ABSTRACT

This assessment deals with “DATA7201 Data Analytics at Scale”. Given a dataset, I used big data analytics techniques to explore the data and to draw some conclusions using pyspark and Hadoop. I have the selected appropriate techniques, imported libraries and justified my choices using supporting evidence from academic literature.

I have written a structured report that describes the approach I have taken to analyse the chosen dataset using big data analytics techniques and presents my main findings. The dataset to be used in this assessment is a collection of sponsored political posts on Facebook targeted at US users during 23 months (03/2020-01/2022). This includes the period preceding the latest US Presidential election in November 2020. The format in which the data is provided by Facebook is JSON files. Each file is the result of a request for active ad campaigns performed every 12 hours during the 23 months period, thus a lot of ad campaigns are duplicated across files.

TABLE OF CONTENT

INTRODUCTION

(generated)

In recent years, the field of big data analytics has exploded, as companies and organizations of all kinds have recognized the tremendous value of leveraging large datasets to gain insights and make better decisions. However, as the size of datasets has grown exponentially, so too has the need for distributed system solutions to effectively manage and process this data. In this project, we will explore the challenges of big data analytics and the practical applications of distributed systems to address them. Specifically, we will examine real-world examples of how distributed system solutions have been used to overcome the limitations of traditional data processing techniques and enable organizations to extract meaningful insights from vast and complex datasets. Through our analysis, we hope to shed light on the critical role that distributed systems play in the world of big data and highlight the potential benefits of these solutions for businesses and researchers alike.

1. Preprocessiong and Data Collection

For the given data in hdfs, after some data-wrangling, I was able to find a lot of information about adds containg information to Trump and Biden for the year 2020 that is during the election period. I collected the data into two different dataframes for the year 2020. For this, I saved all the paths to json files in a text file. Then extracted path name and year from the text file to collect the information. I was able to make 2 dataframes called as trump\_df and biden\_df.

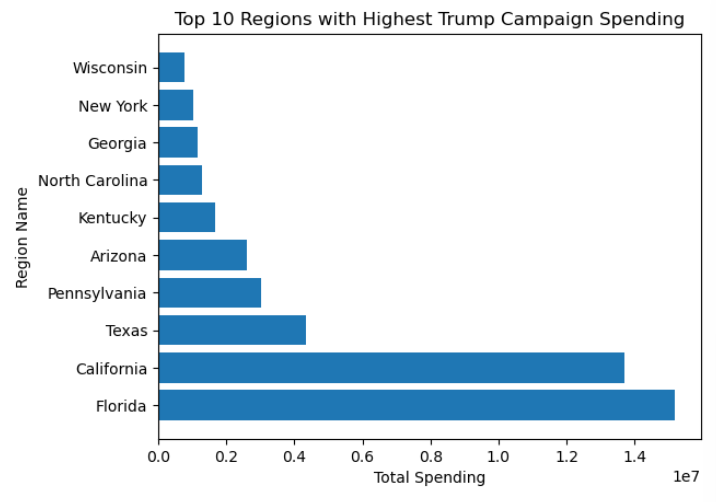
As a part of my data collection, I observed that many of the fields in my data contain information about Biden and Trump. This means they can have negative impact towards Trump so as to promote Biden or vice versa. This type of rows are common to both the tables to do proper sentiment analysis. There are many rows which are duplicates according to ‘id’ but can contain different information related to demographic details, impression etc. This is the reason, I removed duplicates only at the end while doing sentiment analysis.

For the analysis, I made many sub-dataframes and organise the data relevant to the analysis. Along with this, to make visualizations, I converted these tables to padas dataframe for applying visualization libraries on it such as matplotlib, seaborn and plotty. In all, I have carried drill down analysis to understand the key differences between the adds related to Biden and that of Trump.

