Logo, company name

Description automatically generated

**Text Mining Group Project**

**OPIM-5671- Data Mining and Business Intelligence**

**Professor Sudip Bhattacharjee**

**UNVEILING THE VOICES OF TWITTER: Sentiment Analysis of 2020 US Presidential Election Tweets**

**Group 5**  
Rohit Akole  
Ananya Shrivastava

Shivom Moyade

Chinmay Khurana

Rajat Kadam

Table of Contents

[Abstract 4](#_Toc162917693)

[Introduction 5](#_Toc162917694)

[Problem Statement 6](#_Toc162917695)

[Dataset 8](#_Toc162917696)

[I. Introduction to Dataset 8](#_Toc162917697)

[II. Data Source 8](#_Toc162917698)

[III. Original Dataset Schema: 8](#_Toc162917699)

[1. Bidenall2.csv 8](#_Toc162917700)

[2. Trumpall2.csv 9](#_Toc162917701)

[IV. Data Exploration and Pre-Processing: 9](#_Toc162917702)

[V. Post Preprocessing Dataset Schema: 13](#_Toc162917703)

[1. Biden\_Final.xlsx 13](#_Toc162917704)

[2. Trump\_Final.xlsx 13](#_Toc162917705)

[Unsupervised Learning Model 14](#_Toc162917706)

[I. Text Parsing 15](#_Toc162917707)

[II. Synonyms and Multi-Term Words (Emojis) 15](#_Toc162917708)

[III. Text Filter 17](#_Toc162917709)

[IV. Text Cluster 17](#_Toc162917710)

[V. Text Topics 19](#_Toc162917711)

[1. Text Topics For Trump 20](#_Toc162917712)

[2. Text Topic For Biden 21](#_Toc162917713)

[Insights and Observations 22](#_Toc162917714)

[I. Biden being overshadowed under Obama’s leadership. 22](#_Toc162917715)

[Theme 23](#_Toc162917716)

[II. Trump on Suburban Housing: 23](#_Toc162917717)

[Theme 24](#_Toc162917718)

[III. Trump and Biden on absentee voting: 25](#_Toc162917719)

[Theme 25](#_Toc162917720)

[IV. Reactions to Biden and Trump’s candidature for the 2024 elections: 27](#_Toc162917721)

[V. Recent tweets in support of Trump: 28](#_Toc162917722)

[Business Case Recommendations 29](#_Toc162917723)

[I. Political Campaign Strategy 29](#_Toc162917724)

[II. Market Research 29](#_Toc162917725)

[III. Risk Management 30](#_Toc162917726)

[Conclusion 31](#_Toc162917727)

# Abstract

The following project revolves around popular topics and public sentiment of Election 2020 Tweets. The tweets are the public replying to @JoeBiden and @realDonalTrump tweets before the election in 2020. The unsupervised model is used to demonstrate the popular text topic themes around the election period.

***Keywords:*** *Unsupervised, Tweets, Election 2020, Tweeter, text topic, themes, sentiment*

# Introduction

Our text mining project focuses on analyzing sentiment in tweets related to the 2020 US Presidential Election, specifically concerning candidates Donald Trump and Joe Biden. Leveraging the power of SAS Enterprise Miner, we aim to conduct sentiment analysis to uncover insights into public opinion, emotions, and attitudes towards candidates during this crucial political event. We intend to identify the major themes in the tweets to gauge the big talking points and concerns of the citizens. The dataset comprises a rich collection of tweets representing a diverse range of viewpoints, discussions, and reactions from Twitter users across the nation. By utilizing SAS Enterprise Miner's robust capabilities in text mining, we seek to extract meaningful patterns, sentiment trends, and sentiment shifts related to Trump and Biden throughout the election timeline, providing valuable insights into how Twitter users perceive and express sentiments towards the presidential candidates, their policies, campaign strategies, and key election-related events. This analysis contributes to a deeper understanding of the social media dynamics and public sentiment surrounding the 2020 US Presidential Election. Our sentiment analysis can aid political campaigns in message tailoring and geographic/demographic guidance. It could also help companies identify market segments, inform marketing strategies, and manage risks in politically sensitive areas for informed decision-making.

# Problem Statement

Despite the vast amount of data available on social media platforms like Twitter, understanding and interpreting public sentiment towards political candidates and key election-related topics during significant events such as the 2020 US Presidential Election remains challenging. The sheer volume of tweets and the diversity of viewpoints expressed make it difficult to discern prevailing sentiments and identify crucial themes.

