# Australia Federal Election 2022
# Data Analytics Project Report

Lok Yee Joey Cheung

University of Queensland
20 May, 2024

# Abstract

In the year 2022, Australia held its Federal Election in May, resulting in the election of a new prime minister. During the campaign period, there was heightened political activity as parties competed to secure voters' support using various methods, including extensive advertising efforts. By analyzing the political posts and advertisements running on Facebook during the election period, this project aims to investigate text content analysis and the impacts of high ad spending strategy on the election outcome. The results reveal the importance of political campaigns investing in effective advertising strategies and establishing connections with voters.

# Table of Contents

# 1. Introduction

Big data analytics involves diving into vast amounts of data to uncover patterns, trends and insights that can inform decision-making and drive innovation (Sagiroglu et al., 2013). The continuous generation of data from various sources including social media and IoT devices presents challenges for traditional single-server systems. By 2025, global data creation is predicted to reach 180 zettabytes (Taylor, 2023). Traditional systems struggle to keep up with the rapid growth in data volume and its diverse varieties, namely structured, semi-structured and unstructured data. Moreover, traditional batch processing systems encounter challenges in efficiently processing streaming data at high speeds, let alone conducting instantaneous analysis.

Distributed systems address these challenges by offering multiple nodes or servers to support data distribution and parallel computation. They offer horizontal scaling, fault tolerance and parallel processing, which facilitates efficient big data analytics tasks. For instance, Hadoop Distributed File System (HDFS) provides distributed storage across a cluster of commodity hardware and enables effective data processing using MapReduce framework. HBase, a distributed NoSQL database offers low latency and retrieval. These distributed systems facilitate high scalability to accommodate growing data volumes, and their fault tolerance ensures data reliability while minimizing the risk of data loss. Additionally, parallel processing capabilities also enable faster big data analysis, improving efficiency in handling large datasets.

# 2. Data Analytics

## 2.1 Goal

I utilized HDFS for efficient data storage and PySpark for big data processing regarding the Australian Federal Election 2022. The analysis mainly focused on investigating the impacts of political campaigns' spending on the engagement with voters and the election outcome. The stakeholders include political campaigns, media, researchers, government bodies and the general public, who wish to seek insights into political advertising and its impact on election outcomes. Apache Spark was chosen for its efficiency in-memory processing, flexibility, and scalability. PySpark supports SQL functions and Machine Learning features, enhancing pre-processing and analysis of structured data. Moreover, Spark could effectively integrate with HDFS for efficient data handling.

## 2.2 Dataset and Techniques

The dataset comprises political posts obtained from the Facebook Ad Library API. These posts were collected between March 2020 and February 2024 and targeted at Australians. Since our analysis focuses on the Australian Federal Election in May 2022, I extracted data from February to May 2022 which covers both the pre-election and post-election periods. Our analysis centered on columns related to advertisements' content, funding entities, spending, start time, and more relevant metrics.

## 2.3 Pre-processing

First, I loaded the entire *JSON* dataset. I narrowed down the entire dataset between February and May 2022 by filtering the advertisements' start dates. Then, I filtered columns "id", "ad_creative_body", "funding_entity", "ad_delivery_start_time", and "spend" for further analysis, removing any null values. I created a new column named "spend_midpoint" to accommodate the midpoints of range values of the 'spend' column. Midpoints provide a representative value that lies equidistant between the lower and upper bounds of the range, computed by averaging the two bounds. Finally, I removed duplicated campaigns to prevent double counting or skewing of analysis results. This dataset, 'df_clean' resulted in around 180,000 records.

## 2.4 Volume of Ads across the Federal Election Period

To commence, I analyzed the overall trend of political advertisements. I investigated the change in advertisement volume throughout the election period by grouping the pre-processed data by month. After counting the number of advertisements in each month, I ordered the months in ascending order and plotted it in Figure 1. I transformed the processed data to a pandas DataFrame and exported it to a CSV file for plotting using *matplotlib* in *Python*. Figure 1 illustrated a notable increase in advertisement volume when the election approached. This surge was the highest from April to May, suggesting heightened campaign activity and efforts to engage voters as the election date drew near. The peak was reached in May, with approximately 70,000 advertisements.

