DATA ANALYTICS PROJECT

DANITA ANUBHTI PRAKASH

S4745511

SEMESTER 1 2023

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 present my main findings. The dataset 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 CONTENTS

INTRODUCTION

Big data analytics helps to understand raw, unstructured, complex data to find insights that can help in decision making. In recent years, organizations of all kinds have recognized the tremendous value large datasets to gain insights[1]. The field of big data analytics has exploded. Distributed systems divide the data across multiple compuers that help to reduce the cost and time to analyze Big data[2].

These distributed systems have emerged to benefit challeges posed by big data. Real-world examples like social media platforms need distributed system to overcome the limitations of traditional data processing techniques[3]. By leveraging distributed sustems social media companies can process complex data in real-time[4]. Through this project, I hope to shed light on the real-life applications of big data analytics.

1. Pre-processing

For the given data in hdfs, after some data-wrangling, I was able to find a lot of information about ads containing information about Trump and Biden for the year 2020. I saved all the paths to json files in a text file. Then extracted path name and year to collect the information. I was able to make 2 dataframes called trump_df and biden_df.

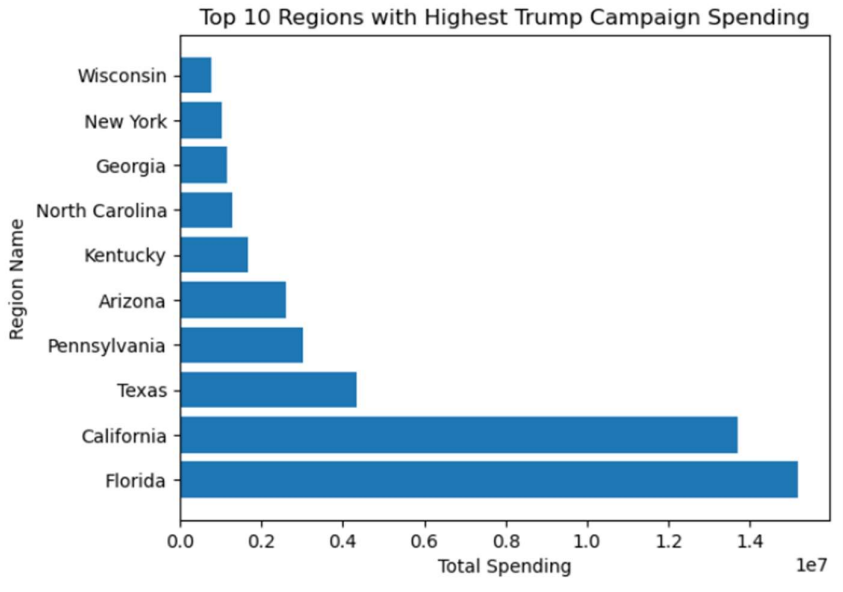
As a part of my data collection, many of the fields in my data contain information about Biden and Trump. This means they can have negative impact towards Trump so as to promote Biden or vice versa. These rows are common to both tables for performing sentiment analysis. There are duplicate rows according to 'id' but can contain different information related to demographic, impression, spending etc. To preserve these values, I removed duplicates before doing sentiment analysis.

For my analysis, I performed a drill down analysis on undertsanding the difference between ads posted for Trump and Biden. Firstly, I started with understanding the spendings of different regions. Which is then, drilled down to demographic distribution, impressions, and add creation time. At the end, I analysed the content and title according to specific gender and region.