Our project aims to address this challenge by leveraging text mining techniques with SAS Enterprise Miner to conduct sentiment analysis on tweets related to the 2020 US Presidential Election, focusing specifically on the sentiments towards candidates Donald Trump and Joe Biden. We seek to extract meaningful patterns, sentiment trends, and shifts in public opinion from this rich dataset, ultimately providing insights into how Twitter users perceive and express sentiments towards the candidates, their policies, campaign strategies, and key election-related events.

By doing so, we aim to contribute to a deeper understanding of the social media dynamics and public sentiment surrounding the 2020 US Presidential Election, enabling political campaigns to tailor messages effectively and assisting companies in identifying market segments, informing marketing strategies, and managing risks in politically sensitive areas for informed decision-making.

Using the unsupervised learning techniques for our research on 2020 US Presidential Election, we will focus on analyzing tweets related to the election without establishing a predefined target variable. Our objective is to explore the sentiments and themes present in the tweets, particularly those concerning candidates Donald Trump and Joe Biden.

During this phase, we will leverage text mining techniques to identify recurring themes, sentiments, and patterns within the vast collection of tweets. By examining the language used by Twitter users across various geographical locations and demographic groups, we aim to gain insights into the prevailing public opinion, emotions, and attitudes towards the candidates and key election-related topics.

Furthermore, we will explore the sentiment trends and shifts throughout the election timeline, providing valuable insights into how Twitter users perceive and express sentiments towards the candidates, their policies, campaign strategies, and significant election events.

# Dataset

## Introduction to Dataset

The source of our dataset is Kaggle, where data was scraped from Twitter website using the Twitter API during the 2020 Presidential Elections. The dataset contains over 2300 tweets referencing Biden and 2700 tweets referencing Trump. The dataset has 2 primary variables, namely the user handle and the tweet associated with it. We intend to run these separately to identify major themes and patterns in the tweets.

## Data Source

Dataset: [Kaggle Dataset Link](https://www.kaggle.com/datasets/noorsaeed/usa-election-sentiment-analysis-dataset/data)

## Original Dataset Schema:

### Bidenall2.csv

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| User | Input | User ID of a Twitter user. This column identifies the unique user associated with each tweet, allowing for analysis of user-level interactions and behaviors. |
| text | Text | Tweet |

### Trumpall2.csv

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| User | Input | User ID of a Twitter user. This column identifies the unique user associated with each tweet, allowing for analysis of user-level interactions and behaviors. |
| text | Text | Tweet |

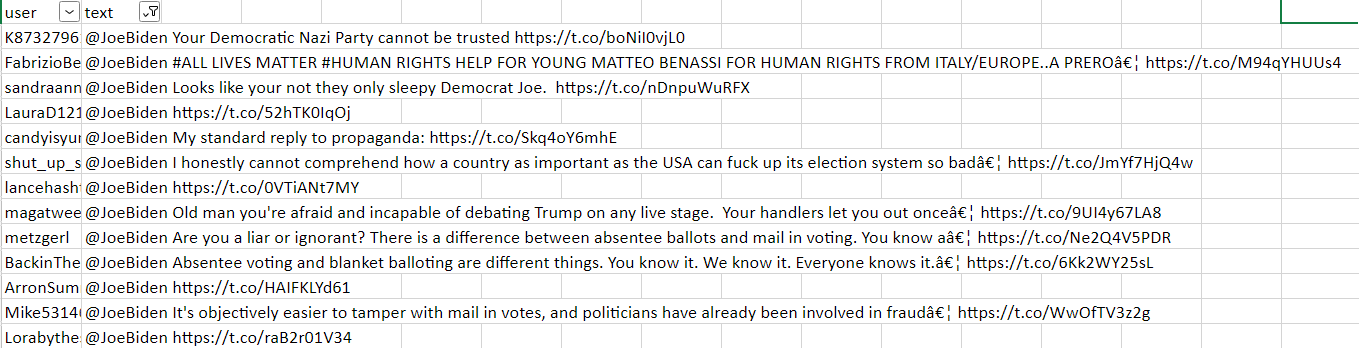
## Data Exploration and Pre-Processing:

We will utilize SAS Enterprise Miner Workstation for our data analysis, complemented by Python for data preprocessing tasks. During the data cleaning phase, we encountered tweets containing appended links leading to images, gifs, or other miscellaneous media. Additionally, our dataset featured a plethora of emojis, which we recognized as crucial elements conveying sentiments and emotions.

