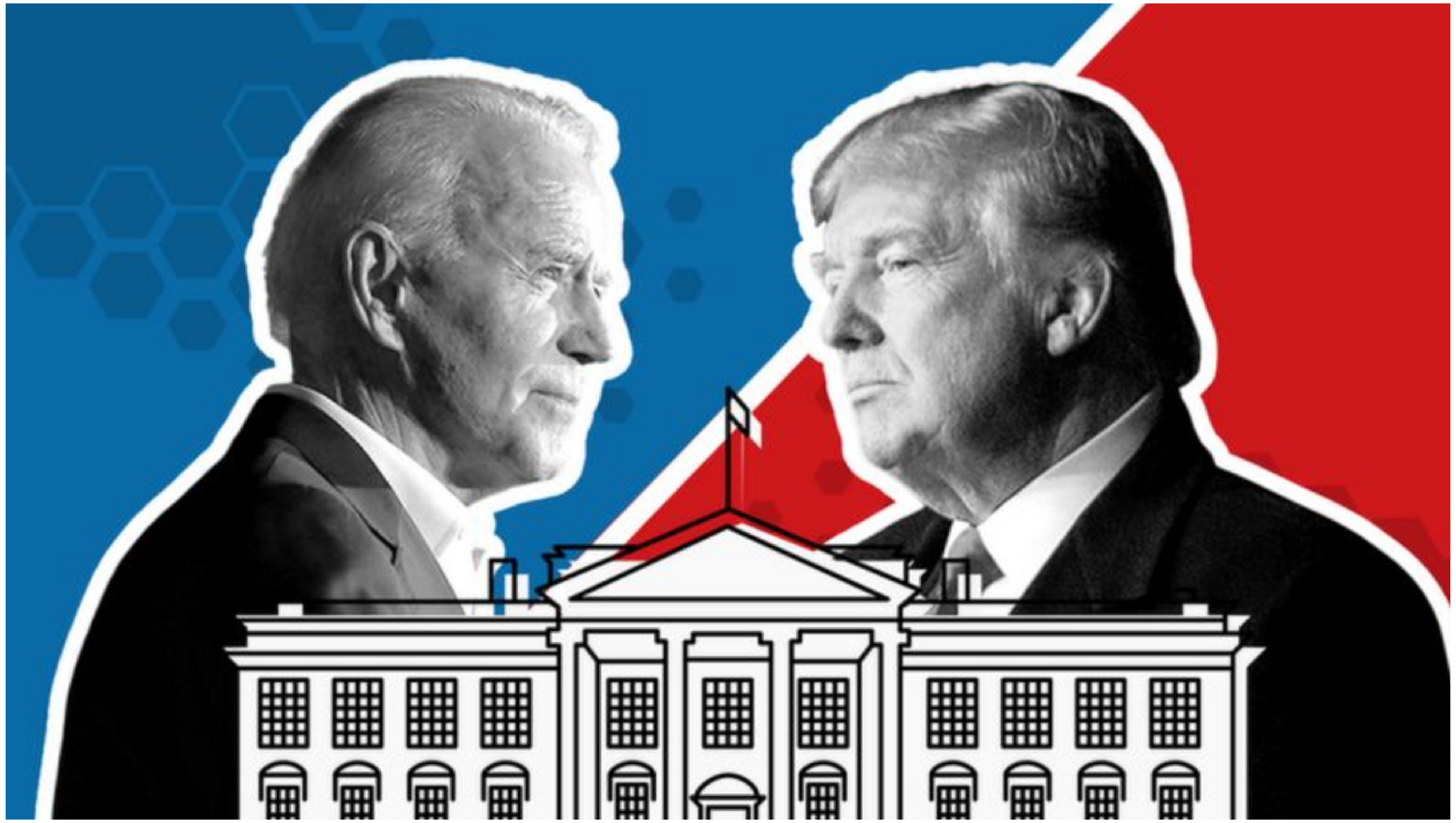
2020 US Elections

Web and Social Media Analysis

Team Members:
Valeria Fumagalli, Adel Kalozdi, Riccardo Pandolfi,
Francesca Ramella, Annel Saavedra

Context

- The 2020 US Election was one of the most impactful events worldwide in recent years. (Election day: 3/11/2020).
- Social media platforms, have not only changed the way we interact with one another but also the way we share news and comment on such world events.
- This is the main reason why we chose to use data collected from one of the main platforms (Twitter) to try to understand individuals' perceptions and what were the recurrent sentiments about each candidate.
- Tweets analysis can thus be used to better understand the voters' opinions and the candidates' performance, but also to improve the prediction of the election outcome.



Data Collection and Cleaning

Tweets collected using the Twitter API *statuses_lookup* and *snsscrape* for the keywords corresponding to the candidates' names.

Source: Kaggle user Manch Hui

Timeframe:
From 15/10/2020
To 08/11/2020

Geographic Area: tweets from the US only

Input data:
2 csv files (one for each candidate).
NLP cleaning: remove punctuation, special characters, lower case, stop words and emojis.
Lemmatization.

Techniques

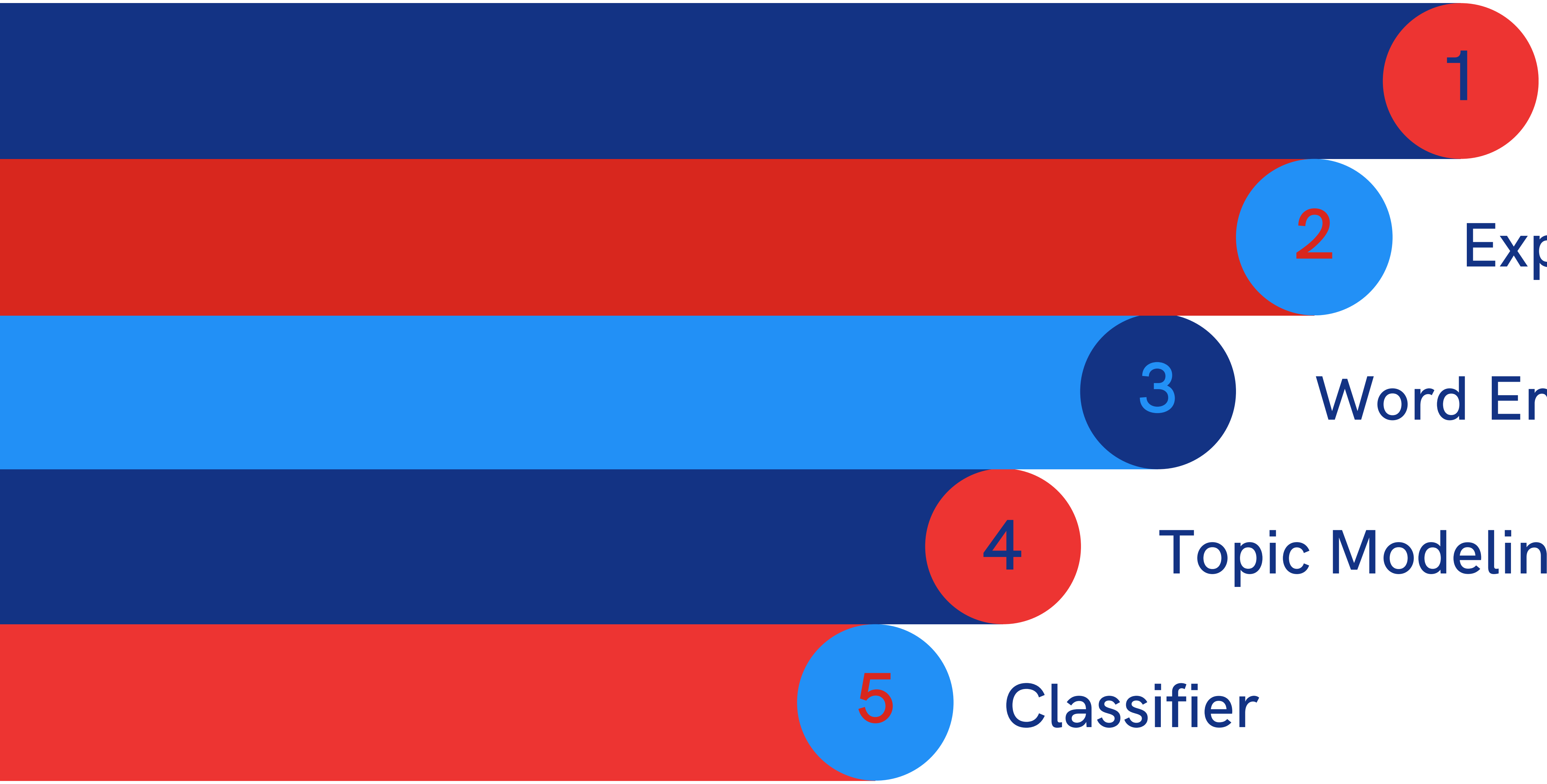
NLP Techniques

- Sentiment Analysis: labels the text in terms of polarity (negative or positive).
- Word Embedding: geometric representation of words that are related, often close to each other.
- Topic Modeling: technique to discover hidden topics in the tweets (in this case).

Why Python?

- Robust collection of NLP libraries such as: SpaCy, NLTK, Gensim. These libraries provide built in functions for text analysis
- More intuitive language
- Provides nice tools for data visualization
- Better for performing Machine Learning

Our analysis: what are the steps?