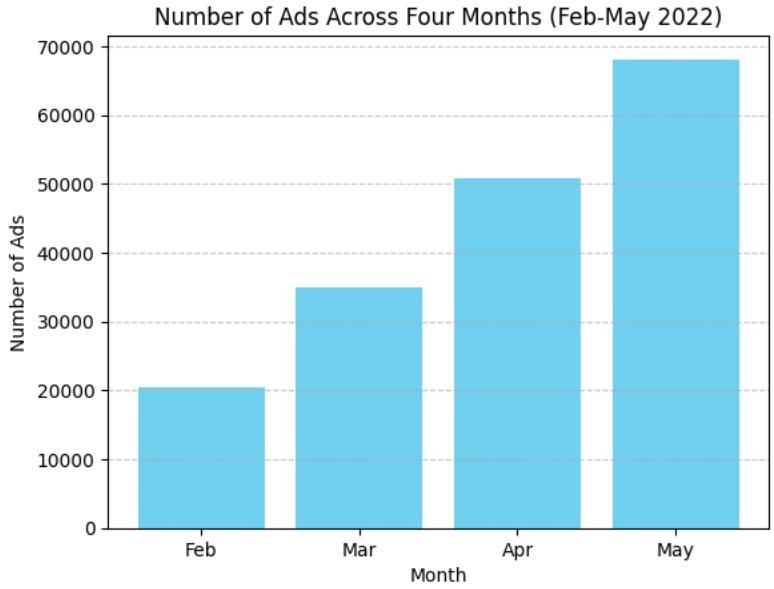


*Figure 1. Bar chart of Ads volume*

## 2.5 Text Analysis in Ads

This section extracted high-frequency words from the advertisements' content and descriptions. Initially, I converted the words in the 'ad_creative_body' column of the pre-processed dataset, 'df_clean', to lowercase, split them into individual words and filtered out empty spaces and punctuations. Then, I counted the frequency of each word and ordered them in descending order. Similar to the previous section, I transformed the processed data to a pandas DataFrame and exported it to a CSV file for plotting using *matplotlib* in *Python*.

By visualizing the frequency in Figure 2, I observed a Zipf's law pattern where fewer words occur very frequently while the majority occur less frequently (Piantadosi, 2014). Stop words appear frequently and dominate the distribution which may skew the result. I remove stop words with *StopWordsRemover* imported from PySpark Machine Learning Library. I then sorted the word frequencies in descending order (see Figure 3) and visualized the top 100 words using *WordCloud* library.

In Figure 4, significant keywords such as 'government', 'people', 'federal', 'local', and 'Australian' were common in political advertisements which highlighted the importance of social awareness and participation during the federal elections in Australia. Frequent action words namely 'support', 'vote', 'help', 'need' etc. stood out which encouraged engagement in the electoral process. The word 'labor' appearing within the top-10 word frequency suggests that the Labor Party may have placed significant emphasis on campaign promotions through advertisement.