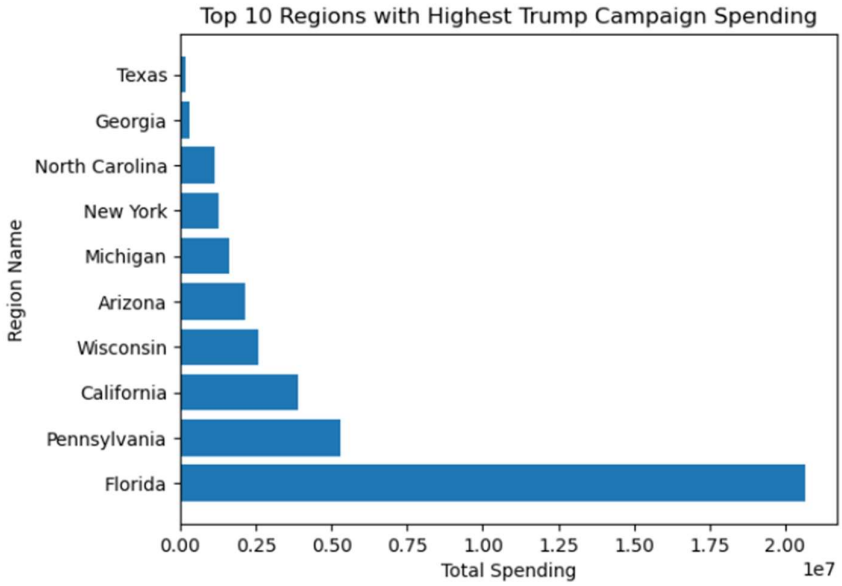
2. Region Analysis

For regional distribution of the ads, 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



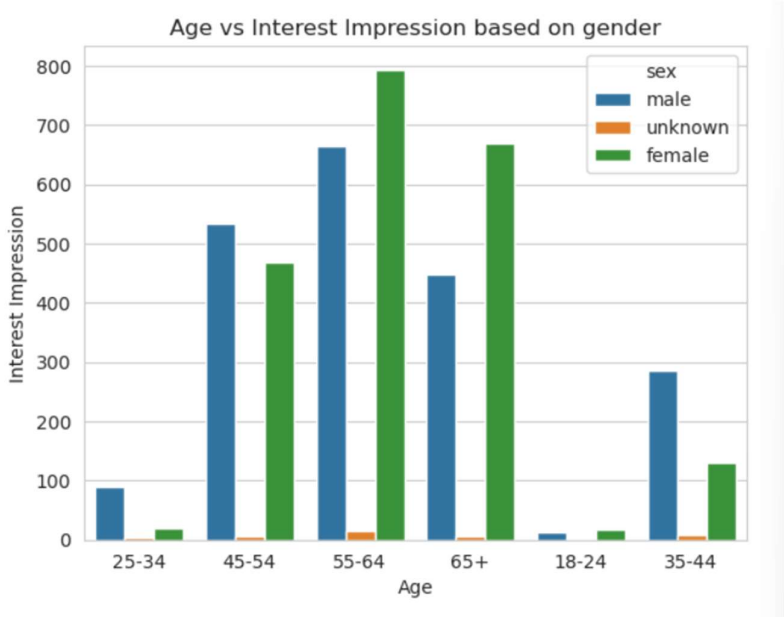
From both the graphs, Trump ads have more spending on the regions- Florida and California while for Biden, Florida and Pennsylvania is on the top. Trump and Biden have Florida as the highest spending region which could be due to larger popullation, diverse demographic groups etc[5]. But the second highest differs significanlty. For trump ads, the second highest is California with very less diference in total spending from Florida.

3. Demographic Analysis Based on Region

To understand more deeply on ads 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.

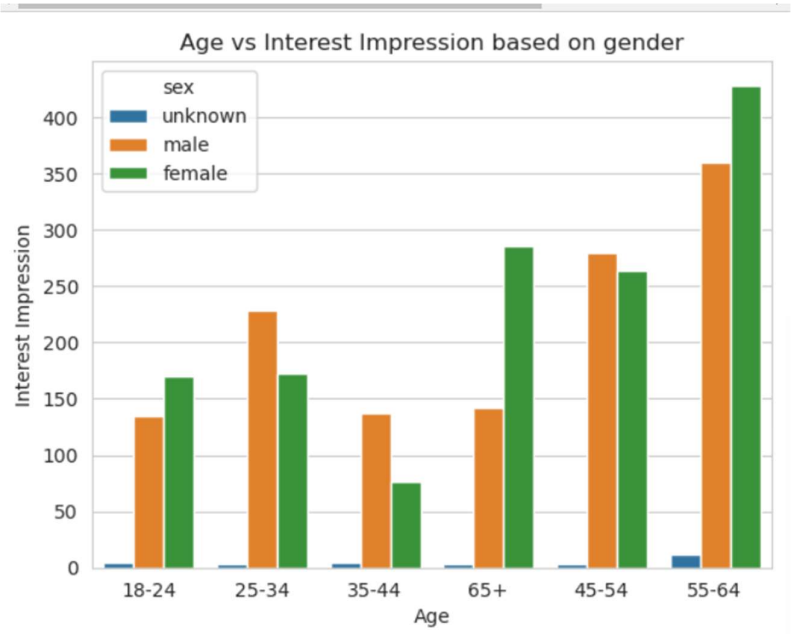
3.1 Trump

CALIFORNIA (2nd Highest)