Understanding the significance of emojis in expressing nuanced feelings and attitudes, we made a deliberate decision to retain this aspect of the data. Emojis serve as valuable indicators of sentiment and can offer deeper insights into the emotional tone of tweets. Therefore, we took meticulous care to preserve the presence of emojis throughout our dataset, ensuring that our analysis captures the full spectrum of sentiment expressed by Twitter users during the 2020 US Presidential Election.

We identified emojis and mapped them using python programming to replace the symbolic emojis with their meaning. We created the list of emojis as given by the website: [Piano World](https://forum.pianoworld.com/ubbthreads.php/topics/2904611/ot-list-of-all-emojis-for-people-who-need-hands-to-talk.html)

We replaced emojis with their meanings in :< meaning>: format.

Figure 1. Raw Data with Links and Emojis

For removing emojis from the code, we used python for emoji mapping, the code for which is below:

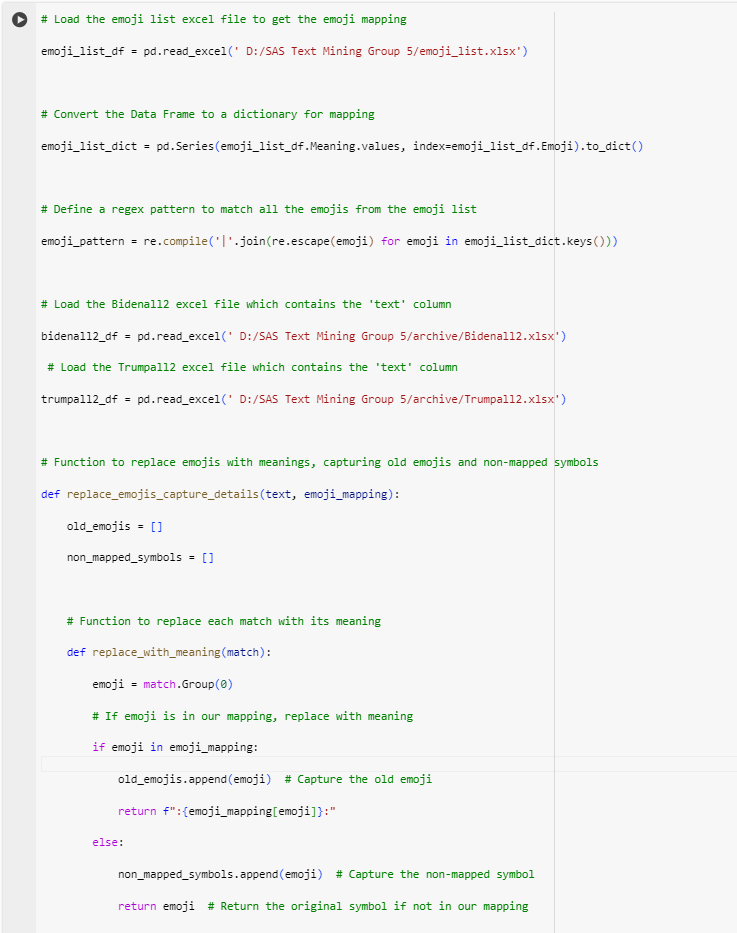


Figure 2. Python Code - Emoji Mapping Part 1

Figure 3. Python Code - Emoji Mapping Part 2

In addition to the following, we also took the following measures for cleaning data:

* Original text included Apostrophe (‘), Double Quotes (“”), Ellipses (...) in symbolic languages, we converted them into appropriate symbol.
* Removed the rows where only @JoeBiden or @realDonaldTrump was mentioned assuming the tweet must be in image format hence the row had no text data in them.
* Saved the final file in “D:/SAS Text Mining Group 5/Final\_Files” after cleaning the data.
* Added Emojis in Multi term from Text Parsing Node.

## Post Preprocessing Dataset Schema:

### 1. Biden\_Final.xlsx

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Role** | **Description** |
| user | Input | User Id of a twitter user |
| text | Text | Tweet |
| old\_emojis | Rejected | Emoji from text (Old format) |
| non\_mapped\_symbols | Rejected | Emoji not mapped |

### 2. Trump\_Final.xlsx

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Description** |
| user | Input | User Id of a twitter user |
| text | Text | Tweet |
| old\_emojis | Rejected | Emoji from text (Old format) |
| non\_mapped\_symbols | Rejected | Emoji not mapped |

# Unsupervised Learning Model

As our data set does not contain any target variable, we chose to do unsupervised learning and explore sentimental opinion of the Twitterati on the two candidates. We ran two models, one for Biden and the other for Trump.

