TASK REPORT

SUMMARY

DATA MINING: 2024 U.S. ELECTION

This task focuses on predicting the possible outcomes of the 2024 U.S. election. It was prepared based on the author's skills and statistical perspective. The project utilized the R programming language. With more experience, better results could be achieved.

INTRODUCTION

Data is a fundamental element that is stored, processed, and analyzed. Data typically represents information in numerical or symbolic formats and is collected from various sources by organizations, businesses, and individuals. Data production refers to the process of gathering, creating, or recording information. This process can occur through sensors, user interactions, software logs, surveys, or other data collection methods. After data production, the steps of data storage, cleaning, analysis, and interpretation follow. Data production and management serve various objectives such as making informed decisions, identifying patterns, tracking trends, and improving business processes. This process enables businesses to gain competitive advantages and make more informed decisions in today's information age.

The 2024 U.S. election is a significant element of democracy, and data mining techniques are used to analyze and understand the election processes. This study aims to conduct a comprehensive data analysis of the 2024 U.S. election using the R programming language and data mining tools. Initially, data related to the election will be collected from various sources, and a dataset will be created. This dataset will include factors such as candidates' campaign activities, public sentiment, social media activities, and poll results. Subsequently, the dataset will be analyzed using the R programming language and data mining libraries. Techniques such as statistical analysis, data visualization, and sentiment analysis will be applied to understand the dynamics of the election process. Various analyses will be conducted to gain an in-depth understanding of candidates' popularity, voter behavior, campaign strategies, and other key factors. The results will evaluate whether the findings and trends obtained through data mining methods contribute to understanding the election process and assess the potential impact of these analyses on future election strategies. This study aims to provide a valuable resource for policymakers, analysts, and researchers by demonstrating how data mining and the R programming language can be effectively used in election analysis.

R PROGRAMMING LANGUAGE

R is an open-source programming language used in the fields of statistical computation, data analysis, and graph creation. As a vector-based language, R is particularly powerful for

manipulating datasets and conducting statistical analysis. Statisticians, data scientists, and researchers can effectively use R in data mining and analytical projects due to its wide collection of packages and statistical functions. Additionally, its rich graphical capabilities make it a popular choice for visualizing results. Being open-source, R is supported and continuously developed by a large community.

YOUTUBE DEVELOPER ACCOUNT

The YouTube Developer Account is a tool for developers who wish to integrate their applications with the YouTube platform using YouTube APIs. These APIs allow developers to programmatically access various features of YouTube and integrate their applications with the platform.

LIBRARIES USED IN R

Figure 1: Libraries used in R

CODE USED IN R

After activating the libraries in R, the environment was prepared as shown in Figure 2 to link the YouTube Developer account to R and pull data from YouTube.

Figure 2: Linking YouTube Developer Account with R

The video from which the data is pulled and the conversion of comments into CSV format is illustrated in Figure 3.

Figure 3: Conversion of video key and comment data into CSV
To ensure the pulled comments are suitable for analysis, a cleaning process was conducted.

Figure 4: Comment Cleaning Process

CONVERTING COMMENTS INTO A TABLE OF WORDS

The cleaned comments were parsed and converted into a table of words, and stop words were removed for further cleaning, as shown in Figure 5.

Figure 5: Converting comments into a table of words and final cleaning process
A command was written to create a word cloud, as seen in Figure 6, to find the most frequently occurring words and make the prepared data meaningful.

Figure 6: Creating a word cloud with word frequencies
The polarity score of the words derived from the comments is shown in Figure 7.

Figure 7: Polarity Analysis

FINDINGS

3.1 WORD CLOUD

The word cloud resulting from the video that discusses the subject in real-time is shown in Figure 8. The number of comments obtained is 20,021. According to this word cloud, the most frequently occurring word is "Trump," followed by "Biden."

Figure 8: Word Cloud

3.2 SENTIMENT ANALYSIS

The sentiment of the election results is depicted in Figure 9. Values close to +1 indicate positive sentiment, 0 indicates neutral, and values close to -1 indicate negative sentiment. Based on the analysis result, the sentiment analysis shows a score of -0.0345, indicating that the sentiment is neutral but slightly leaning toward the negative. The red line represents the overall sentiment, while the blue bars represent each word's sentiment score, ranging from -1 to 1.

Figure 9: Sentiment Analysis Result

The mathematical breakdown of the analysis is provided in Figure 10.

Figure 10: Mathematical Results

CONCLUSION AND DISCUSSION

The popular candidates in the election have shifted over time, and candidates' media appearances have influenced changes in opinions. Biden, who once ended Trump's era, has slowly lost his trend status due to his social media presence, causing people to gradually turn toward Trump. For this reason, the most frequently repeated word is "Trump." However, when we look at the most frequently repeated words in order, we can infer that people still desire a Democratic administration. Thus, although Trump, a Republican, seems to be gaining momentum, a difficult election awaits him. Based on the average sentiment score, it can be said that there is a slightly negative sentiment development, indicating a slow decline in confidence toward Trump.

This study was conducted using comments from the period when both candidates were experiencing turbulent times, focusing on videos that received the most engagement. A total of 2,542 comments and 20,021 words were analyzed. Results may change significantly if more up-to-date data is considered. As the study is based on a global news source, people

from different countries who do not favor either candidate also expressed their opinions. More accurate results can be obtained through more localized surveys.