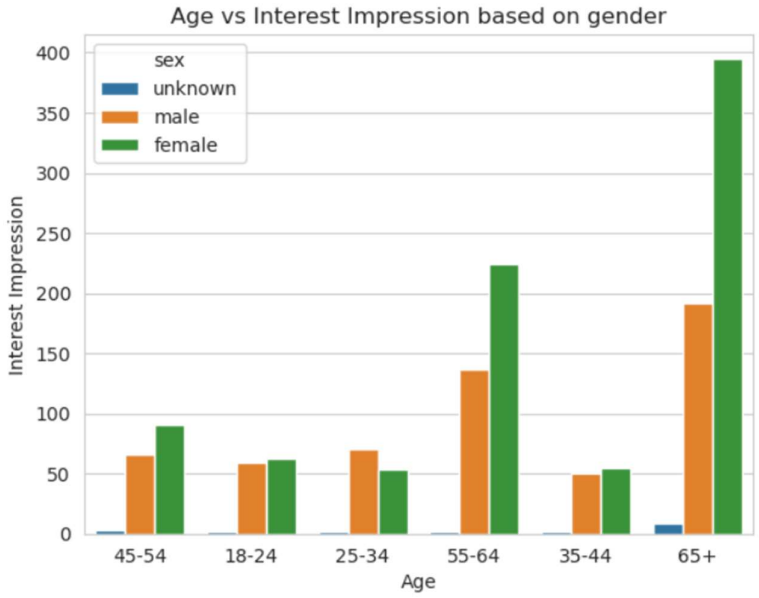
(Note the colour is male – blue, female- green, unknown - orange)

FLORIDA (1st Highest)

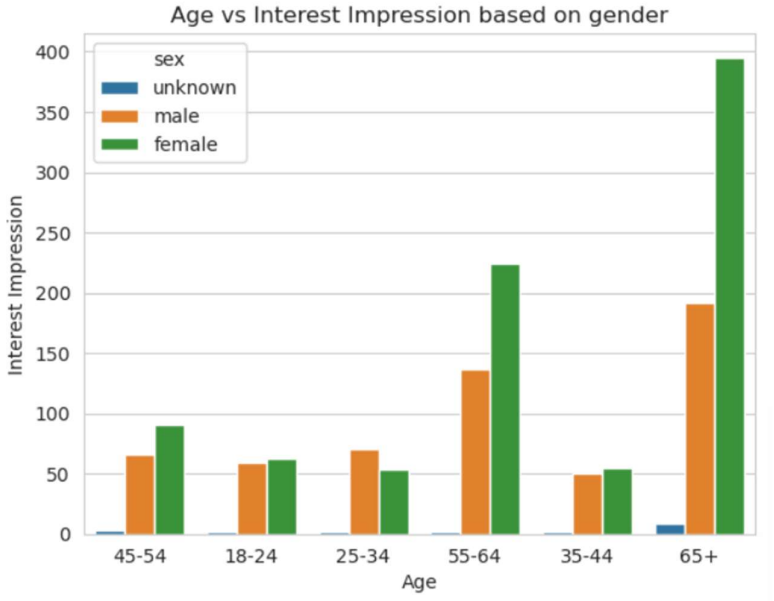


3.2 Biden

Pennsylvania (2nd Highest)



Florida (1st Highest)



(Note- The order of age-group representation in the graphs may differ)

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

Age	Florida for Trump	California for Trump	Florida for Biden	Pennsylvania For Biden
18-24	Female	Female	Almost equal	Almost equal
25-34	Male	Male	Male	Male
35-44	Male	Male	Almost equal	Almost equal
45-54	Male	Male	Female	Female
55-64	Female	Female	Female	Female
65+	Female	Female	Female	Female