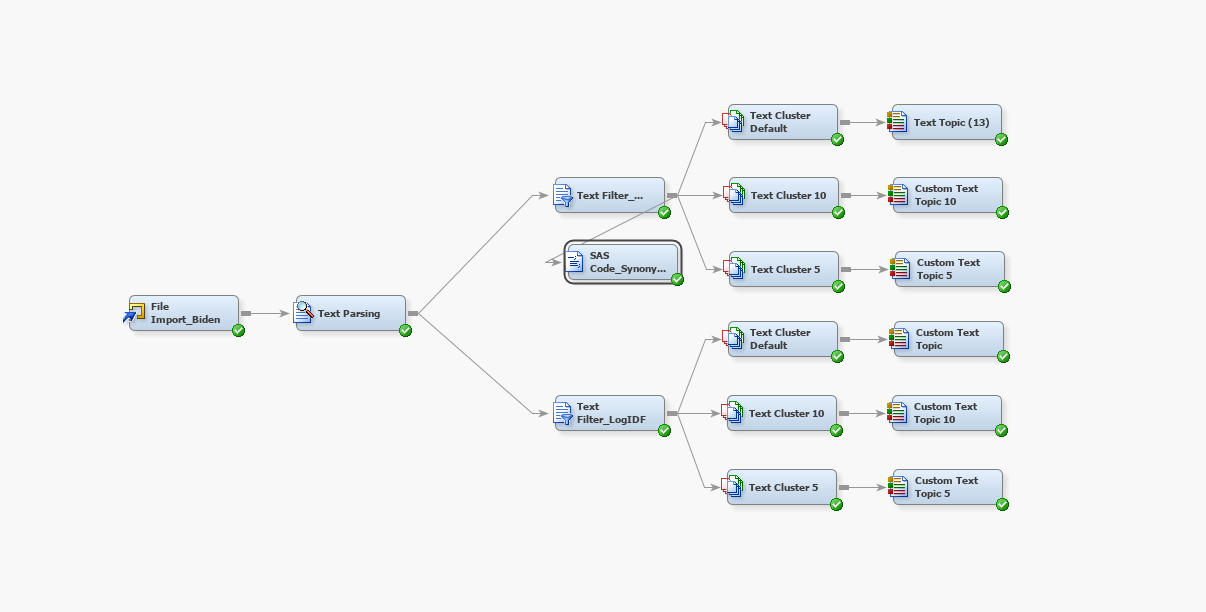


Figure 4. Unsupervised Model for Joe Biden

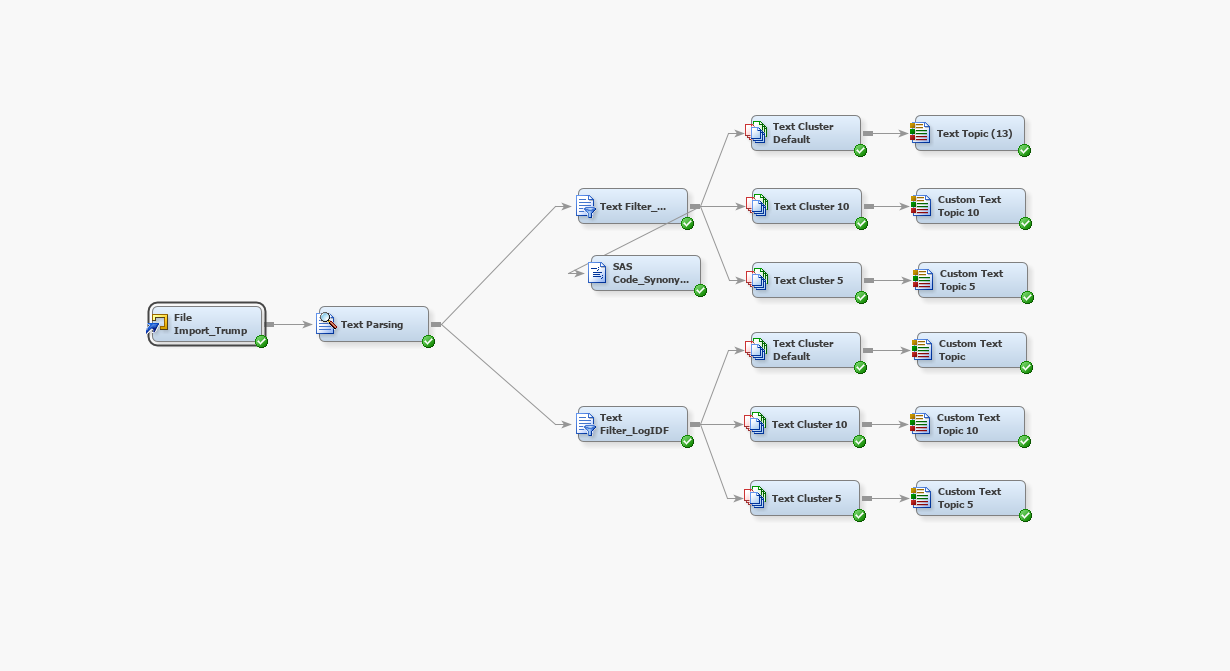


Figure 5. Unsupervised Model for Donald Trump

## Text Parsing

In the text parsing node, we made changes to the default settings. We created synonyms using the table we got from Text Filter’s Spell-Checking Results table. We exported the synonyms table to a SAS table using SAS Code node. We added processed emojis into Multi-Word Terms as shown below. We also created a Stop list by going through all the tweets given in the data.

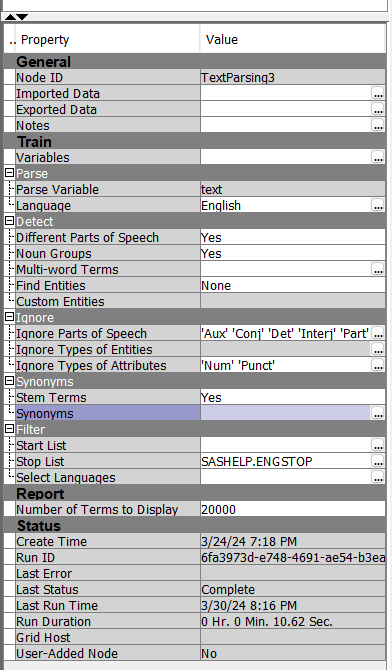


Figure 6. Text Parsing Node Properties

## Synonyms and Multi-Term Words (Emojis)

We had numerous tweets that ended with emojis. To simplify our analysis and improve readability during text parsing, we converted emojis into text using the text translation available on the '[Piano World](https://forum.pianoworld.com/ubbthreads.php/topics/2904611/ot-list-of-all-emojis-for-people-who-need-hands-to-talk.html)' website, which covers the translation of all emojis. Subsequently, we incorporated these phrases into the default list.

|  |  |
| --- | --- |
| Figure 7. Declaring synonyms | Figure 8. Defining emojis in multi word terms |

## Text Filter

We tried using different Term weights like Entropy and IDF to observe different results in our dataset.

|  |  |
| --- | --- |
| Figure 9. Text Filter Node Log-Entropy | Figure 10. Text Filter Node Log-IDF |

We tested different term weights like Entropy and IDF to see how they affected our dataset. After comparing the results, we found that both approaches produced similar outcomes in terms of text topics. As a result, we decided to move forward with using Entropy as our preferred term weight.