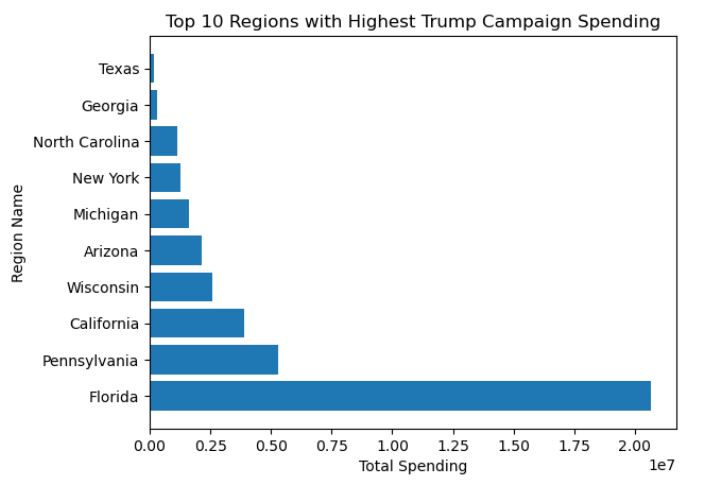
2. Region Analysis

I started my anaysis by understanding the regional distribution of the adds. I created a sub dataframe which contains the region information and it’s spending. The region data was in a form of a tuple containing region name and percentage of the add corresponding to the region such that for all the regions in the add the percentage sums up to one. To get the spending of each reagion I multiply the given percentage to the region spending.

2.1 Trump



2.2 Biden



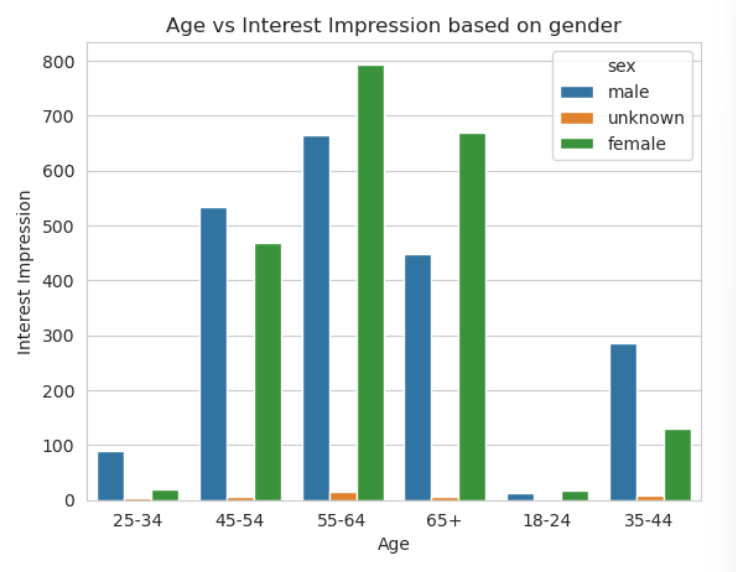
Based on the above information, Trump adds have more spending in Florida and California while for Biden, Florida and Pennsylvania is on the top. It can be noticed that Trump and Biden have Florida as the highest region they have spent on. But the second highest differs significanlty. For trump adds, the second highest is California with very less diference in total spending from florida. But for Biden, Pennsylvania is the second highest.