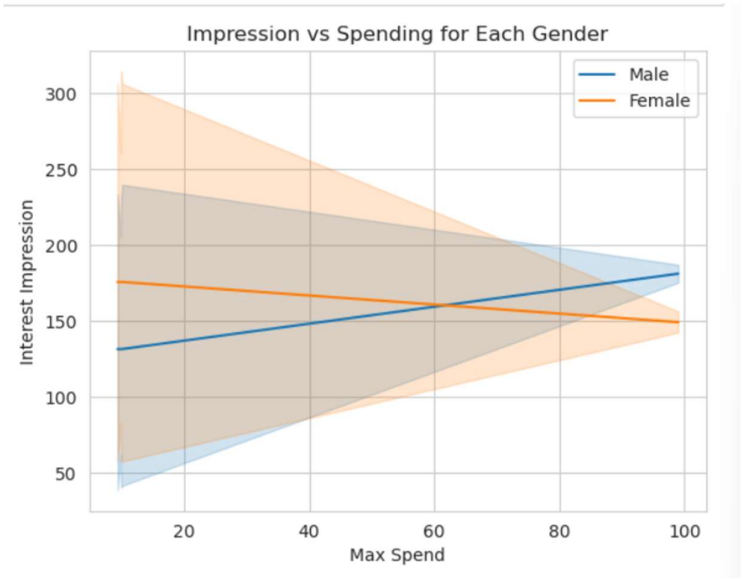
(made using excel*)

The ads 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

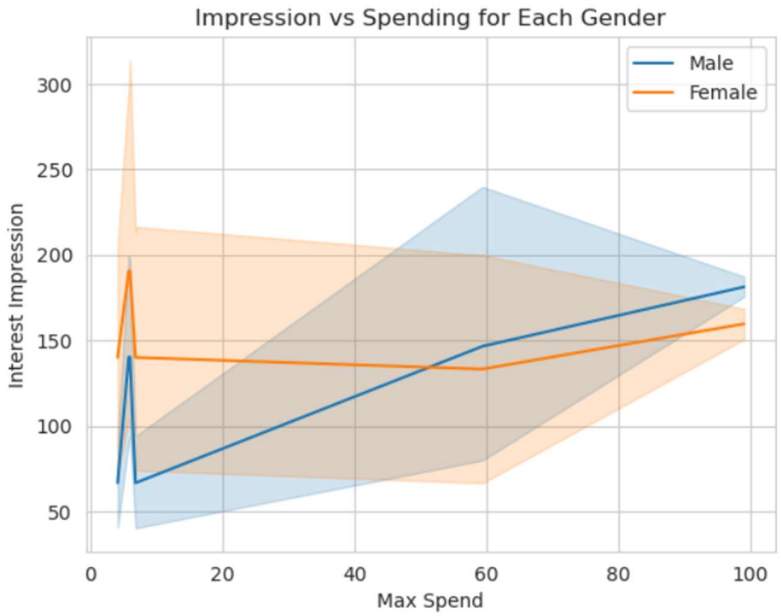
Next, I choose to understand the spending vs impression for each gender as given below:

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

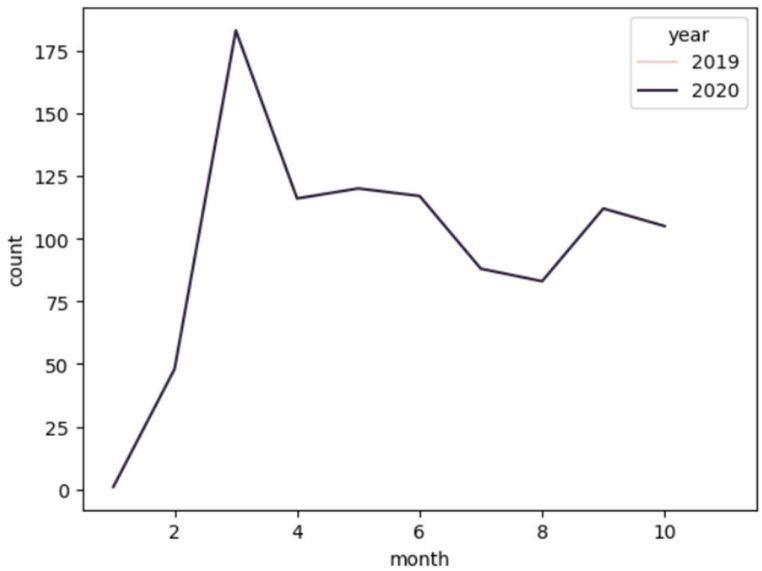


In case of Biden, there are fluctuations in the lineplot. For both male and female, there is a sudden increase in the initial stages, but the plot for males has incremental pattern. The way male and female respond, the amount of money spent on the add differs for gender and the regions.