## Text Cluster

In the text cluster we have considered different SVD resolutions like default, 10 and 5, number of clusters default, 10, and 5 respectively and the multi term topics as default, 10, and 5, respectively. The following were settings for Biden and Trump:

|  |  |
| --- | --- |
| Figure 11. Text Cluster SVD 10  **Number of Clusters:** 10  **SVD Resolution:** High  **Max SVD Dimensions:** 10 | Figure 12. Text Cluster SVD 5  **Number of Clusters**: 5  **SVD Resolution:** High  **Max SVD Dimensions:** 5 |
| Figure 13. Text Cluster SVD Default  **Number of Clusters:** 40  **SVD Resolution:** Low  **Max SVD Dimensions:** 100 | Figure 14. Text Cluster SVD 5  **Number of Clusters:** 5  **SVD Resolution:** High  **Max SVD Dimensions:** 10 |
| Figure 15. Text Cluster SVD 5  **Number of Clusters:** 25  **SVD Resolution:** High  **Max SVD Dimensions:** 5 | Figure 16. Text Cluster SVD Default  **Number of Clusters:** 40  **SVD Resolution:** Low  **Max SVD Dimensions:** 100 |

## Text Topics

The topics clearly identifiable from the text topic nodes are highlighted in the below images:

As we were going through different variations of the models, we determined that the model with 10 SVD Dimensions and 10 text topics were giving us unique and better results compared to the default model with 25 multi-term text topics and 5 multi-term text topics. We were getting many duplicate topics in 25 text topics and too much saturation in 5 multi-term text topics.