1 Sentiment Analysis

2 Exploratory Data Analysis

3 Word Embedding

4 Topic Modeling

5 Classifier

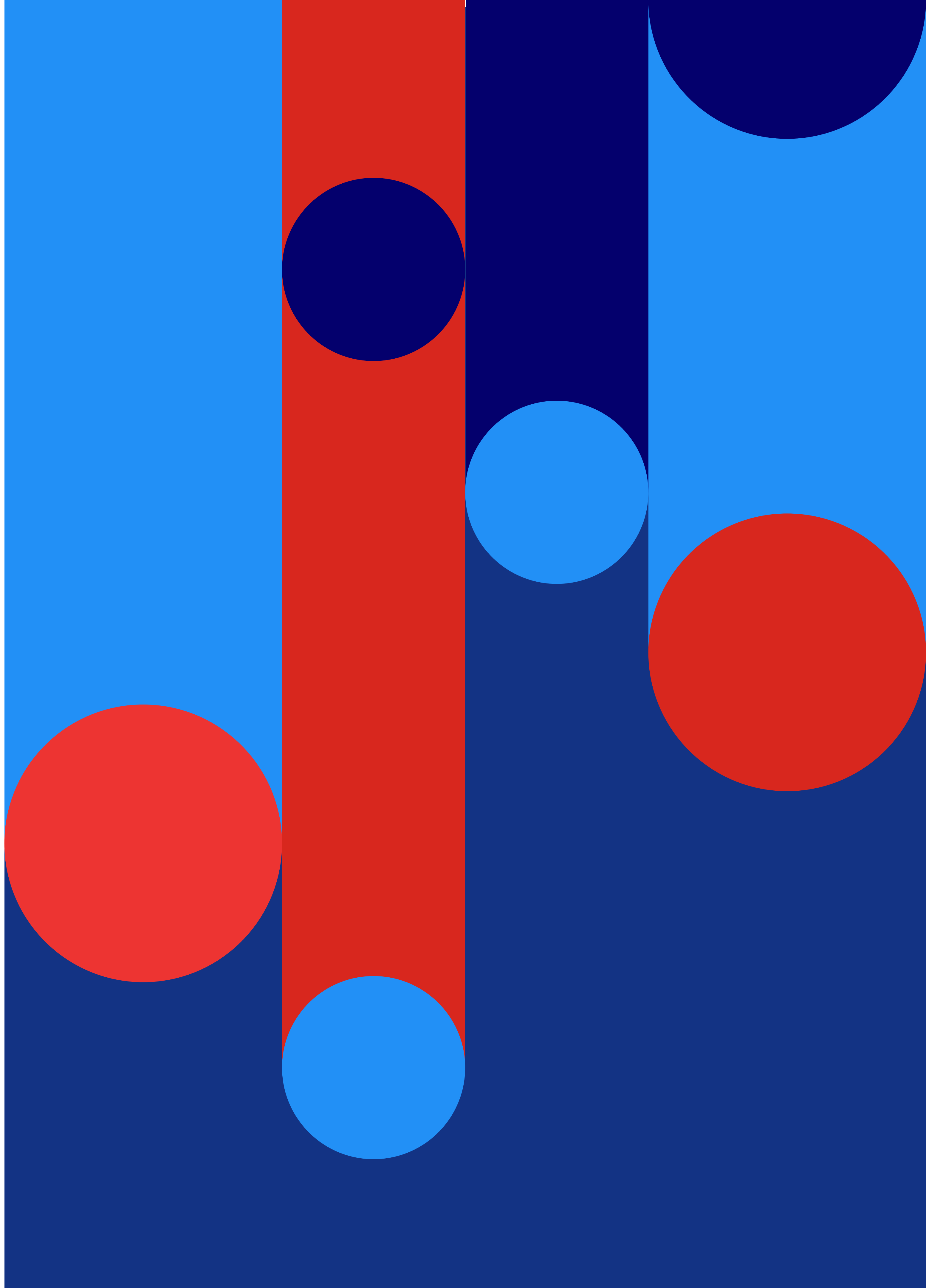
Sentiment Analysis

Sentiment Analysis

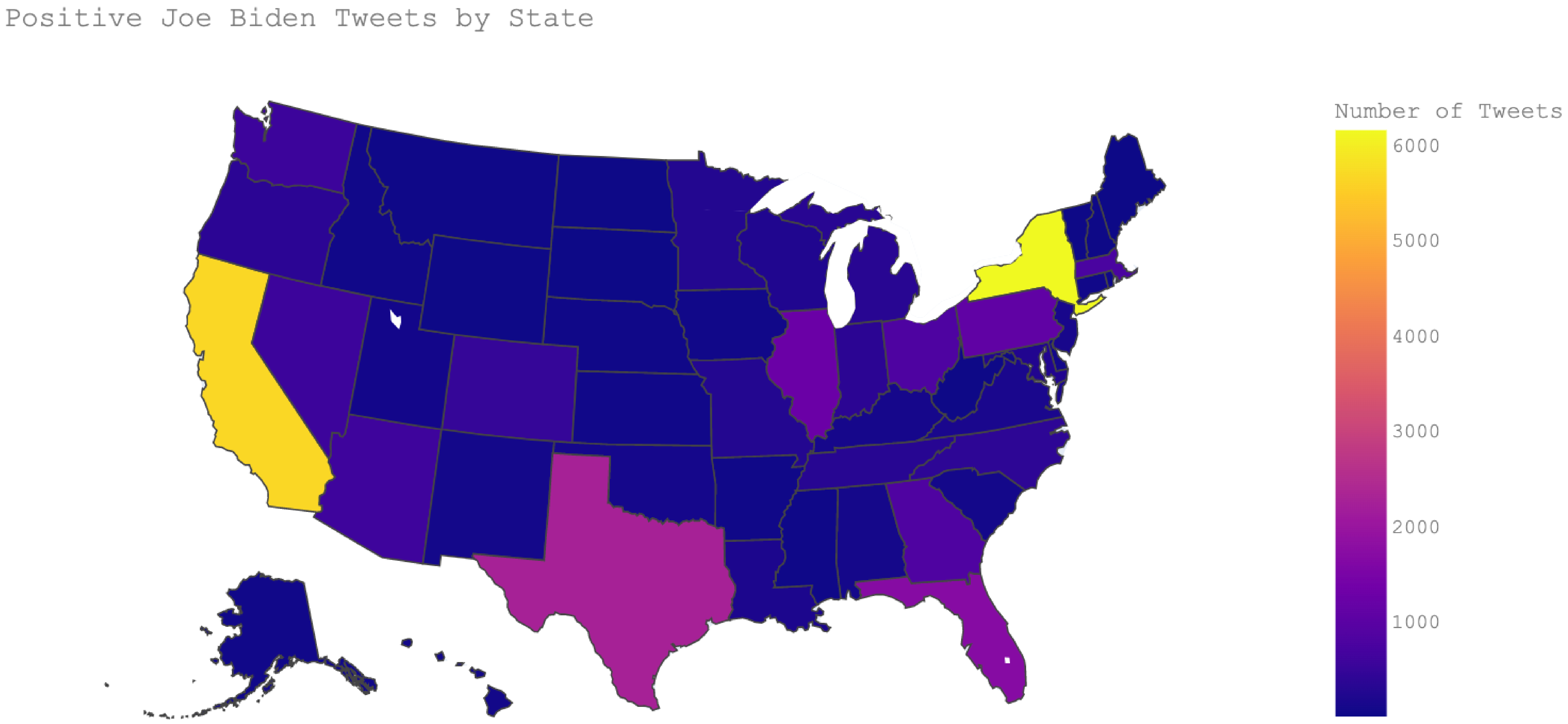
1. The SentimentIntensityAnalyzer from the Natural Language Toolkit library in Python was used to classify the sentiment of a given text as positive, negative, or neutral, and assign a polarity score to each tweet.

2. The compound score was then computed to express the overall sentiment and the intensity expressed in the tweet.

3. Finally, the results obtained by this analysis, were used to study the distribution of the users' sentiment over time and geographical areas. Sentiment analysis can help in understanding the public opinion and predicting the outcome of the election, but it can also be useful for political campaigns and strategy-making.

