DATA7201 Project Report

Ruyun Qi (4450606)

May 26, 2022

Abstract

In 2021, Facebook made political ad data available to academic researchers. Analyzing the advertisements that ran in the month prior to the US presidential election allows people to discover how the advertising influenced the outcome of the election. This project gives a quick analysis of the high-frequency words in ad text, as well as the campaigns keen to push high-paying advertising. The findings show the prominence of keywords in political ad content and campaigners' power to advertise on Facebook.

Contents

1 Introduction

1.1 General area of big data analytics

Big data analytics is the application of related algorithms to mass data processing, analysis, and storage, intending to extract value from a large amount of data in our daily life and industry production. The availability of more data on everything, as well as advancements in storage and processing technology, are driving the big data trend. Big data analytics has an influence on all aspects of life, including food, telecommunications, economics, entertainment, sports, and other fields.

The volume of data is the first characteristic of big data. The development of hardware technology cannot keep up with the rate of growth of data capacity, resulting in a data storage and processing problem (Marr, 2016). However, the big data storage dilemma is impacted not only by the exponential growth of data but also by the variety of data types, which is the second characteristic. With the emergence of multimedia applications on the internet, the proportion of unstructured data such as sound, pictures, and video is increasing. The third characteristic is high speed. Files that used to take a long time to transfer may now be completed in a few seconds, but this brings a significant challenge to the accuracy of the data and the server's seamless operation.

1.2 Motivation of distributed system solutions

A distributed system is composed of a collection of computer nodes that interact through a network and collaborate to complete shared tasks. When a single node's processing power is insufficient to fulfill the needs of ever-increasing computing and storage tasks, hardware upgrades are too expensive, and the program cannot be further optimized, we can consider distributed systems. Computing and storage are complementary in that computing requires data, either real-time or stored data and computing outputs must also be saved. As a result, there are a variety of distributed system solutions that boost performance in these aspects.

Distributed File System (DFS) provides a logical tree file system structure for resources that are distributed throughout the network, making it easier for users to access shared files. Hadoop Distributed File System (HDFS) is well-suited for distributed computing and storage, and it can be run on low-cost machines because of its high fault tolerance and scalability. HBase is designed for the storing of unstructured data, which is different from normal relational databases. Hadoop HDFS provides HBase with high-reliability low-level storage, while Hadoop MapReduce provides HBase with high-performance computational power. The distributed computing system is built on distributed storage, which divides long-running activities into multiple jobs and processes them concurrently, increasing computation efficiency. For different scenarios, the distributed computing system is grouped into three categories: off-line (e.g. Hadoop), real-time (e.g. Spark), and streaming (e.g. Storm, Flink/Blink). These advanced distributed system solutions contribute significantly to big data analytics.

2 Dataset Analytics

2.1 Description of dataset

Political advertising on the internet is getting more popular since political campaigns realize the potential of social media platforms like Facebook in reaching and engaging voters. The datasets used in this project derive from the Facebook Ad Library API and include the Facebook political ads data from March 2020 to January 2022. This project focuses on four columns: ad creative bodies, ad ID, funding entity, and spend, because of the massive datasets. The data types and data descriptions are also listed in Table 1. Since the 2020 United States presidential election was held on 3rd November 2020, I would be more interested in the data on political ads throughout the month before the election. Therefore, all October 2020 files are loaded for further data analytics. The ad text and campaigns that are willing to pay a premium for their ads will be analyzed. The stakeholders may include political campaigns and researchers, looking to learn about the trends of political advertising on social media platforms.

Column Name	Data Type	Description
id	string	ID for the archived ad object.
ad_creative_body	string	Text displayed in the ad.
funding_entity	string	The name of the person, company, or entity that provided funding for the ad.
spend.lower_bound	integer	Lower bound of the money spent running the ad.