5. Time Series Analysis

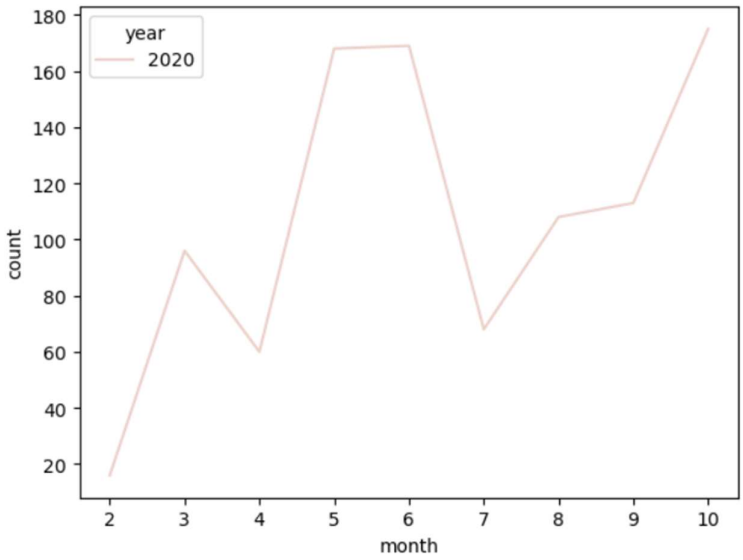
The following is an analysis on the ads in Florida for females using the column add_creation_time. The pattern of ads creation differ significantly for Trump and Biden.

5.1 Trump



Most of the ads are created during the initial months of the year 2020, For the later months the no of ads created seem to be decrementing.

5.2 Biden



Most of the ads for Biden are created all through out the year, which seem to be increasing towards the later years. This may be a good stratergy to get interest of people as ads are released consistently throughout the year.

6. Sentiment Analysis for Trump Ads

Based on the analysis above, the impact of political ads varies based on region, demographics, and creation time. To gain a deeper understanding of these ads, I decided to focus on ads targeted at females in Florida. By examining the type of content and title, I wanted to determine whether they produced a positive or negative sentiment. This sentiment analysis can help provide valuable insights into the effectiveness of political advertising.

This can be done using natural language processing (NLP)[6] techniques, which involve breaking the text into smaller words and phrases, and then assigning score based on the tone of the text. These sentiment scores can then be used to identify patterns and trends in the data, such as which type of ads tend to produce a more positive(1) or negative sentiment(0) among the audience.

I classified Trump-ads in two categories based on sentiment. The following are the number of ads in each section.

sentiment	count
1	279312
0	78419

Next, I divided the ads into two separate dataframes, then checked the add_creative_link_title and total ads under each title. The title tells a lot of details about the type of impact it has

6.1 Positive Ads

The following dataframe counts the number of ads by a perticular add-title. From here we can see that 'United Research Group' had the most amount of ads. Ads with positive sentiment tend to have positive titles such as 'MAKE LIFE GREAT AGAIN', 'OFFICIAL APPROVAL POLL', etc.

	title	count
0	United Research Group	29
1	OFFICIAL TRUMP 2020 CAMPAIGN STORE	20
2	TAKE THE SURVEY	10
3	MAKE LIFE GREAT AGAIN	9
4	JOIN US!	8
5	FREE Trump 2020 Bandanas	8
6	WISH PRESIDENT TRUMP A HAPPY BIRTHDAY	8
7	OFFICIAL APPROVAL POLL	8
8	CONGRATULATIONS, President Trump!	7
9	OFFICIAL TRUMP vs BIDEN POLL	7
10	PRESIDENT TRUMP IS YOUR PRESUMPTIVE NOMINEE	7
11	Order your Gold Card NOW	7

6.2 Negative Ads

The titles itself are negative and may contain information which can cause a bad perspective for Trump such as 'This must end', 'Vote Him out', etc.

	title	count
0	This must end	9
1	BREAKING NEWS: DEMOCRATS SUED TO STOP TRUMP RALLY	7
2	Join 2.5 million Americans sending this powerf...	5
3	ATTENTION: Biden Announced His VP Pick	4
4	Polls: Mike beats Trump	4
5	FAKE NEWS	4
6	Vote Him Out	3
7	Should Presidents Be Able To Sell Off National...	3
17	PRESIDENT TRUMP IS COUNTING ON YOU	2
26	DEADLINE TODAY: Help us BEAT Joe Biden	2
25	Cases Are Rising Again	2

