, as it does not seem to have any strong trends.

EXTREME WORD COUNT - IMMIGRATION



Conclusions

Immigration has been a hot topic in recent years and this trend is interesting because it coincides with a political decision by Trump to build a wall on the mexican border at the end of 2017 so by 2018 and on it was an extremely controversial topic in America. Also with covid and the borders being shut the trend continued to be controversial and therefore a higher extreme word percentage in the tweets pulled.d a

WORD CLOUD ANALYSIS

All Years

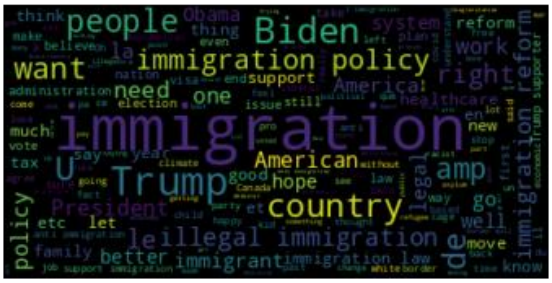
2016

2017

2018

2019

2020



When people discuss the topic of immigration, especially in the United States, in all probability they discussing **illegal immigration** which pops up in every cloud**.** First and foremost **Trump** is also featured who famously began his campaign with a controversial **immigration policy** which is also featured in all clouds. Indeed, **different visions** is featured in 2016, likely a result of the **vastly different** approaches the candidates had toward immigration.

Law and **immigration law** are also heavily featured, as this issue is about how the United States Government is legally obligated to treat immigrants. **Canada** is also featured here, indicating that while the southern US border is one of the hottest contested topics of immigration it is not the only one.

Conclusion

Based on the data we collected, and specifically focusing on the extreme word counts, it is hard to deny that there is quality evidence for continuing polarization among the American public, at least when it comes to the politics around the topics of Abortion, Education, Gun Control, Healthcare, and Immigration.

With the exception of Abortion, the extreme words used in tweets doubled or more than doubled from 2016 to 2020. Almost uniformly, we saw the most extreme words being used in 2018, and having a slight drop as it moves toward 2020, but maintaining levels well above 2016 and 2017.

We do not have as strong evidence for continued division when examining sentiment analysis, but we do see some. Negative sentiment on Education is way indicating more division and passion in that area (educations extreme words also doubled) and we saw much less neutrality and more opinionated opinions surrounding the area of healthcare (extreme words more than doubled here).

Combining these observations, we can confidently conclude that Healthcare and Education show strong evidence that they are quickly becoming more controversial/polarizing, whereas Gun Control, Immigration, and Abortion have shown decent evidence for becoming more controversial/polarizing.

Some limitations of this study were that we were not able to get as many tweets as we wanted due to fear of getting an IP address banned by twitter, and our twitter sampling as such was not as random as it could have been. Rather than get a random sampling of all tweets over a month, all tweets were pulled from the newest to oldest. If a day had a specific political event, then that could have skewed our data.

This project has highlighted the continued division of American among political lines. Though each human is indeed entitled to their own opinion and thoughts, we must be careful that we do not allow our differences to define us more than our similarities. The next political century will see us more connected and as such, able to fight or come together more than ever before. It will be each of our responsibilities to try and handle this brave new world with as much kindness and empathy as possible, and to avoid polarization and division where possible.