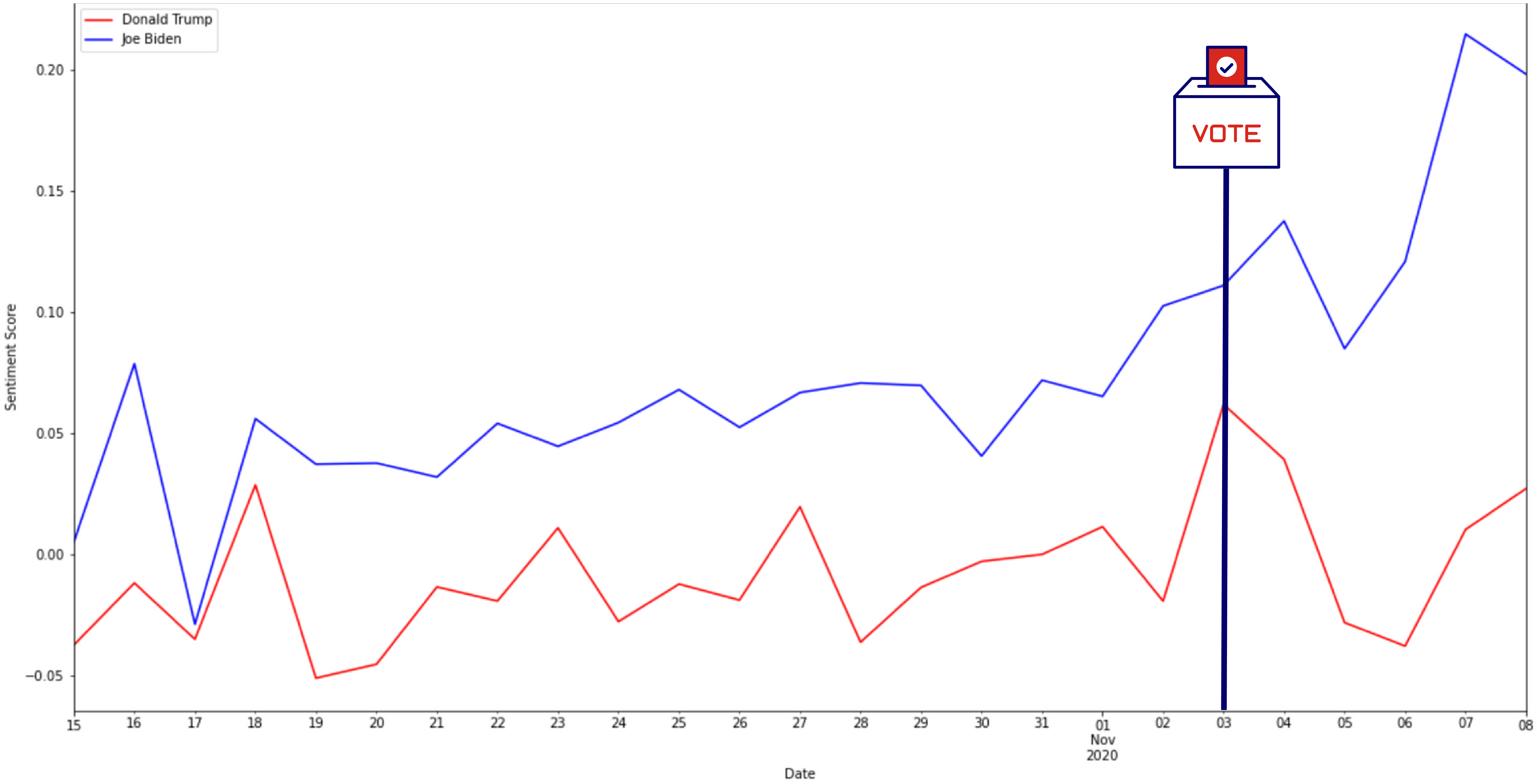


Positive Biden Tweets



In this map, we wanted to illustrate the distribution of positive Biden tweets per state. The top states are California and New York, both of which indeed are considered to be safe blue states. On election day, Biden won overwhelmingly with 63% and 61% of the votes respectively. Surprisingly Texas is in third place for positive Biden tweets, a state that has been considered a red state for a long time. Despite the fact that Trump could win in Texas in 2020 also, Biden has gotten the biggest percentage of a Democratic candidate since Jimmy Carter in 1976, which explains the higher-than-expected amount of positivity in this state.

Sentiment over time



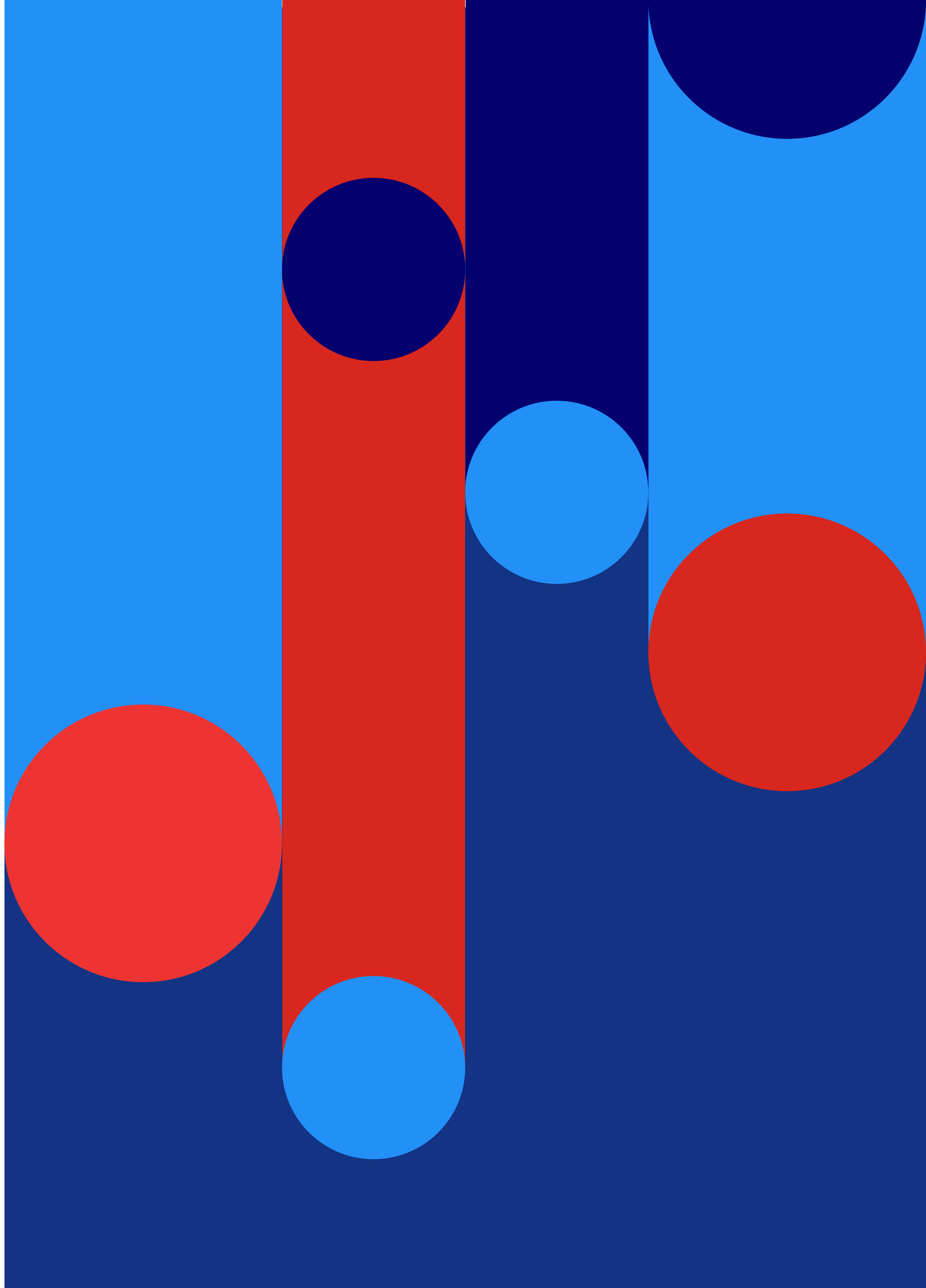
In this graph, we can clearly see how the sentiment towards both candidates has changed over time. It is interesting to notice how the sentiment towards Trump has actually always been lower (less positive) than the sentiment towards Biden. In both cases, the sentiment has changed over time, particularly for Biden. Another interesting thing that we can notice by looking at the graph, is that the sentiment towards Biden starts decreasing right after election results. This might be due to tweets of disappointed from Trump supporters.