3. Demographic Analysis Based on Region

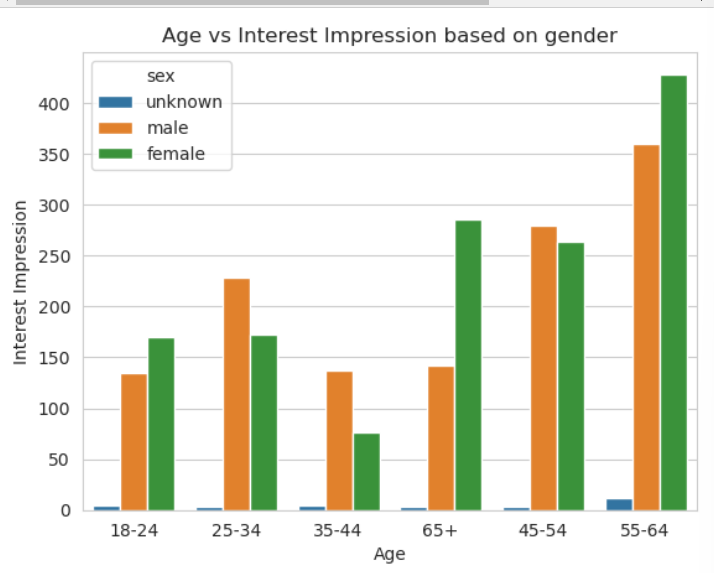
To understand more deeply on adds in perticular region, I exploited the demographic details based on the regions. I created a dataframe called demographic\_df which contains information such as age, gender, percentage interest, impressions of each add.

3.1 Trump ---Graph 1

CALIFORNIA (2nd Highest)

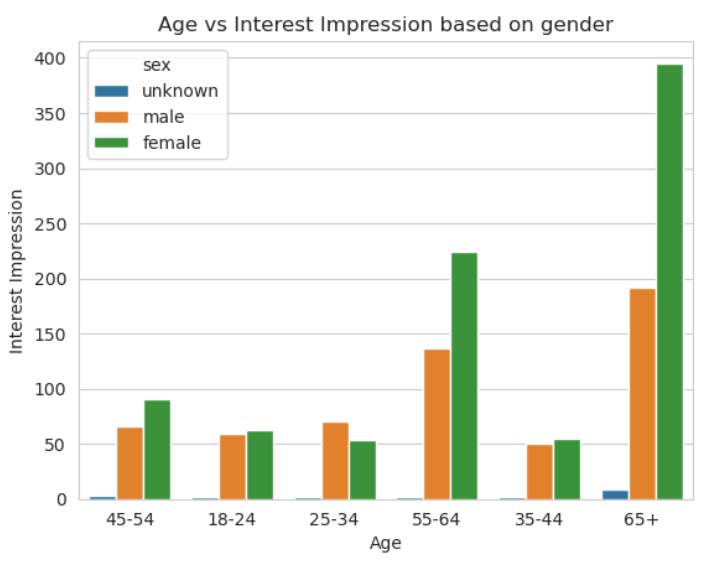
 

FLORIDA (1st Highest)

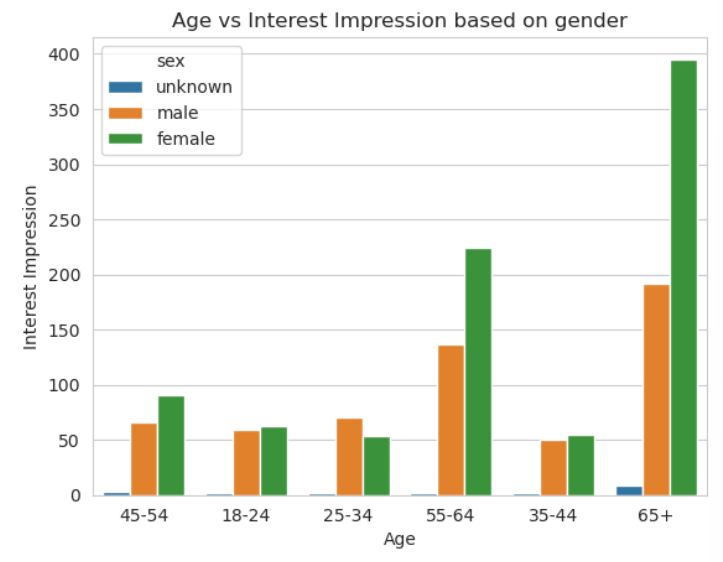


3.2 Biden --- Graph 1

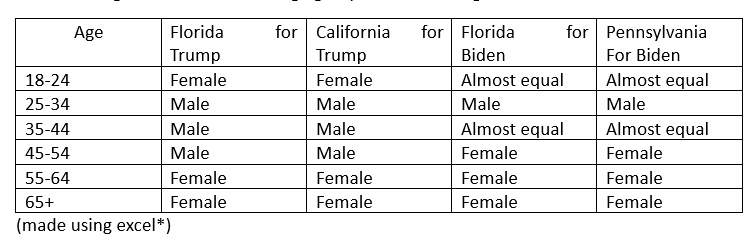
Pennsylvania (2nd Highest)