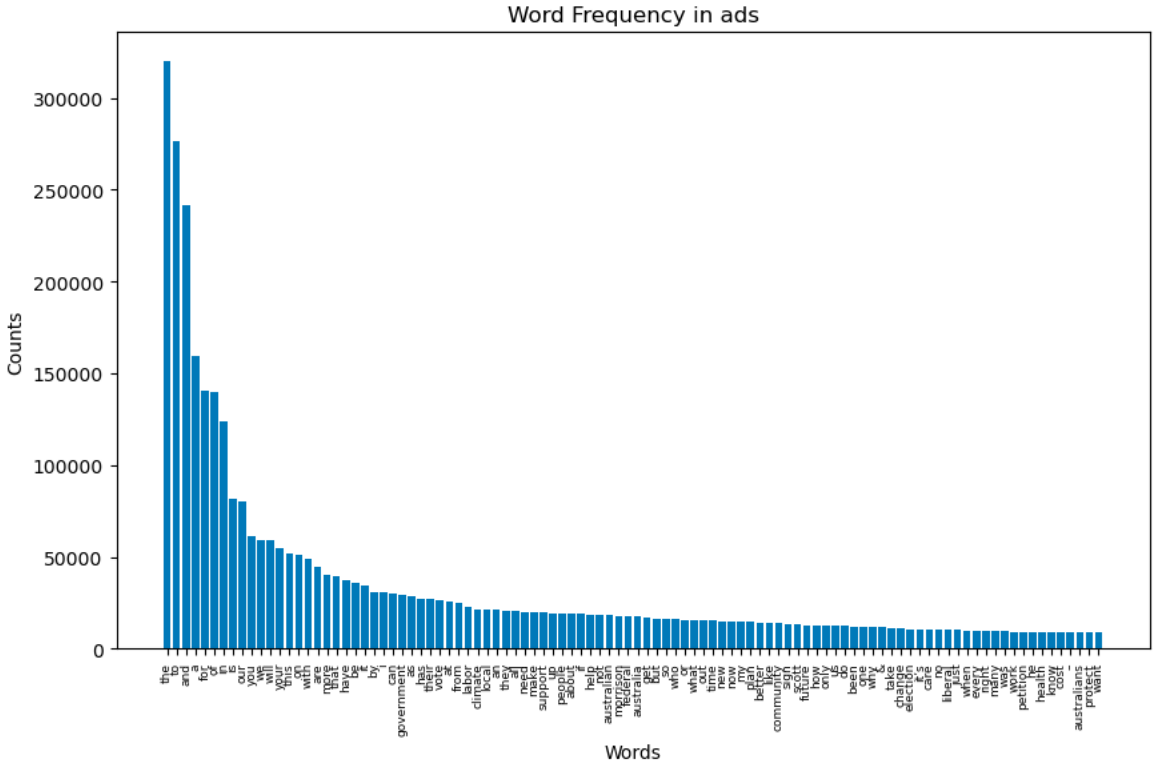


*Figure 2. Word Frequency in ads (without stop removal)*

| word | count |
|---|---|
| government | 263113 |
| help | 253264 |
| make | 229714 |
| need | 228298 |
| support | 222382 |
| people | 217059 |
| climate | 215630 |
| labor | 199297 |
| vote | 198007 |
| local | 193480 |
| sign | 181036 |
| get | 178485 |
| australian | 176257 |
| time | 163952 |
| australia | 161882 |
| like | 157178 |
| federal | 154833 |
| new | 150654 |
| community | 149056 |
| us | 141102 |

*Figure 3. Word Frequency Table (only showing top 20 rows)*


*Figure 4. Word cloud of the Ads*

## 2.6 Top Spending Political Campaigns Analysis

This analysis identified the funding entities with the highest spending and discerned any notable trends in spending patterns as the election date approached. Using data in the pre-processed dataset 'df_clean', I first calculated the monthly total spending of each entity. Then, I aggregated the total spending for each entity across the four months. Subsequently, by sorting the total in descending order, I selected the top 10 highest-spending funding entities across 4 months, as shown in Figure 5. I merged this Dataframe with these entities' monthly spending using inner join, exported it as a CSV file, and plotted their spending trends across the election period in Python.

Figure 6 shows that the Labor Party, United Australia Party and the Liberal Party emerged as the sole political parties among the top 10 funding entities. Notably, the Labor Party exhibited the highest advertisement spending, experiencing a significant surge from March (nearly 0) to April (almost 4 million AUD) and further to May (over 8 million AUD). The United Australia Party followed, peaking at 4 million in May, while the Liberal Party observed a relatively smaller increase. This evidence supports the fact that the Australian Labor Party achieved a majority government in the federal election of 2022 by spending the most on advertising for effective campaigning.

| funding_entity | sum(total_spend) |
|---|---|
| Australian Labor Party | 1.26940215E7 |
| United Australia Party | 7406032.5 |
| Australian Electoral Commission | 4619260.5 |
| Liberal Party of Australia | 2287917.5 |
| Solutions for Australia | 2110110.5 |
| Liberal Party of Australia (Victorian Division) | 1762441.5 |
| Liberal National Party of Queensland | 1692410.5 |
| Amnesty International Australia | 1657240.5 |
| Greenpeace Australia Pacific | 1650746.0 |
| Climate 200 | 1565760.5 |