Exploratory Data Analysis

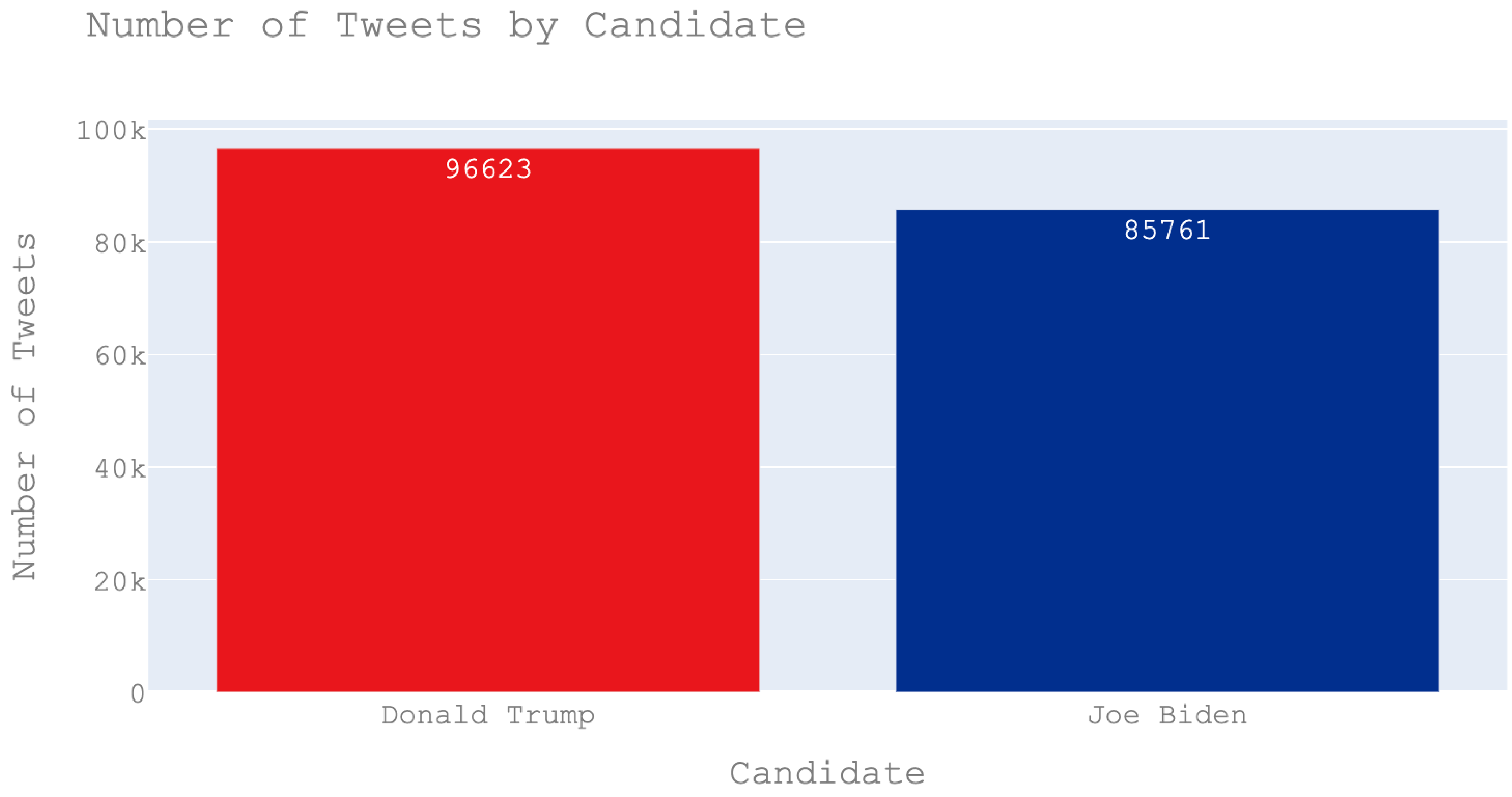
Exploratory Data Analysis

1 The distribution of the number of tweets among the two candidates was analyzed from both a geographical and temporal persepctive.

2 Word clouds were used to visualize the most common words used in Tweets related to the two candidates. Looking at these words, we can gain insights about common vocabularies and topics that appears in the users' tweets during the election days.

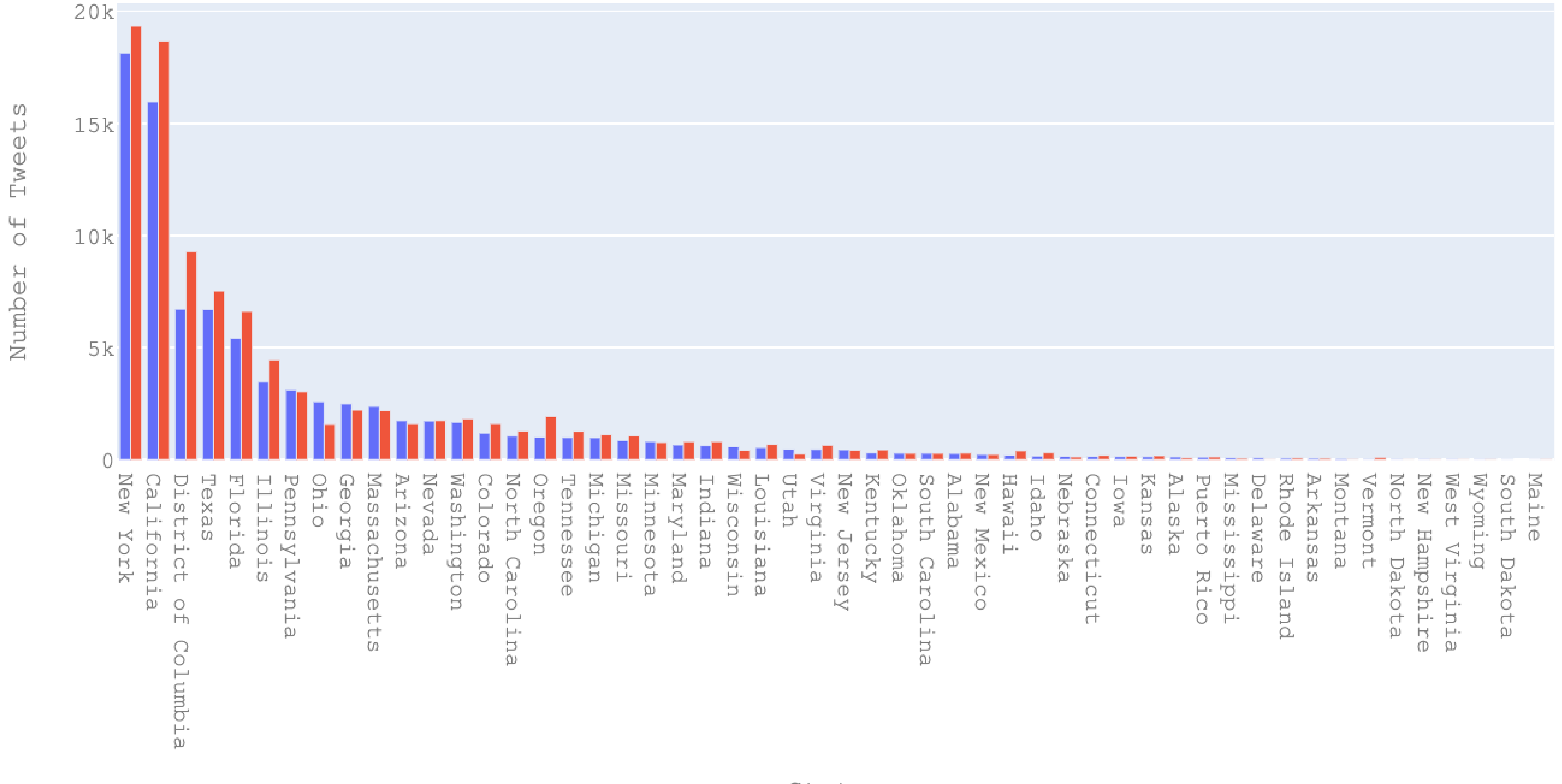


Number of Tweets by Candidate



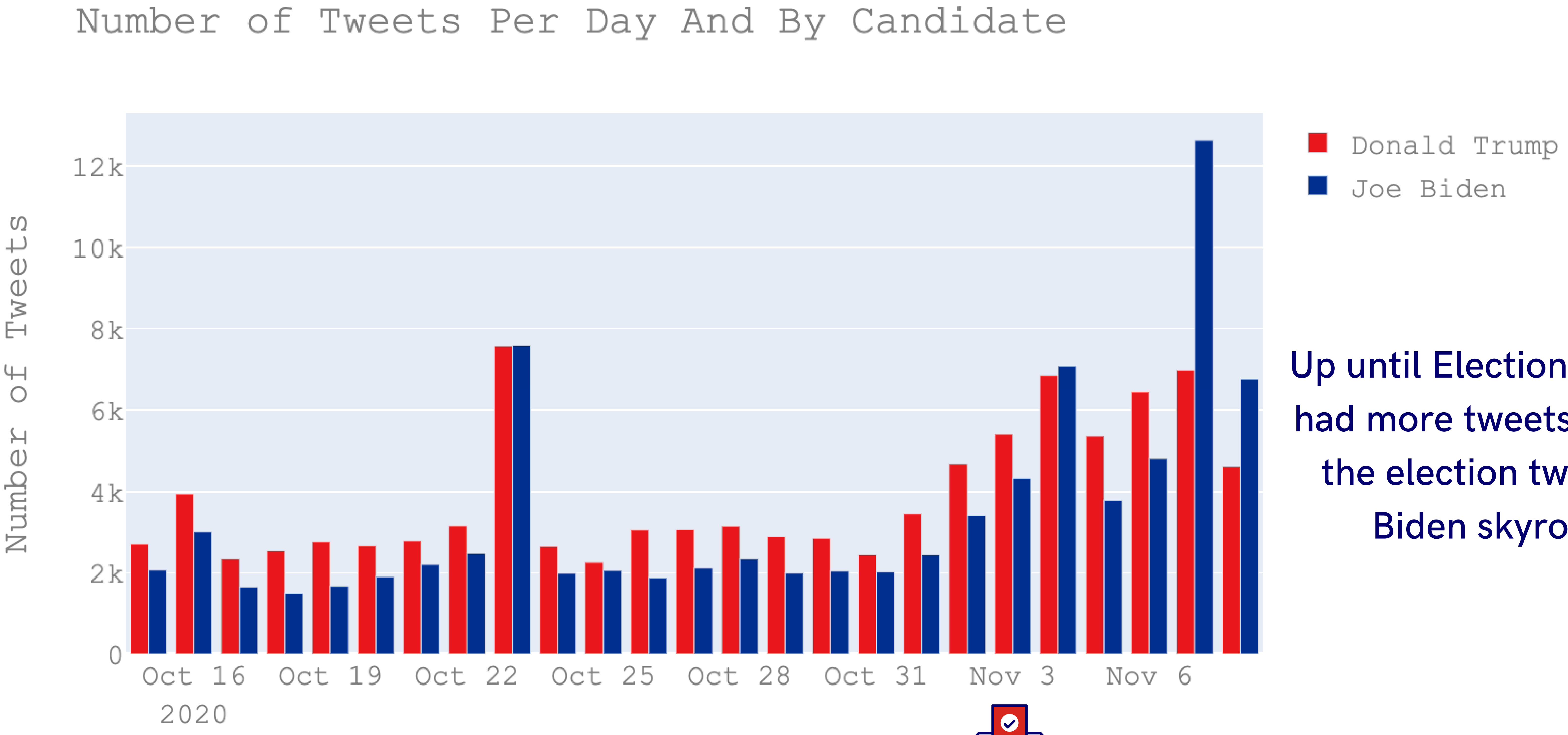
Our dataset includes approx. 95K tweets about Trump, and 85K tweets about Biden

Geographic distribution of tweets within the U.S.