7. Sentiment Analysis for Biden Ads

Similar to the above section, this section does the analysis for Biden ads. The following are the number of positive and negative sentiment on ads for Biden.

```
+---------+-----+
|sentiment|count|
+---------+-----+
|        1|70974|
|        0|26528|
+---------+-----+
```

7.1 Positive Ads

From here we can see that 'OFFICIAL TRUMP vs BIDEN POLL' had the most amount of ads.

	title	count
0	OFFICIAL TRUMP vs BIDEN POLL	43
1	TAKE THE SURVEY	37
2	We are SO close. Rush a contribution. Help us ...	21
3	LAST CHANCE TO DONATE	18
4	Public Opinion Research Group	16
5	Donate Now to Help Us Beat Trump	13
6	BREAKING NEWS: Lamestream Media is trying to R...	13
7	OFFICIAL APPROVAL POLL	10
8	We can't let a SOCIALIST take over our Country	10
9	Joe Biden is not fit to be President	10
12	Presidential Smackdown. Who Are You Picking?	9
13	Your Response Missing: Are you voting for Joe ...	9

7.2 Negative Ads

Most of the titles are very negative, which depromote Biden. Some of the negative titles which are eye catching are 'JOE BIDEN IS BAD FOR AMERICA', 'JOE BIDEN IS DANGEROUS FOR AMERICA', etc.

	title	count
0	JOE BIDEN IS BAD FOR AMERICA	33
1	JOE BIDEN IS DANGEROUS FOR AMERICA	30
2	Joe Biden's policies are RADICAL and RECKLESS	29
3	Vote For Biden	17
4	FAKE NEWS	16
5	BREAKING NEWS: Biden Announced His VP Pick	13
6	Joe Biden is SLIPPING	9
7	RELEASE the records now!	9
8	ATTENTION: Biden Announced His VP Pick	9
9	THIS JUST IN: Trump to designate ANTIFA & the ...	7
10	Joe Biden's Liberal Ideas would CRUSH our economy	6
13	Joe Biden has embraced the RADICAL Left	5

In all, interesting observation can be noted on how ads tend to gain attenion from audience and influence their interest towards a perticular candidate.

CONCLUSION AND DISCUSSION

This project helped me to understand some of the key concepts of big data and it's real world usage. I was able to apply many concepts of Pyspark and Hadoop and theirunderstand differences from traditional warehousing methods.

Overall, through the drill down analysis, the ads related to Biden performed better in terms of consistency and for attracting a bigger audience from different age, gender and region. Targetting Florida was a common stratergy for both the parties. The spending and impressions by both genders showed interesting patterns. Lastly, the title and content of ads produce a strong positive or negative impact on the interest and opinion of people.

In conclusion, there is much difference between ads and it's stratergies for both Trump and Biden. The analysis showed that a well-planned approach coupled with target spending captures interest of people that causes a significant impact in the performance of political campaigns.

REFERENCES

[1] Author links open overlay panelNataliya Shakhovska a, a, b, and AbstractThe analysis of Big data technologies was provided. An example of MapReduce paradigm application. “Big Data Processing Technologies in Distributed Information Systems.” Procedia Computer Science, November 21, 2019. https://www.sciencedirect.com/science/article/pii/S1877050919317478.

[2] (PDF) distributed computing in big data analytics: Concepts ... Accessed May 15, 2023. https://www.researchgate.net/publication/317427131_Distributed_Computing_in_Big_Data_Analytics_Concepts_Technologies_and_Applications.

[3] Mining Social Media: A brief introduction - pubsonline. Accessed May 15, 2023. https://pubsonline.informs.org/doi/abs/10.1287/educ.1120.0105.

[4] Mining Social Media: A brief introduction - pubsonline. Accessed May 15, 2023. https://pubsonline.informs.org/doi/abs/10.1287/educ.1120.0105.

[5] “Demographics of Florida.” Wikipedia, May 3, 2023 https://en.wikipedia.org/wiki/Demographics_of_Florida

[6] Genç, Özgür. “The Basics of NLP and Real Time Sentiment Analysis with Open Source Tools.” Medium, April 21, 2019. https://towardsdatascience.com/real-time-sentiment-analysis-on-social-media-with-open-source-tools-f864ca239afe.

(TOTAL WORD COUNT 1663)