*Figure 5. Top 10 highest-spending entities*

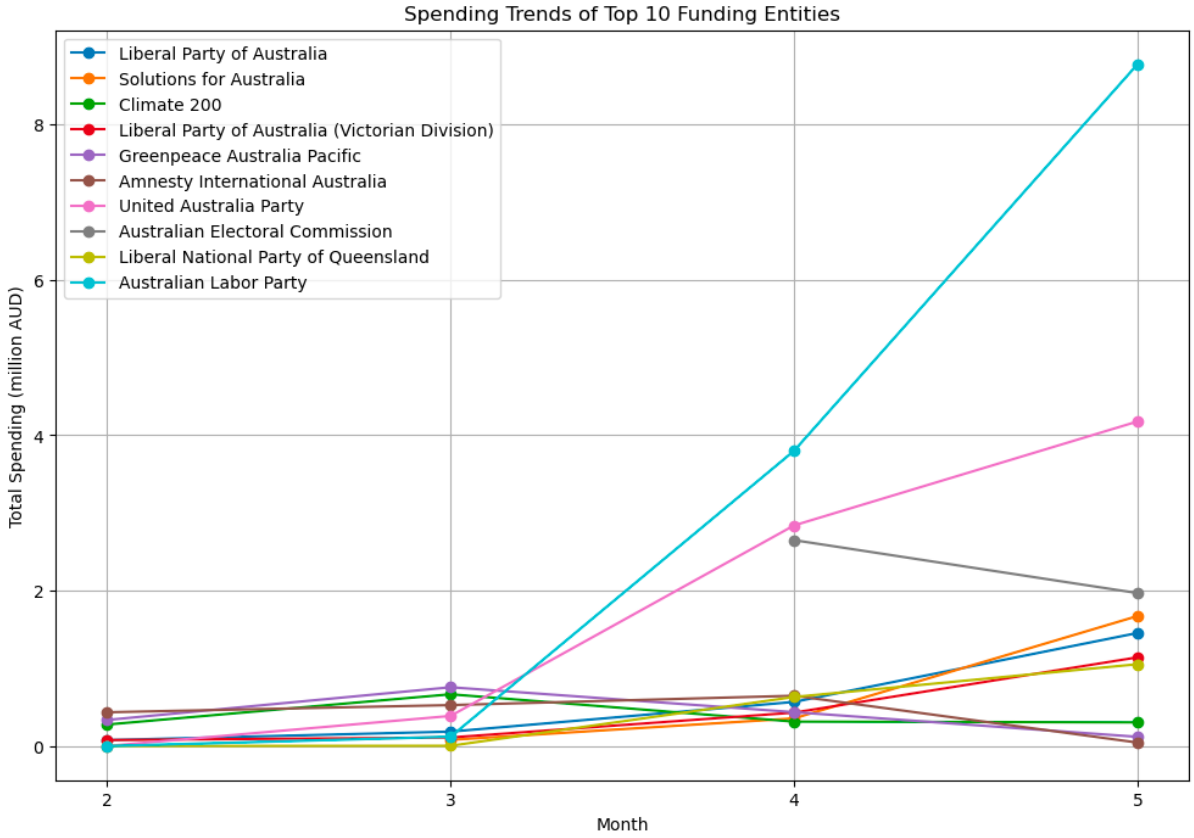


*Figure 6. Spending trends of top 10 highest-spending entities*

## 2.7 Spending Analysis of Seated Political Parties

Previously, only part of the political parties were analyzed, namely the Labor Party, United Australia Party and the Liberal Party. In this section, I conducted a more comprehensive analysis of the relationship between the spending patterns of seated parties and their positive outcomes. Based on the pre-computed aggregated spending of each entity in each month, I filtered the 'funding_entity' column to only include campaigns representing parties that secured seats in the Australian House of Representatives. Seven parties were seated,

including the Labour, the Liberal, the Greens, the National, the Centre Alliance, the Independent and the Katter's, as shown in labels in Figure 7.

In Figure 7, political campaigns representing the Labor had spent the most throughout the election period, exceeding 10 million AUD in May. It doubled that of the Liberal which ranked second, followed by the National. The skyrocketing surges in spending of the former two were the most obvious. This result positively correlated to the election outcome, with the Labor winning 77 seats and the Liberal-National Coalition winning 58 seats. By aggregating the total spending of each entity in 4 months, Figure 8 reveals the top 10 highest spending campaigns amongst the seated parties and their corresponding total spending. As expected, the Labor Party was the top spender, accounting for 12 million AUD. While the Labor and the Liberal dominated the ranks, their strategies had an important impact on shaping the electoral outcome.