Florida (1st Highest)



The following table illustrates the age groups in which the gender is dominant:

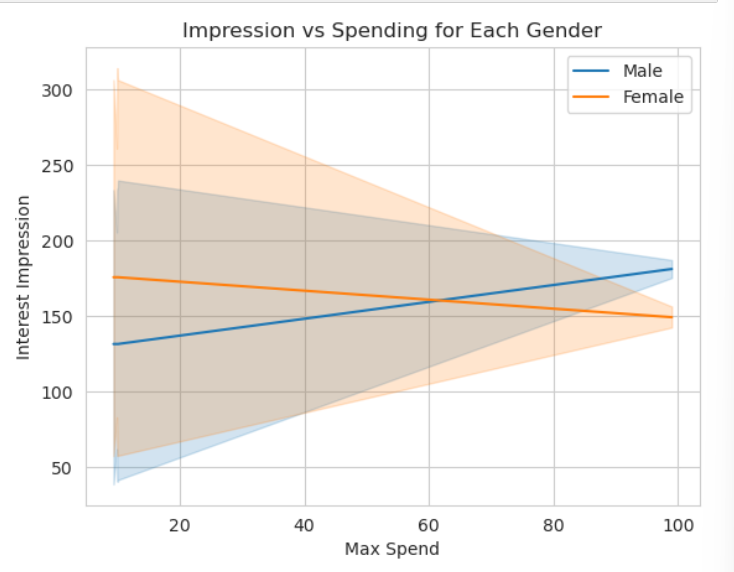


From the table above, it is easier to state that females have a better interest impression in most of the genders. The adds related to trump have a higher impressions for females especially in the younger and older age groups. While for the Biden, there are age groups with equal impressions but females are more dominant.

4. Gender based analysis

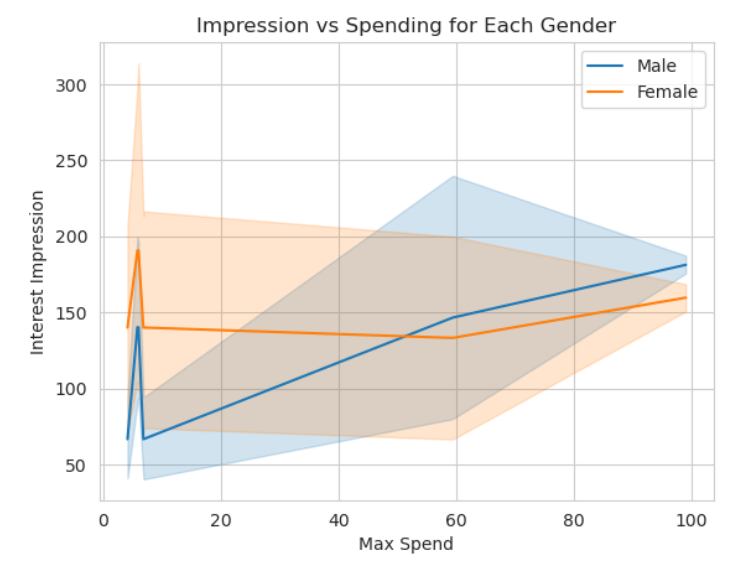
It is clear that most of the age groups in their respective regions have female with higher interest compared to men, this makes me to dive deeper into finding the impression vs spending of Florida according to gender.