### Text Topics For Trump

Figure 17. Text Topic 10

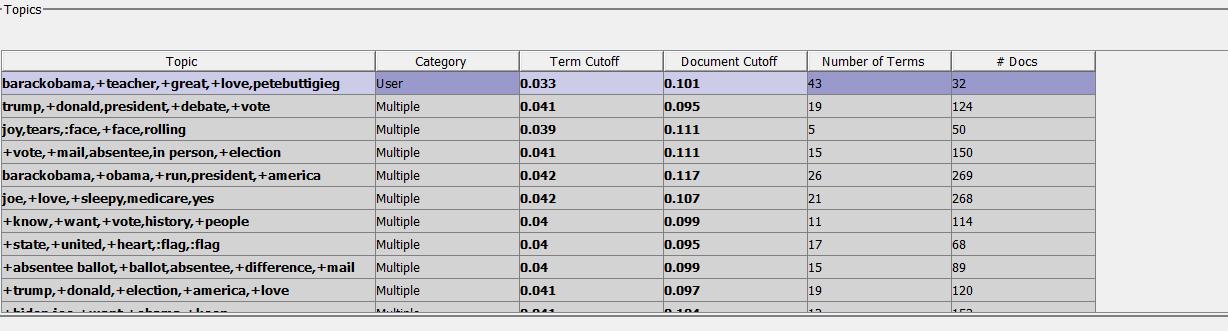
Figure 18. Text Topic 5

Figure 19. Text Topic Default

### Text Topic For Biden

Figure 20. Text Topic 10

Figure 21. Text Topic 5

Figure 22. Text Topic Default

# Insights and Observations

As we tried different combinations for SVD dimensions, weights, and frequencies, we got some interesting insights from our observations. From the tweets that were garnered from our data, we decided to backtrack and look at the source tweet for each candidate and were able to gain deeper context on the topic and theme in concern and how the audience was inclining to one side or another.

Some key examples of those tweets and the relevant group is given below:

## Biden being overshadowed under Obama’s leadership.



Figure 23. Joe Biden’s Original Tweet

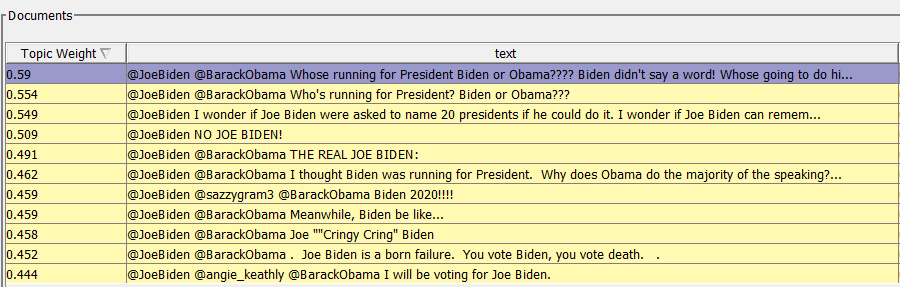


Figure 24. Public Reaction Tweets related to the above tweet

After reviewing tweets across clusters generated through Text Clustering analysis, we found this tweet. The prominent keywords associated with this tweet included: "Biden," "teacher," "run," "VP," "president," and "Obama."

### Theme

This cluster's theme revolves around the perception of Joe Biden being overshadowed by his association with Barack Obama. Twitter users expressed criticism towards Biden for potentially lacking independence in decision-making, contrasting this with admiration for Obama's leadership style and principles. Obama is lauded as a great teacher, highlighting his influence and perceived superiority in guiding political ideologies. The juxtaposition of Biden's role and Obama's stature as a mentor formed a compelling narrative within this cluster, offering intriguing insights into public perceptions and discourse surrounding the 2020 US Presidential Election.