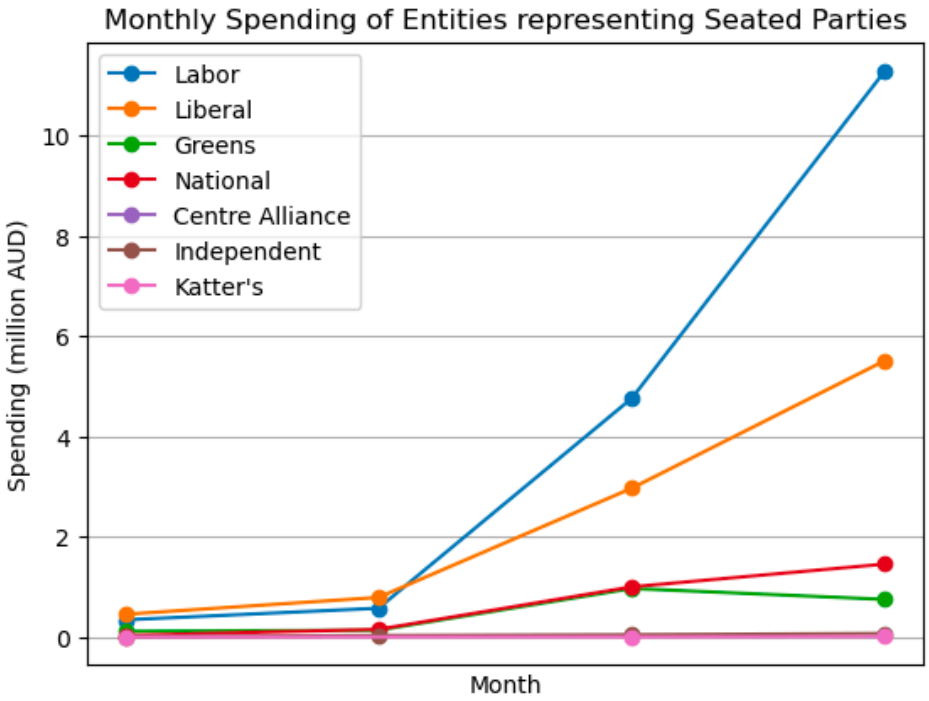


*Figure 7. Spending trends of entities representing various seated political parties*

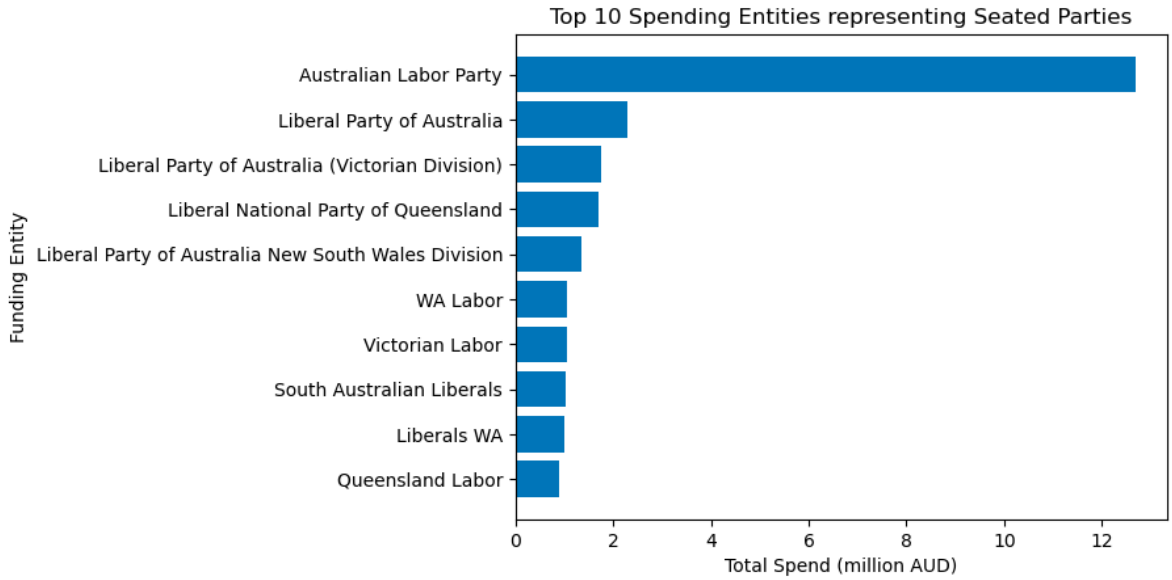


*Figure 8. Total spending of top 10 entitles representing seated parties*

# 3. Discussion and Conclusion

In conclusion, this project reveals the significance of political campaign spending in influencing voter engagement and election outcomes, particularly in the Australian Federal Election of 2022. By leveraging techniques like HDFS and PySpark, we achieved efficient data storage, big data processing and comprehensive analysis of campaign dynamics.

One notable observation is the trend of escalating investment in advertising by political campaigns as the election date approached. This strategic move suggests the critical importance of maximizing outreach before the election date. Moreover, campaigns could also draw inspiration from keywords and action words in advertisements to connect with and mobilize voters. An illustration was observed in the significantly high spending and frequent appearance in advertisements by the Labor Party during April and May, indicating an increase in their exposure through advertisement campaigns. This strategy coincided with the party's remarkable success in securing the majority of seats and ultimately attaining the position of prime minister. The advancements in technology boost political campaign reach via targeted advertisements, engaging voters directly, and spreading messages efficiently.

Total words: 1504

# Reference

Taylor, P. (2023, November 16). *Data Growth Worldwide 2010-2025*. Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, *21*, 1112-1130.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

# Appendix

Please see next page.