4.1 Trump



The observations obtained are quite interesting. It appears that as the maximum spending in the region increases, the interest impression also increases for males, but decreases for females.

4.2 Biden

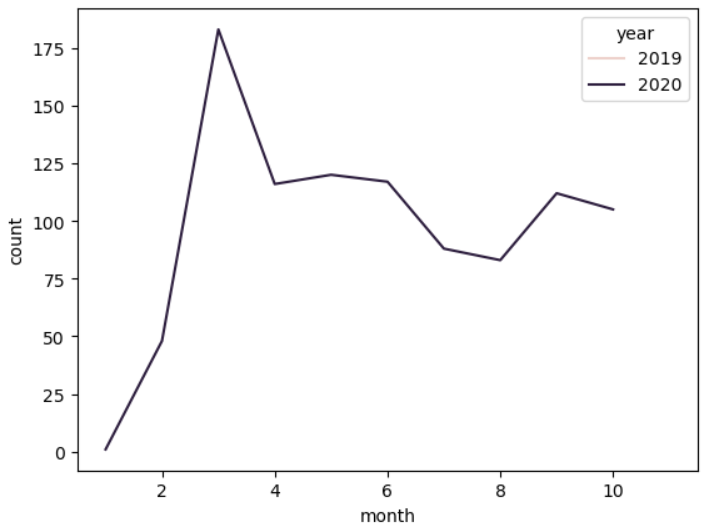


In case of Biden, there are fluctuations in the line graph of female and male. Although the females add impressions are more stable throughout. The lineplots for each gender increase after the threshold 60 of max spend. This suggests that the way male and female respond to an add and the amount of money spent on the add differs for both the groups and the regions.

5. Add content analysis on Females in Florida

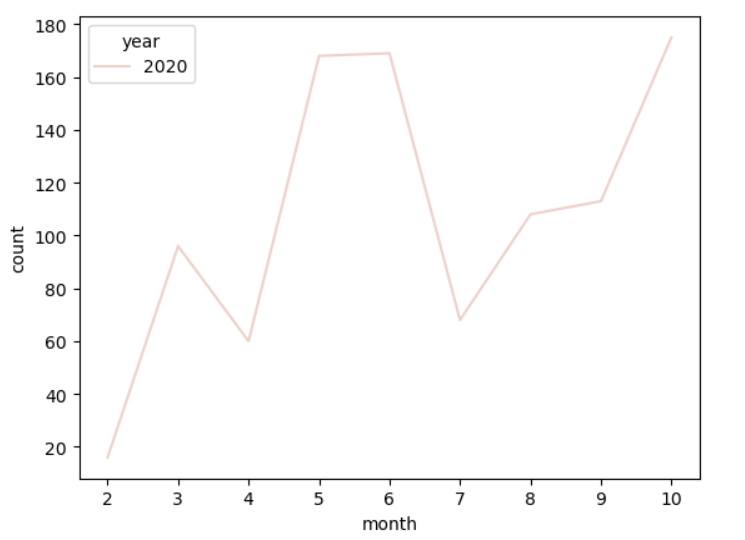
Next, I performed sentiment analysis on the given data for both the candidates. The library ‘SentimentIntensityAnalyzer’ helps to understand the sentiment of an add depending on weather it could contain positive or negative content and labeling it as a fraction between 0 and 1. For better understanding, this decimal number is converted to 0 if it is negative sentiment or 1 if it is positive sentiment. I then filtered data for females in Florida for certain fields like content, creation date, body, etc. which I converted to a csv. Next, I make a graph to understand the add creation time for adds having sentiment for females in florida.

5.1 Understanding the creation time analysis for Trump adds



Most of the adds are created in the year 2020, during the initial months.