Table 1: Interested Columns of Facebook Political Ads Data, Adapted from Facebook Political Ads Data Warehouse Tables *(n.d.)*

2.2 Pre-processing steps

The first step of pre-preprocessing will be to load the data files in October 2020. Because the JSON files provided by Facebook are updated every 12 hours, it is required to load several files using wildcards, such as FBads-US-202010, which matches all October 2020 files. However, the entire dataset is still too huge, making the loading and processing procedures time-consuming and memory-intensive. Thus, the second step is to select the columns that will be used for analysis, as well as to filter out null values and duplicates. Since the data is structured, PySpark is used for data pre-processing and analysis in this project. Spark is more efficient than Pig when dealing with large amounts of data, because of its in-memory computing and distributed processing using parallelize. Furthermore, Pyspark includes some essential packages, such as PySpark DataFrame and PySpark SQL, that make analysis easier to code and data more readable. PySpark SQL functions such as pyspark.sql.Column.alias, pyspark.sql.functions.col, and pyspark.sql.functions.filter are used in the pre-processing steps.

2.3 Ad text analysis

The goal of ad text analysis is to extract high-frequency words from the text descriptions of hundreds and millions of ads. This analysis includes the following steps:

1. Lowercase the words in the ad text and split them up into words.
2. Count the number of times each word appears in all the selected ads.
3. Sort the data in descending order by word frequency.
4. Remove the "stop words," which are the most commonly used, such as"is", "at", "which", and "on".
5. Export a CSV file with the top 100 words and their frequencies.
6. Remove some meaningless words manually.
7. Make a word cloud for visualization.

As shown in the word cloud (i.e. Figure 1), some action verbs, such as vote, support, and help, are frequently used in political advertising to encourage people to vote in the presidential election. Moreover, significant keywords such as country, state, city, and people intend to inform voters of critical issues and decisions that must be made. "Trump" appears in the top 50 words in political commercials, which is an interesting fact. This reveals that Trump's 2020 election campaign spent a significant amount of money. Other keywords include health, education, work, and tax, which are the most pressing considerations in society.

Figure 1: Word cloud of high-frequency words in ad text

This analysis did not use complex information retrieval techniques but extracted and counted the words. The requirement of manual keyword filtering is one of this method's drawbacks. Also, classifying high-frequency words may provide a more in-depth understanding of ad keywords.

2.4 High-paying ad analysis

High-paying ad analysis aims to determine which campaigns are willing to pay a premium for their political advertisements. This analysis includes the following steps:

1. Extract the ads with a lower bound of more than $5,000 in spend.
2. Count how many high-paying ads each funding entity has.
3. Sort the data in decreasing order by the number of high-paying ads.
4. Export a CSV file including the funding entities and the number of high-paying advertising they have.
5. Use Tableau for visualization.

As shown in Figure 2, the top 15 funding entities took 26.54 percent of high-paying ads in October 2020, indicating that huge financial groups are willing to spend a lot of money on advertising. Since Biden and Trump are campaigners in the 2020 presidential election, it's not surprising to see their campaigns play a significant part in Facebook ads.