## Trump on Suburban Housing:

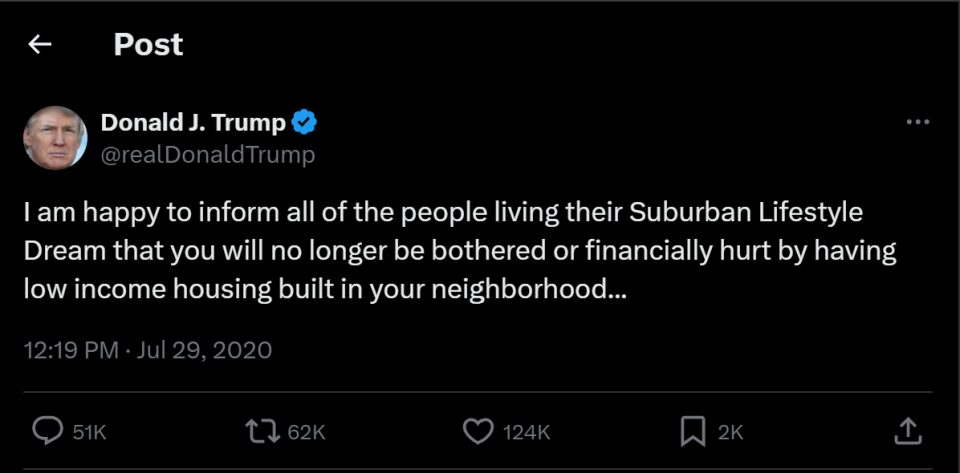


Figure 25. Donald Trump’s Original Tweet

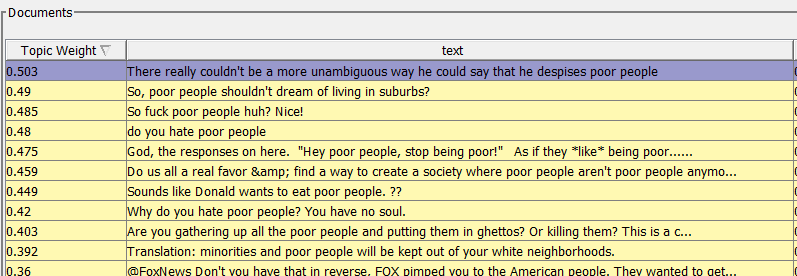


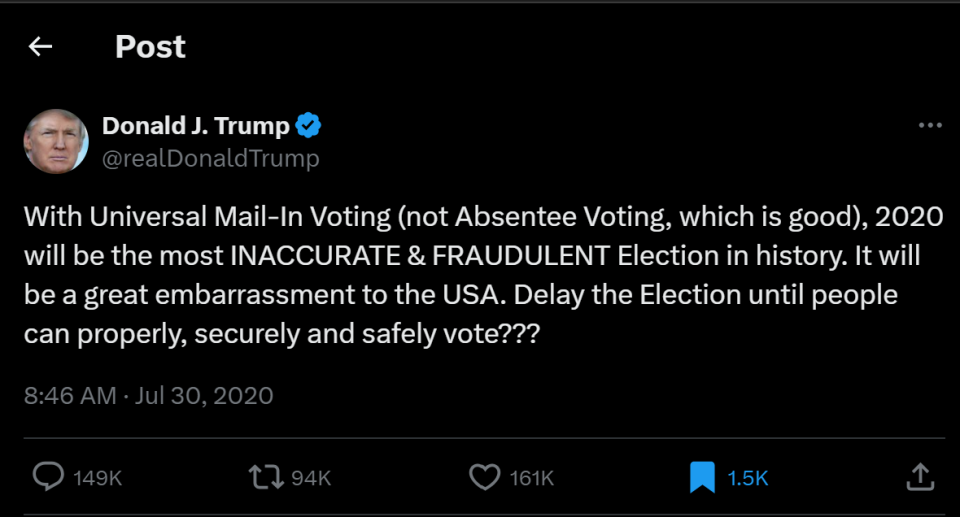
Figure 26. Public Reaction Tweets related to the above tweet

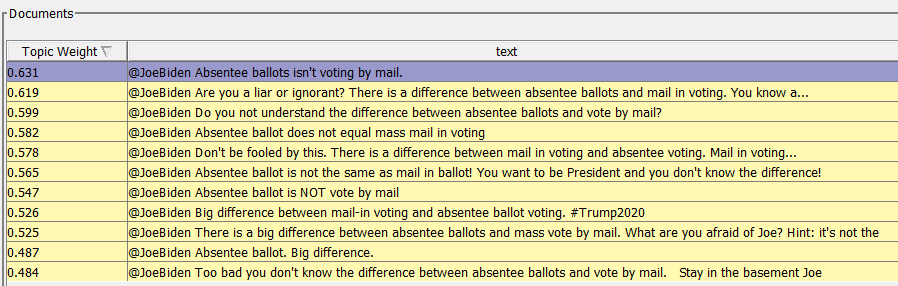
The tweet in question features prominent keywords such as "Trump," "poor," "people," "hate," and "housing."

### Theme

The overarching theme of this tweet and others within its cluster revolves around mounting concerns regarding escalating housing prices, particularly impacting low-income individuals and residents supportive of Trump's suburban lifestyle. These tweets shed light on the challenges faced by individuals belonging to economically disadvantaged backgrounds and highlight the frustration felt by those grappling with rising housing costs. The issue of soaring housing prices emerges as a central focus within this cluster, reflecting broader societal concerns and implications for marginalized communities.