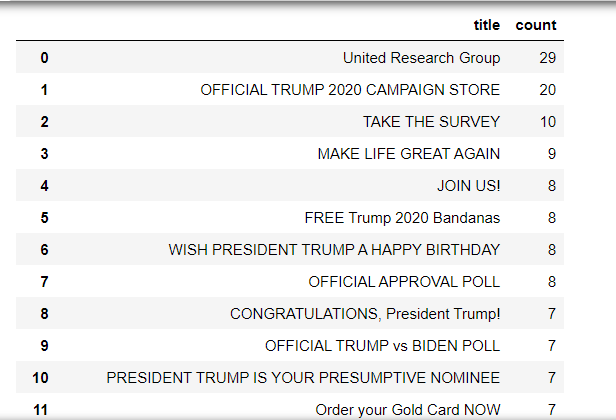
5.2 Understanding the creation time for Biden adds



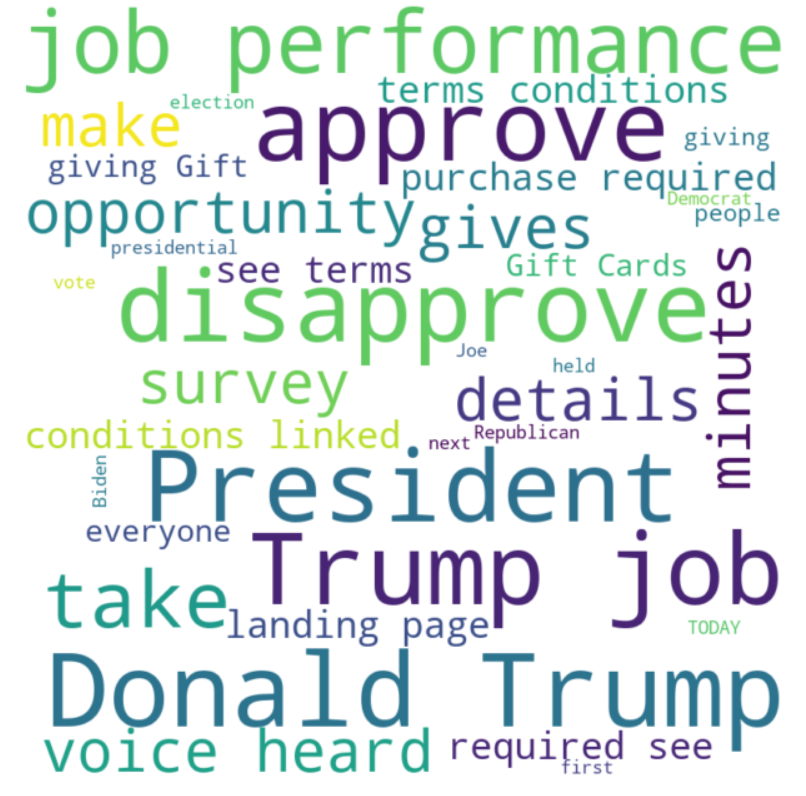
Most of the adds for Biden are created all through out the year, which seem to be increasing towards the later years. This may also mean that that more and more people are gaining interest in Biden related content. The pattern of add creation differ significantly for Trump and Biden.

6. Sentiment Analysis for Trump Adds – Positive Adds for Trump

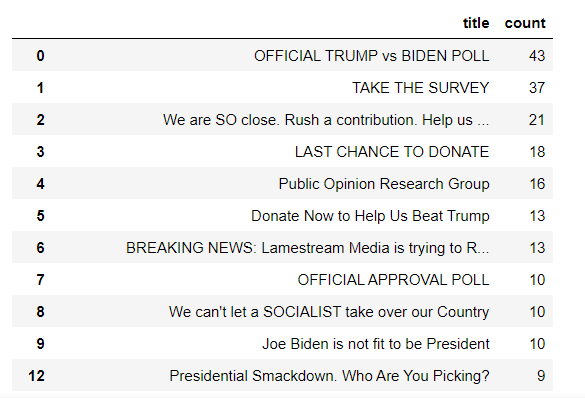
There are around 279312 adds having a positive sentiment towards trump. The following dataframe counts the number of adds by a perticular add-title. From here we can see that ‘United Research Group’ had the most amount of adds.