Unsurprisingly, New York and California had the most tweets, the two states together take up 20% of the country's population.

Date distribution of tweets

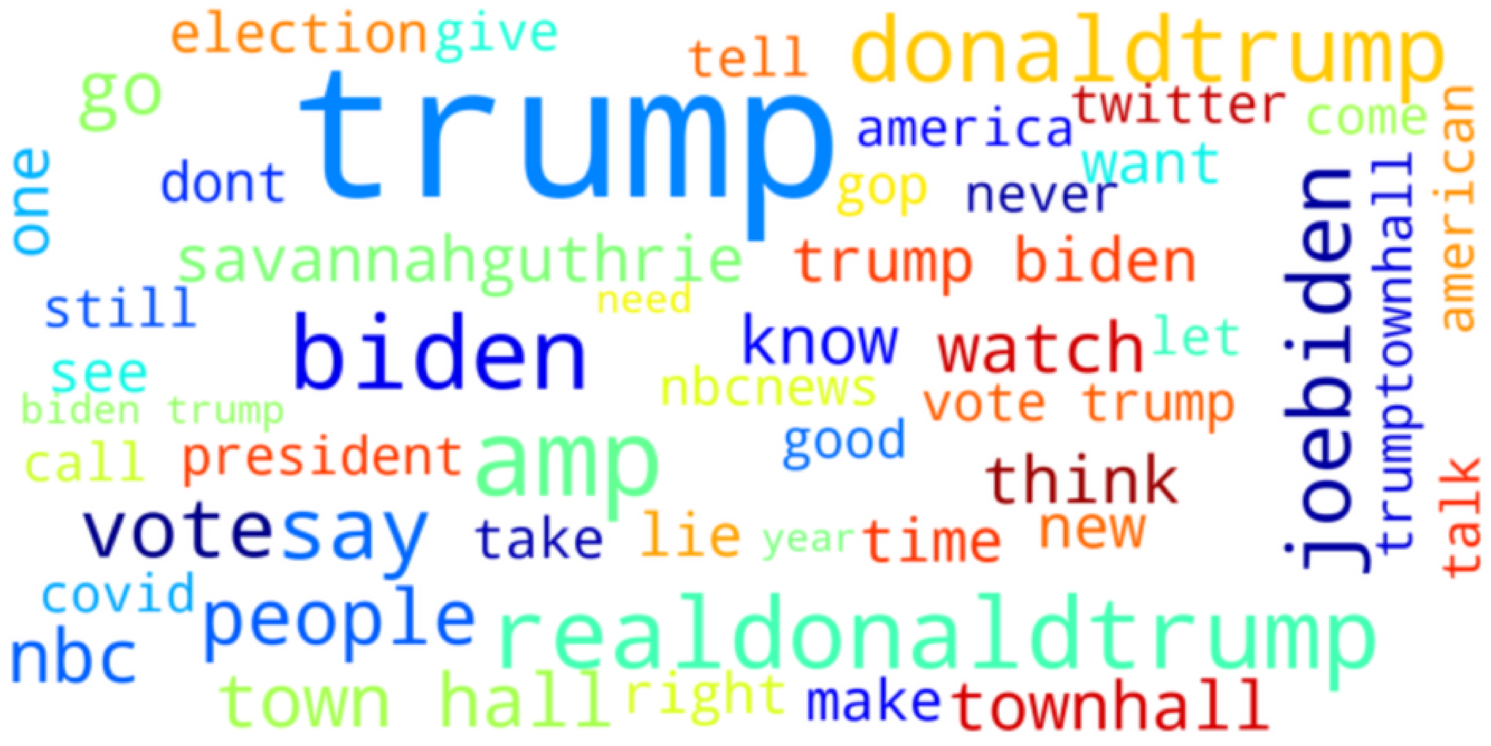


Up until Election Day, Trump had more tweets, while after the election tweets about Biden skyrocketed.

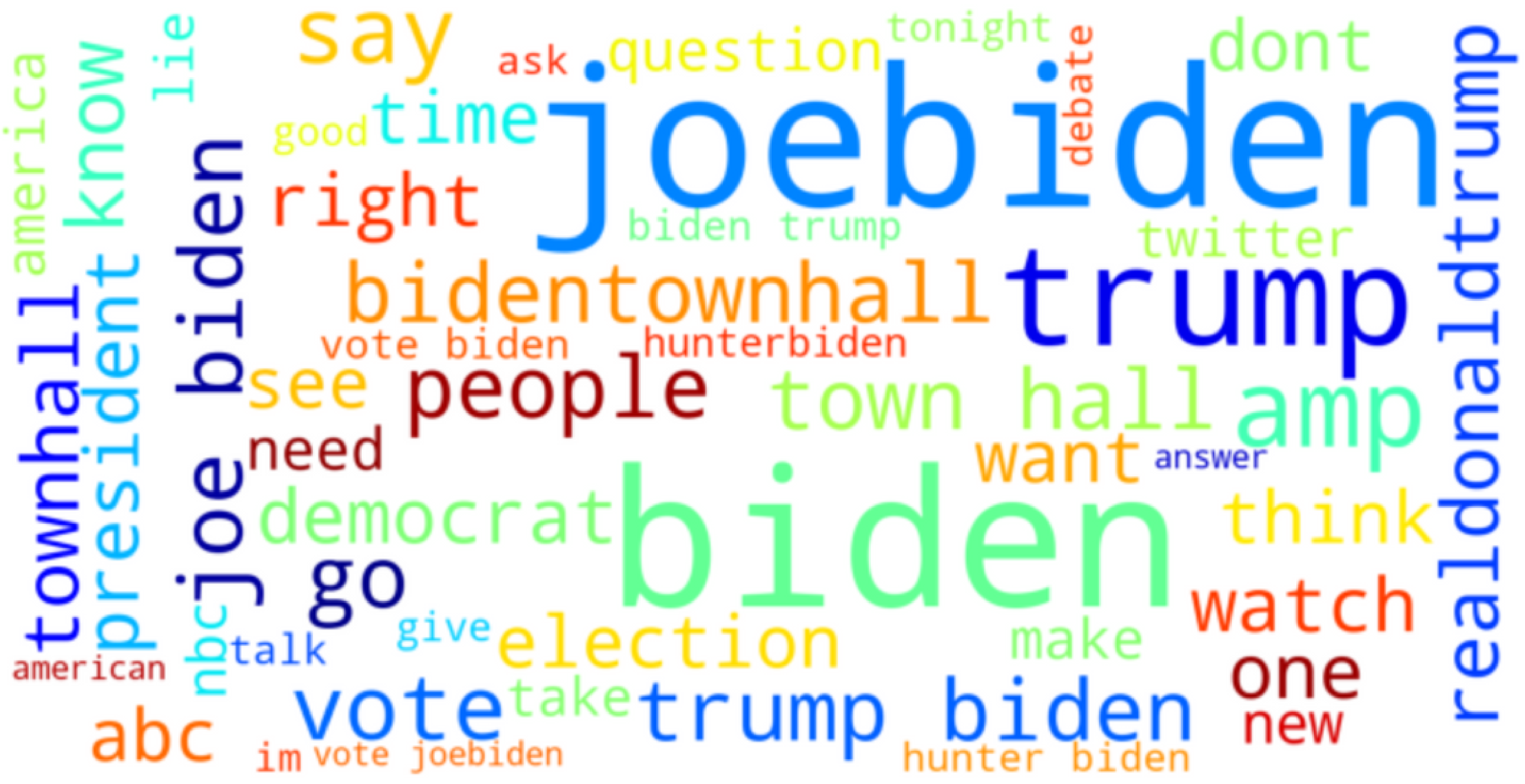
VOTE

- Here we can see two different word clouds. The one at the top includes the most frequent words appearing in tweets about Donald Trump whereas the one at the bottom represents the same but with Tweets related to Biden.
- One may notice that the sizes of the terms are different. This is because the size represents the frequency of the words: the bigger a word in the word cloud, the more often it appears in tweets.
- By looking at both clouds, we may notice that in tweets related to Joe Biden, the word "Trump" appears a lot, and vice-versa. In fact, in the tweets, people would often compare the two candidates. This is a new trend: Americans tend to compare the two candidates on a personal level, more than comparing the different ideas.