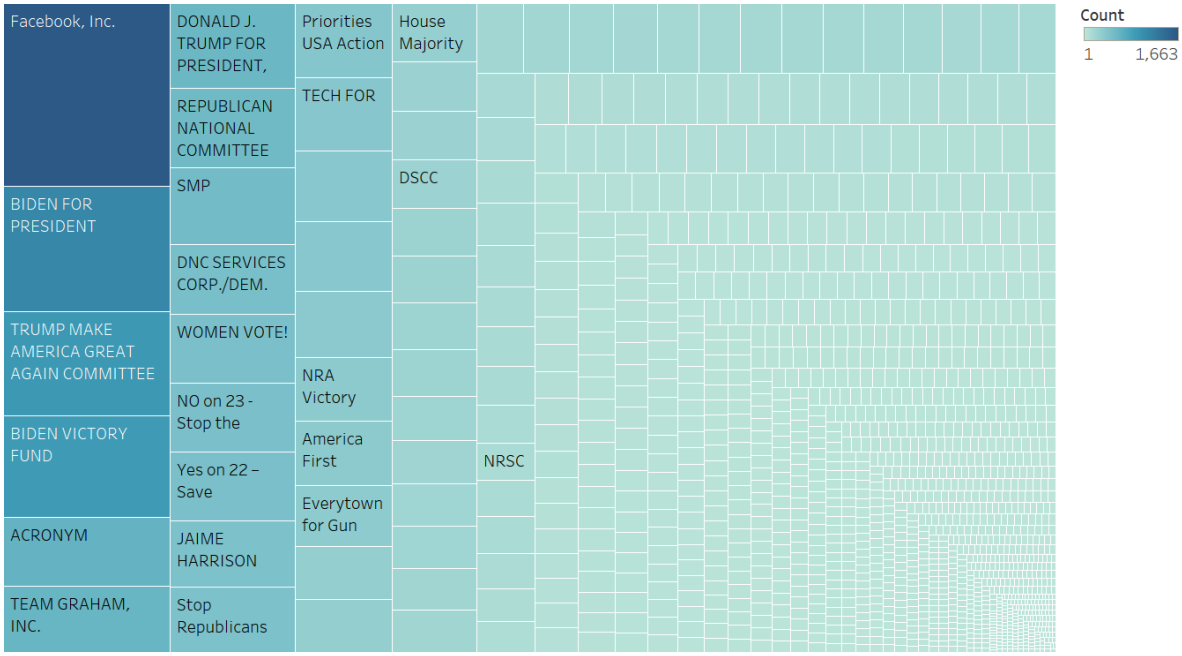

Figure 2: Funding entities with count of high-paying ads

When we look at the top 10 funding entities shown in Figure 3, we can see that, except for Facebook, Inc., Biden's campaigns have more high-paying advertising than Trump's. It's intriguing that while "Trump" appeared more in ads text, Biden's campaigns spent more money on ads. This might reveal their preferences for political ad marketing.

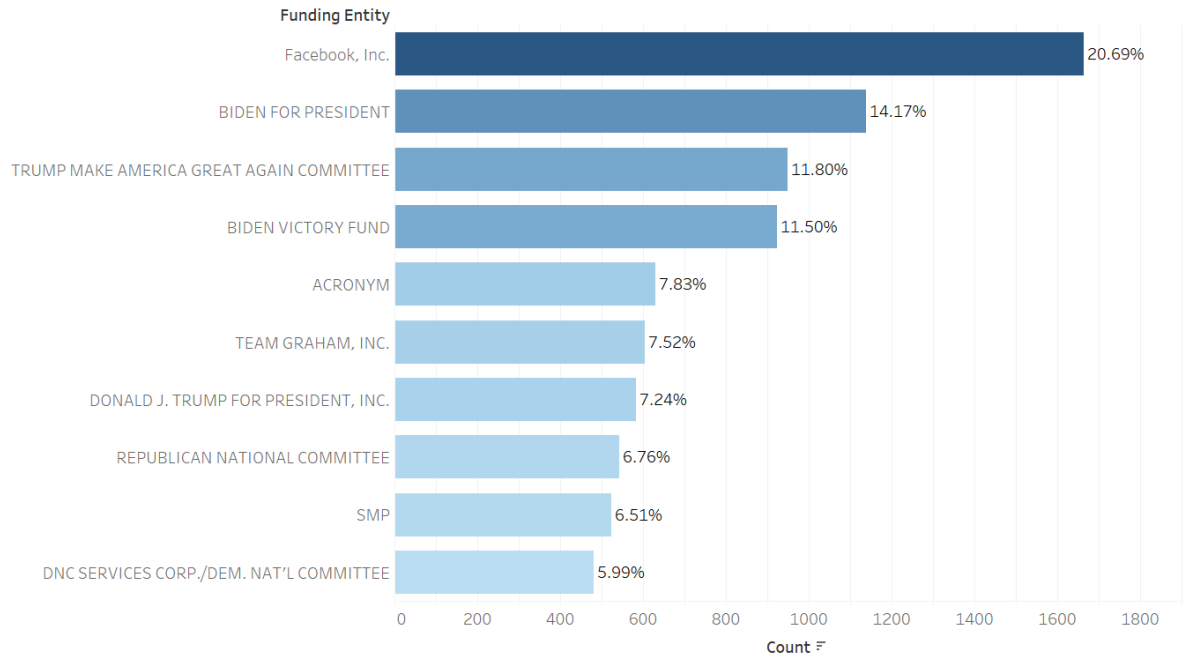

Figure 3: Top 10 funding entities for high-paying ads

This analysis only examines ads having a lower bound of more than $5000 in spend. It may be preferable to design a scoring system by multiplying the weights of various expenditures by the ad counts. The weighted scores should reflect the purchasing capabilities of political ads for the top campaigns.