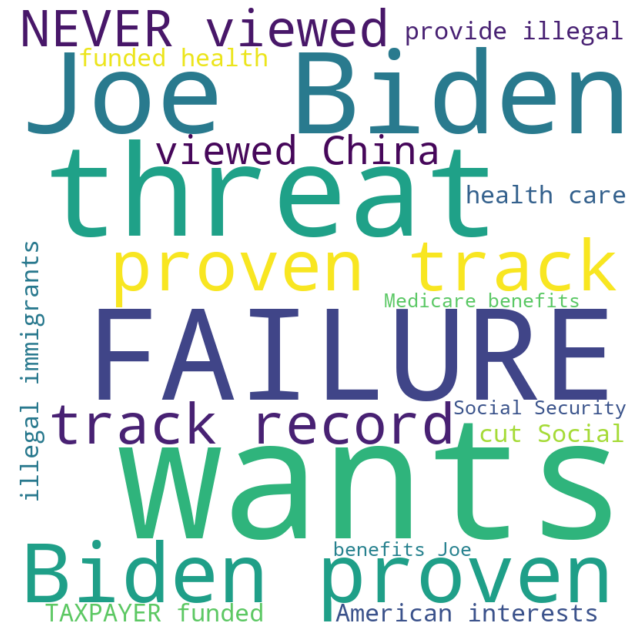
I then filtered out the adds with ‘United Research Group’ and found the frequency of the word distribution of the adds which may effect their performance. From the obtained visualization, words such as job, performance , voice heard which give people a positive impact are expressed more to promote Trump.



Adds which have a negative sentiment for Biden may be promoting Trump. There are ---- adds having negative sentiment towrds biden. The following is a table describing the total count of titles such adds.

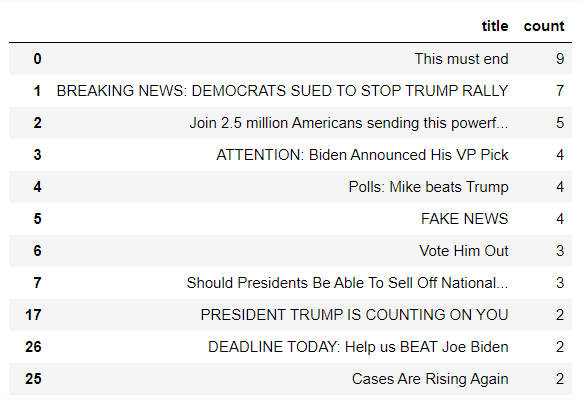


To understand this better, I found the word frequency of words realted to adds containing 26528 as the title.

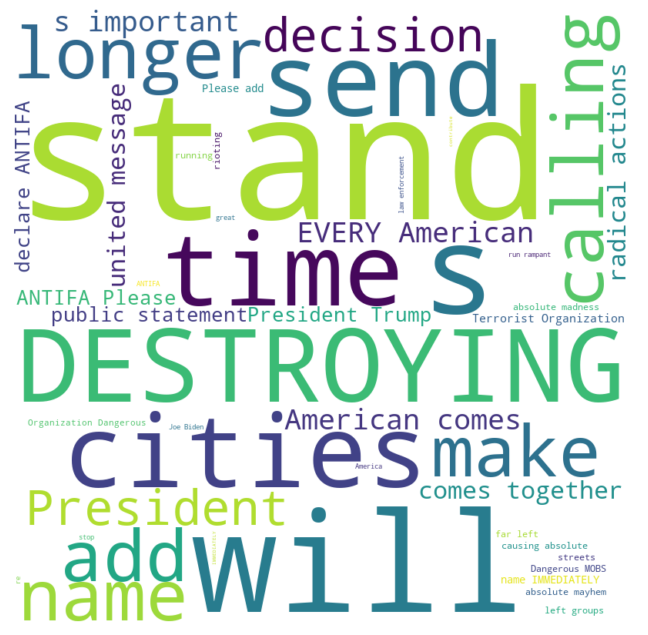


7. Sentiment Analysis for Biden Adds - Positive Adds for Biden

Adds which have a negative sentiment for trump may be promoting Biden, the results for such senarios are mentioned below. There are 78419 total adds bearing a negative impact towards Trump. These adds can be grouped accourding to title as:



From the table above, it can be observed that the title itself are negative and may contain information which can cause a bad perspective for Trump. From these titles I extracted the add content of ‘This must end’ to check the word frequency used by them:



There are words which cause negative impact on the audience like Destroying, terrorist organization, dangerous mobs, ANTIFA, etc.

8. Performance

(generated) --- written for Biden adds in florida

There could be several reasons why most of the ads related to Biden have the most spending in Florida. One possible reason is that Florida is considered a battleground state, meaning that it is not reliably Democratic or Republican, and is therefore crucial to winning the presidency. As a result, political campaigns may focus more of their efforts and resources on such states to try and sway undecided voters or win over voters from the opposing party. Additionally, Florida has a large and diverse population, including a significant Hispanic population, which could make it a strategic target for political campaigns seeking to reach specific demographics. Finally, it is also possible that the Biden campaign simply believed that Florida was a state where they had the best chance of winning and therefore invested more heavily in advertising there.

It's possible that the ads related to Biden are targeting female voters in Florida more heavily than male voters. Political campaigns often use demographic data to target specific groups of voters with their ads. It's also possible that there are simply more female voters in Florida who are interested in Biden, or that female voters in Florida are more likely to engage with political ads in general.

However, without more data and analysis, it's difficult to draw any firm conclusions about why females in Florida may be more interested in Biden ads than males.

There could be multiple reasons why male ads for Biden have more spending in Florida. One possible reason could be that the campaign team has identified that males in Florida are more likely to be swing voters or undecided voters, and hence they are targeting them with more resources in order to sway their vote towards Biden.Another reason could be that the male demographic in Florida is more responsive to certain types of messaging or issues that the Biden campaign is focusing on, and hence the campaign team is investing more in ads that are tailored to appeal to this demographic.

It is also possible that the higher spending on male ads is simply due to a strategic decision by the campaign team based on the available budget and the cost of reaching each demographic in Florida. The team may have found that it is more cost-effective to target males with a higher budget, as it yields a higher return on investment in terms of votes gained.

It's possible that the campaigns have different strategies for targeting each gender based on their analysis of voter demographics and behavior. Additionally, there could be other factors at play, such as differences in ad content or placement, that may be influencing spending patterns. Further analysis and exploration of the data may be necessary to gain a deeper understanding of these patterns.

The results produced by Biden adds prove to be more consistent and as the election comes closer the content for positve impact towards Biden seem to be growing.

CONCLUSION AND DISCUSSION

This project helped me to understand some of the key concepts of big data and it’s usgae especially for real word data that is sourced from Facebook adds. I was able to apply my knowledge of analytics in Pyspark and see many differences from using traditional warehousing methods.

For my analysis, I worked on undertsanding the difference between Adds posted in favour of Trump and that of Biden and do a drill down on various factors. I started this by understaning the regional ditribution of the adds and their spending, which is carried forward to demographic distribution and gender. The patterns shown by both genders and the spending on the adds can significantly impact the way people cast votes. To further analyse, I choose to drill down on the add related content for females in Florida for both the parties.

In futhur discussion, I would like to take a deeper look into how each add conte makes an impact on the audience and how much it effected the election.

In conclusion, there is much difference between add and it’s content for both Trump and Biden. The way they effect people’s opinion depends on various factors such as gender, spendings, impressions, age group which differ from region to region.

REFERENCES