Most frequent words in tweets about Donald Trump



Most frequent words in tweets about Joe Biden



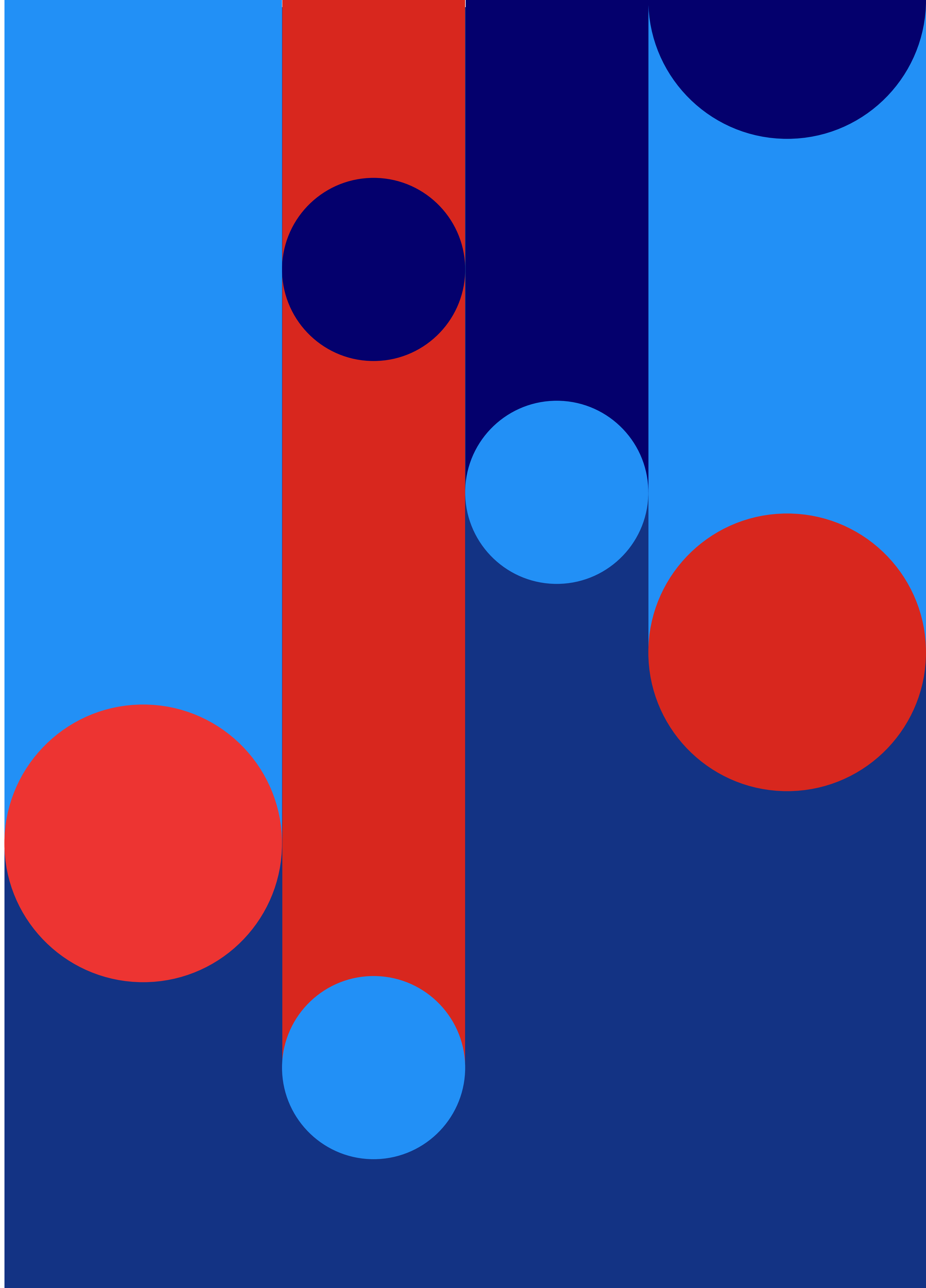
Word Embedding

Word Embedding

1 Word embedding was implemented using the Word2Vec model on election-related tweets. The cleaned tweets were first tokenized and then used to train the model.

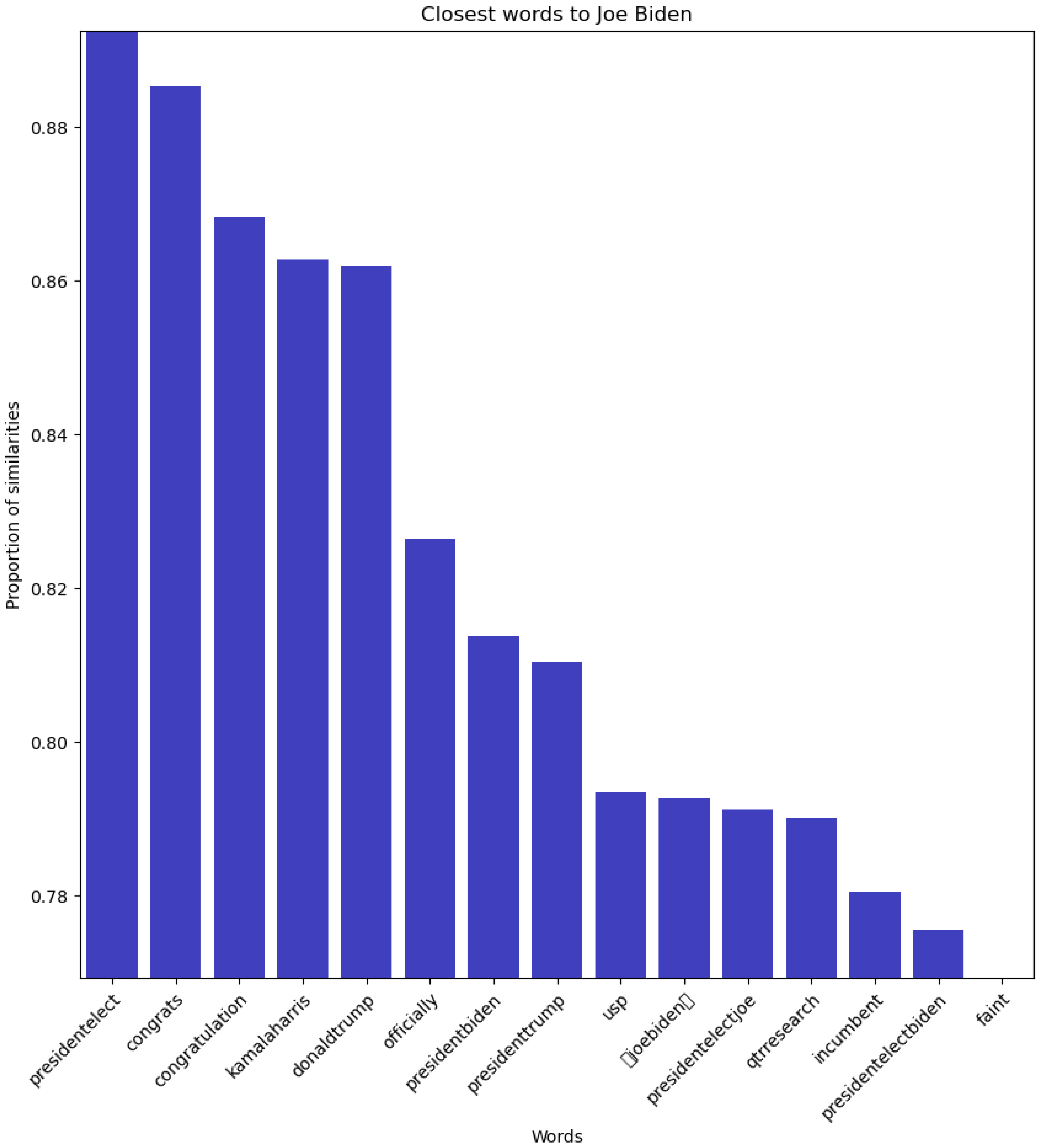
2 Then, the 15 most similar words to "joebiden" and "donaldtrump" were extracted and visualized using a bar plot. Words with similar meanings or contexts are located closer to each other.

3 This technique helped us in identifying semantic similarities, and to capture the underlying meaning and topics discussed in the tweets.



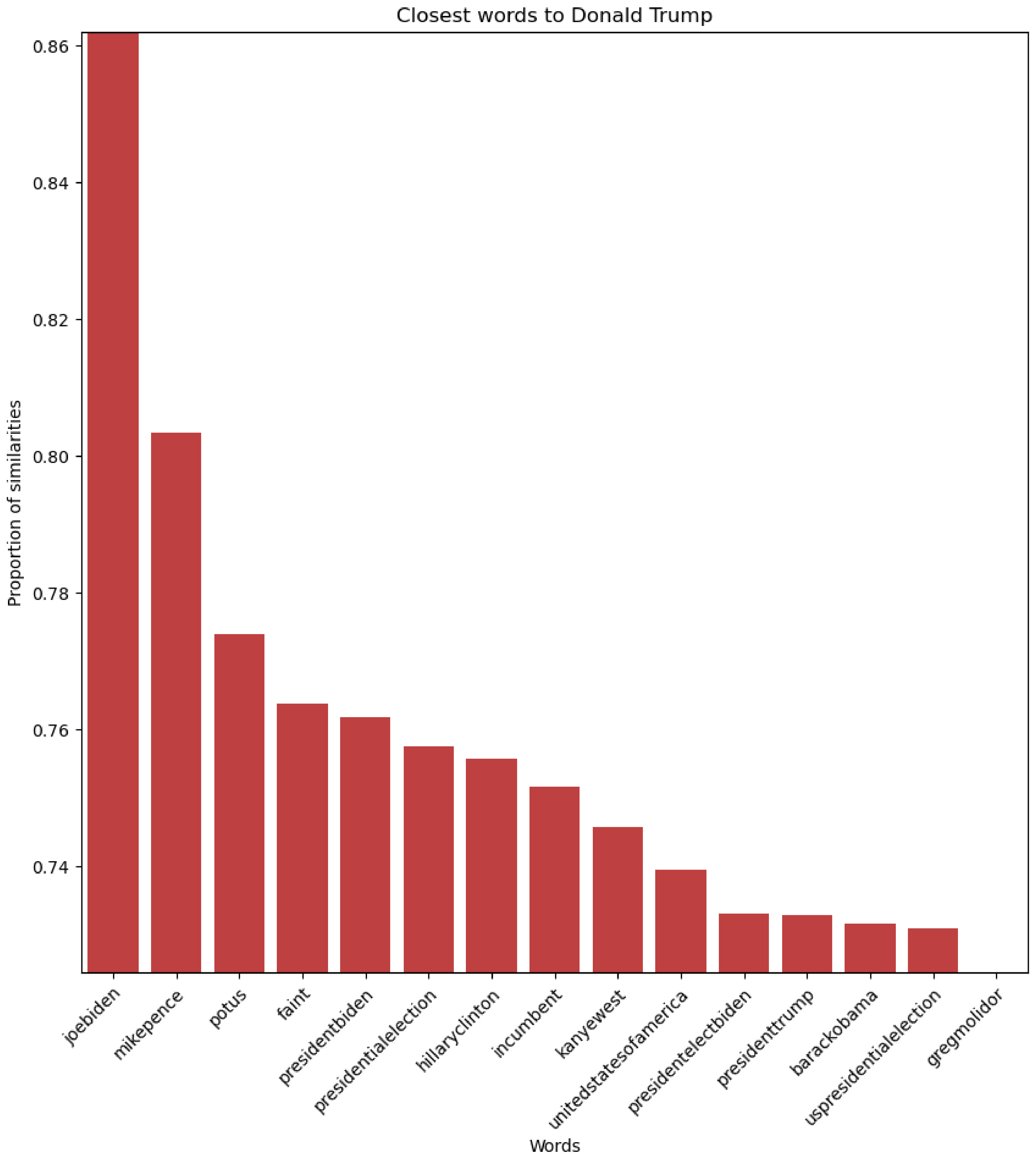
Top 15 closest words to "Joe Biden"