3 Conclusion

This project provides a brief review of the high-frequency words in the ad text, as well as the campaigns eager to promote high-paying ads in the month before the 2020 US presidential election. Encouraging action verbs and social topic keywords are commonly used in political ads. Campaigns could be more inspired by learning from the ad titles by understanding which and why some keywords are more influential. Aside from the title's impression, the ad budget can make a major difference. The campaigners' financial support for these high-priced ads affected much in the presidential election. This reveals the impact of political advertising on social media, as well as the relationship between campaign results and marketing spending. More facts and insights are expected to emerge from other studies on the regional or demographic distributions of political adverts on social media platforms. These findings are significant in terms of showing political finance transparency and further exposing the mystery surrounding online political advertising (Gitomer et al., 2021).

Word count: 1415

References

Facebook Political Ads Data Warehouse Tables. (n.d.). Retrieved 2022-05-25, from https://supermetrics.com/docs/integration-facebook-political-dwhtables

Gitomer, A., Oleinikov, P. V., Baum, L. M., Fowler, E. F., & Shai, S. (2021, February). Geographic impressions in Facebook political ads. *Applied Network Science*, *6*(1), 18. Retrieved 2022-05-25, from https://doi.org/10.1007/s41109-020-00350-7 doi: 10.1007/s41109-020-00350-7

Marr, B. (2016). *Big data in practice: how 45 successful companies used big data analytics to deliver extraordinary.* West Sussex, United Kingdon: John Wiley and Sons Ltd. Retrieved from https://onlinelibrary-wiley-com.ezproxy.library.uq.edu.au/doi/book/10.1002/9781119278825

A PySpark Code

```
"""
DATA7201 Project
Ruyun Qi (44506065)

"""

import pyspark
from pyspark.sql import SparkSession, SQLContext, DataFrame
from pyspark.sql.types import *
from pyspark.sql.functions import mean, min, max, lit, col, first, last, lower, explode, sz
import pyspark.sql.functions as f
import pandas as pd

sc = pyspark.SparkContext("local")
sqlContext = SQLContext(sc)

# Load data files in Oct 2020
data = spark.read.json("/data/ProjectDatasetFacebook/FBads-US-202010*")
data.printSchema()

# Select columns & Filter out NULL values
df = data.select("id",
                 "ad_creative_body",
                 "funding_entity",
                 col("spend.lower_bound").alias("lower_bound")),
         .filter('ad_creative_body IS NOT NULL')\
         .filter('funding_entity IS NOT NULL')\
         .dropDuplicates()


### Ad text analysis
# Lowercase ad text
df1 = df.select(lower(col("ad_creative_body")).alias("ad_creative_body"))\
        .dropDuplicates()

# Word frequency
word_freq = df1.withColumn('word', explode(split(col("ad_creative_body"), ' ')))\
               .groupBy('word')\
               .count() \
               .sort('count', ascending=False)\
               .toPandas()
```

```
# Install stop-words package
import sys
print(sys.executable)
!{sys.executable} -m pip install stop-words

# Filter out stop words
from stop_words import get_stop_words
stop_words = get_stop_words('en')

# Top 100 words
word_freq = word_freq[~word_freq['word'].isin(stop_words)].head(100)

# Export CSV file for making a word cloud
word_freq.to_csv('Word Frequency.csv', encoding='utf-8', index=False)


### Ad (£5k+) analysis
# Group by funding entity
df2 = df.filter("lower_bound > 5000")\
        .groupBy('funding_entity')\
        .count() \
        .sort('count', ascending=False)\
        .toPandas()

# Export CSV file for data visualization in Tableau
df2.to_csv('lower_bound_5k.csv', encoding='utf-8', index=False)
```