## Trump and Biden on absentee voting:

Figure 27. Donald Trump’s Original Tweet

Figure 28. Public Reaction Tweets related to the above tweet

The tweet in question highlights significant keywords such as "absentee ballot," "vote," "mail," and "in-person."

### Theme

The theme surrounding absentee voting emerged as a prominent aspect of the 2020 elections, drawing attention from both Trump and Biden supporters. Twitter users engaged in discussions and debates regarding the merits and pitfalls of absentee and in-person voting methods, with each candidate espousing their respective views. Trump's attempts to delay the elections, perceived as motivated by fear of losing, drew criticism from Twitter users, while Biden faced scrutiny for allegedly misunderstanding the distinction between absentee ballots and mail-in voting.

Furthermore, our analysis revealed notable trends in sentiment towards both Trump and Biden. During Trump's tenure, Biden received approximately 30% more positive and supportive tweets. However, in the current election scenario, we observed a shift, indicating that Biden, as the incumbent President, faces heightened scrutiny from the public. This observation suggests that the current administration is subject to more criticism compared to the opposition party, reflecting Twitter users' expectations for improvement and change.

Additionally, we have compiled reactions to Biden and Trump's candidature for the 2024 elections, providing further insights into public sentiment and anticipation surrounding future political developments.

## Reactions to Biden and Trump’s candidature for the 2024 elections:

|  |  |
| --- | --- |
| Figure 29. Biden’s Original Tweet | Figure 30. Biden’s Original Tweet |
| Figure 31. Biden’s Original Tweet | Figure 32. Biden’s Original Tweet |

## Recent tweets in support of Trump:

|  |  |
| --- | --- |
| Figure 33. Public on Trump 2024 Tweet | Figure 34. Public on Trump 2024 Tweet |

# Business Case Recommendations

## Political Campaign Strategy

Scenario: During the 2020 elections, sentiment analysis reveals a significant volume of negative tweets about Donald Trump's approach to COVID-19 policies, indicating public concern over the administration's response to the pandemic.

* **PR and Messaging Shift:** The Trump campaign could have responded by adjusting its public relations strategy to directly address these concerns. This might include producing advertisements and social media content that highlight specific actions taken to combat the pandemic, such as securing vaccines or implementing relief measures.
* **Policy Emphasis:** Shift campaign focus to emphasize plans for pandemic response, including clear, step-by-step recovery plans, healthcare support, and economic recovery measures, communicated through targeted emails, social media campaigns, and televised addresses in areas with the most expressed concern.

## Market Research

Scenario: A sentiment analysis of tweets related to the 2020 elections uncovers positive sentiments towards Joe Biden's stance on healthcare reform, particularly among middle-aged voters in suburban areas.

* **Healthcare Product Alignment:** Companies in the healthcare sector, especially those offering insurance or telehealth services, can use this insight to tailor their marketing strategies. Campaigns could emphasize how their services align with the ideals of healthcare reform, such as affordability and accessibility, highlighting benefits that resonate with the values of this demographic.
* **Segmented Marketing Campaigns:** Develop segmented marketing campaigns that address the specific healthcare concerns identified in the sentiment analysis, such as coverage for pre-existing conditions or affordable prescription plans. Utilize platforms popular with suburban middle-aged demographics, like Facebook and local television stations, for these campaigns.

## Risk Management

Scenario: Sentiment analysis indicates widespread concern over economic policies and their impact on small businesses among tweets from urban business owners regarding the 2020 election outcomes.

* **Proactive Communication:** Businesses in financial services or consultancy can leverage these insights to offer targeted advice and products designed to help small businesses navigate uncertain economic times. This could include webinars on financial planning post-election, special service packages for risk assessment and management, and direct consultations to prepare for policy changes.
* **Advocacy and Partnership:** Align with or create advocacy groups to lobby for policies that support small businesses, using data from sentiment analysis to underscore the importance of such policies. Partner with local chambers of commerce to offer workshops and resources that help businesses adapt to the new economic policies, reassuring them through proactive support and guidance.

# Conclusion

The analysis highlighted fluctuating public sentiments towards Donald Trump and Joe Biden, reflecting the dynamic nature of public opinion throughout the election period. Sentiment trends were influenced by campaign events, policy announcements, and the candidates' public statements. The sentiment analysis revealed the key differences in public opinion. These variations underscore the importance of customized campaign strategies and the need for targeted messaging to address the concerns of specific voter segments.  
The unsupervised learning approach identified major themes and concerns among the electorate, such as healthcare, economy, and social justice issues. Understanding these themes is crucial for aligning campaign messages and policies with voter priorities.