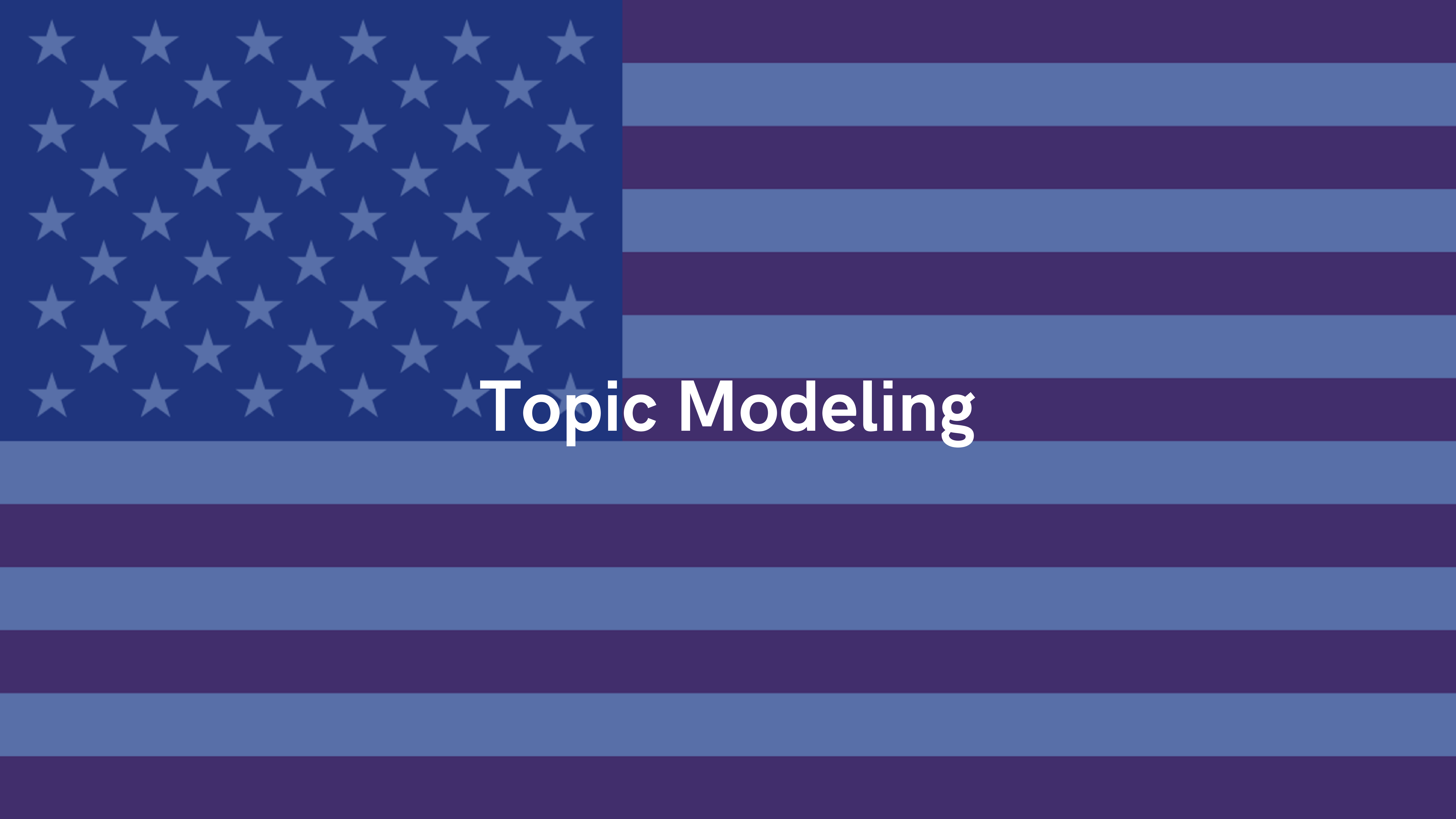
- From the list we can see that some words are related to Joe Biden either because of the name or the context, such as:
  - The words containing "President" or "POTUS", "congrats" and "officially", make sense because he won the elections and tweets were collected until the 8/11.
  - The words containing "Kamala Harris", make sense because she is the now Vice president.
  - Finally, the words related to Trump may be associated to Biden, because they were both the candidates, and people tweeted about them in the same tweet.



Top 15 closest words to "Donald Trump"

- From the list we can see that some words are related to Joe Biden either because of the name or the context, such as:
  - The words containing "Donald Trump" or "Mike Pence" make sense because he was the other candidate or his candidate for Vice President.
  - "Kanye West" may be associated to Trump because Kanye West supported Trump.
  - Finally, the words related to Biden may be associated to Trump, because they were both the candidates, and people tweeted about them in the same tweet.





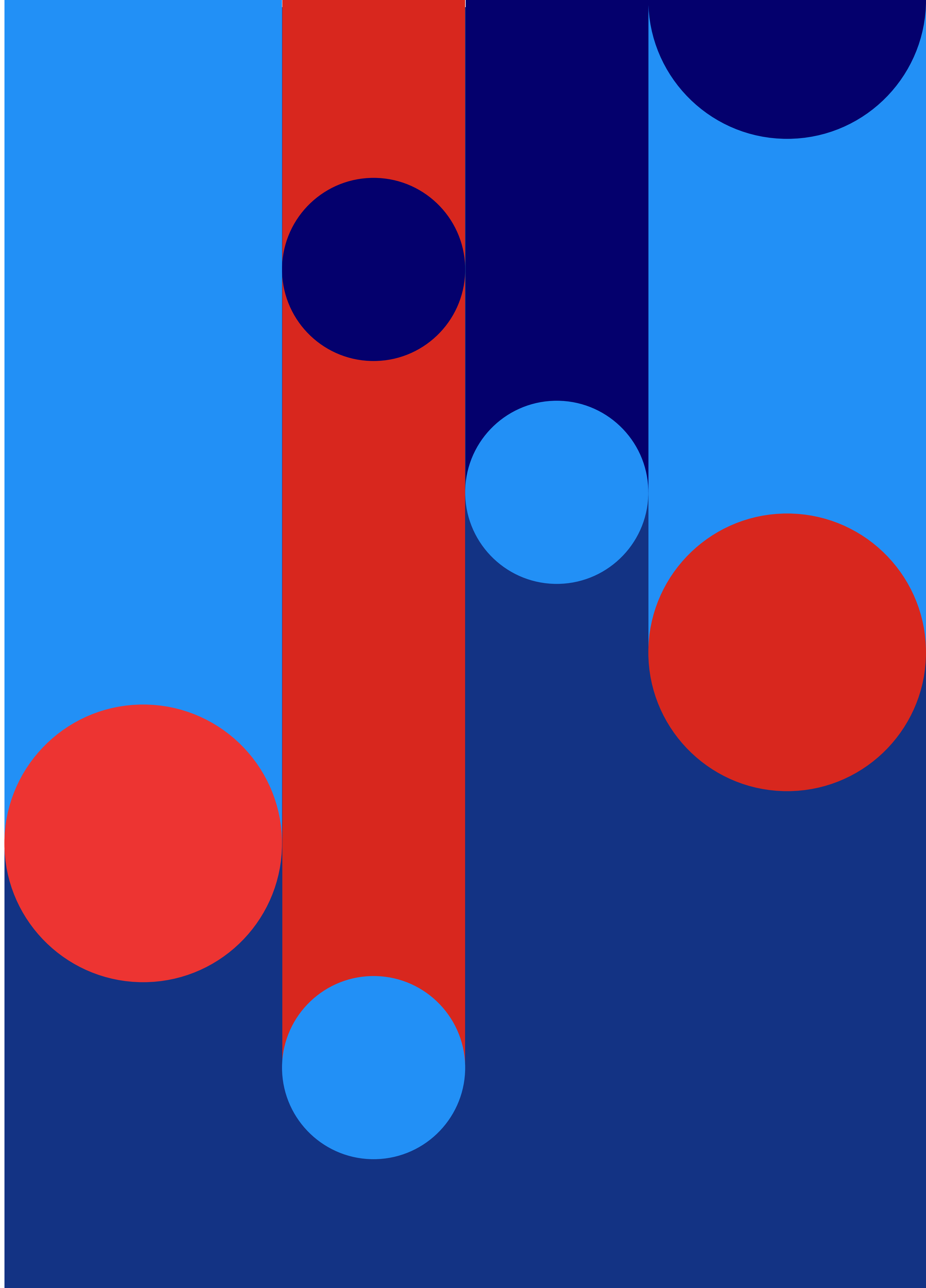
Topic Modeling

Topic Modeling

1 Topic modeling was implemented using Latent Dirichlet Allocation (LDA). The CountVectorizer is used to convert the text into a document-term matrix to fit the model.

2 The top 15 words for each topic are extracted and this procedure is repeated for both tweets and users profile's descriptions .

3 This technique is used to identify the underlying topics in the election-related tweets to provide valuable insights into the key issues and concerns of the public, but also to distinguish different groups of voters based on their charactersitcs, descriptions, and political preferences.



Topic Modeling

The main topics identified and some of the corresponding top words are reported for both tweets and users' descriptions. Regarding tweets, the two main themes of discussions seem to be anti Trump or related to the policies undertaken by the candidates to mitigate the effects of Covid. Regarding the profiles' descriptions, we can identify two opposite groups (republicans and liberals), but also a topic related to the users' jobs.

Tweets

1) Anti Trump: "demcast", "dump trump", "blue wave", "vote blue", ...

2) Covid Policies: "coronavirus", "covid", "campaign", ...

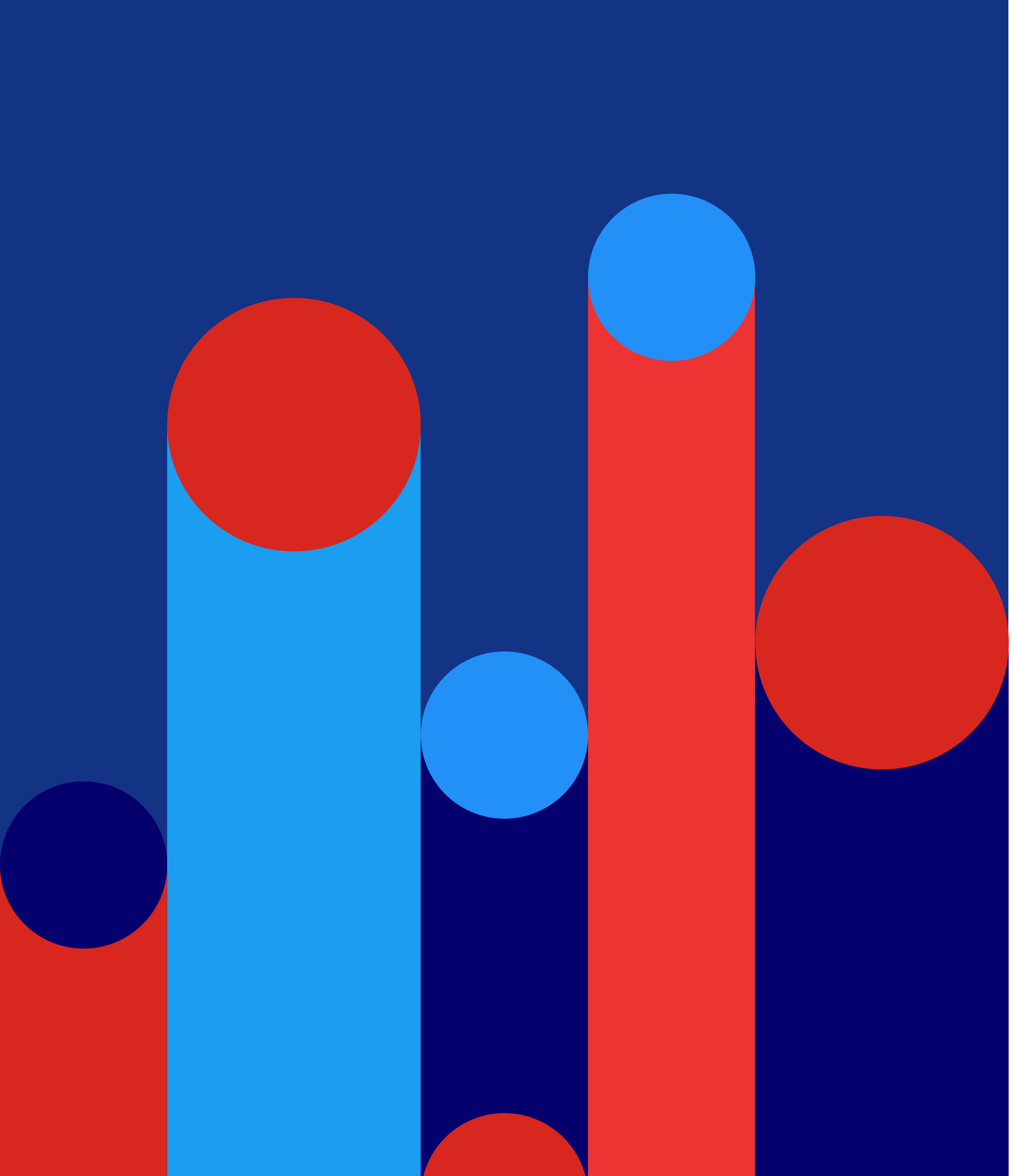
Users

1) Republican: "trump", "make america great again", "patriot", "life", ...

2) Liberal: "never trump", "black lives matter", "resist", ...

3) Jobs: "doctor","author", "music", "correspondent", ...

Classifier



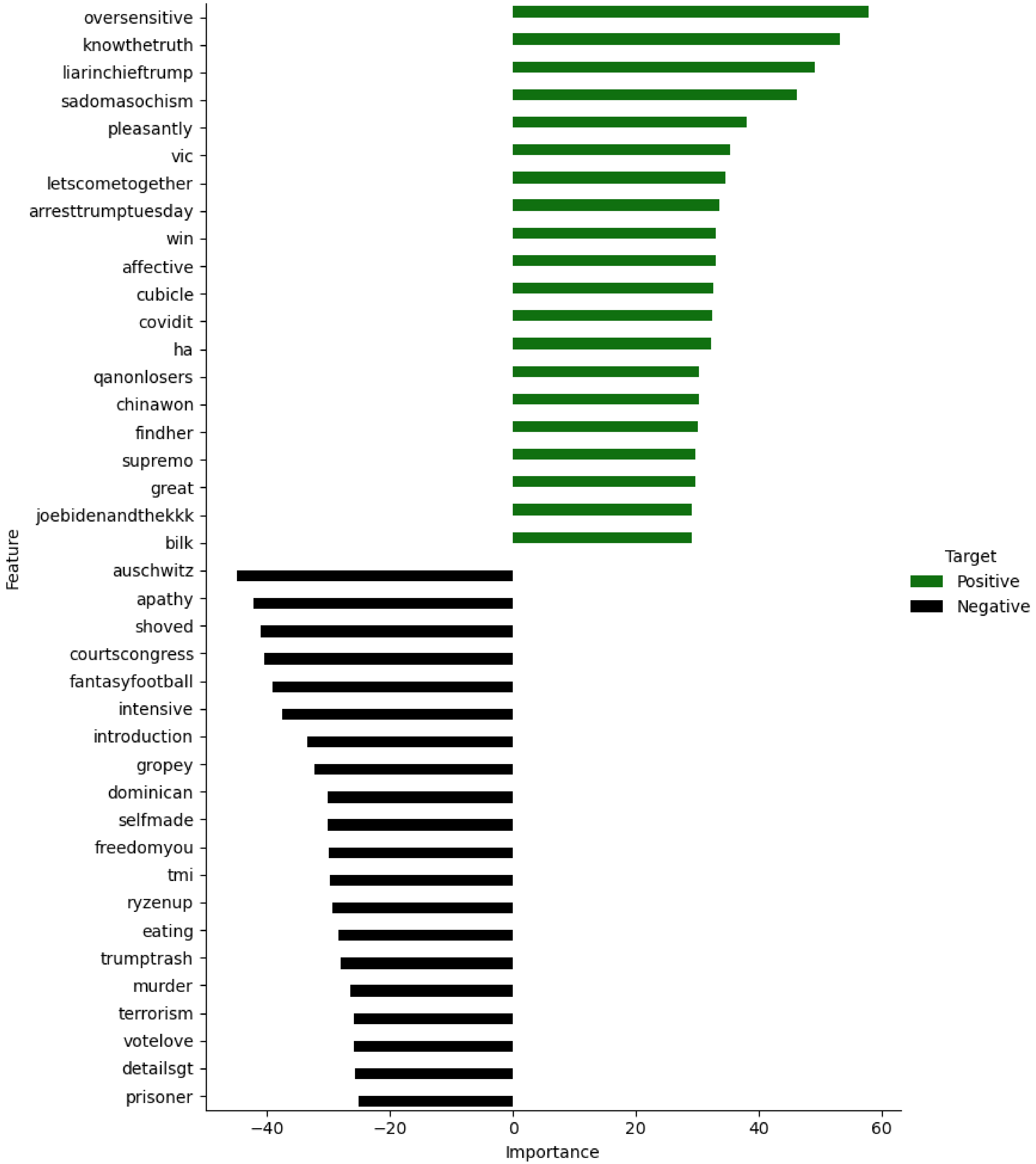
Classifier

1 Classifier used to predict the sentiment of the tweet (if positive or negative). Both a logistic regression and a decision tree were built but the regression performs better.

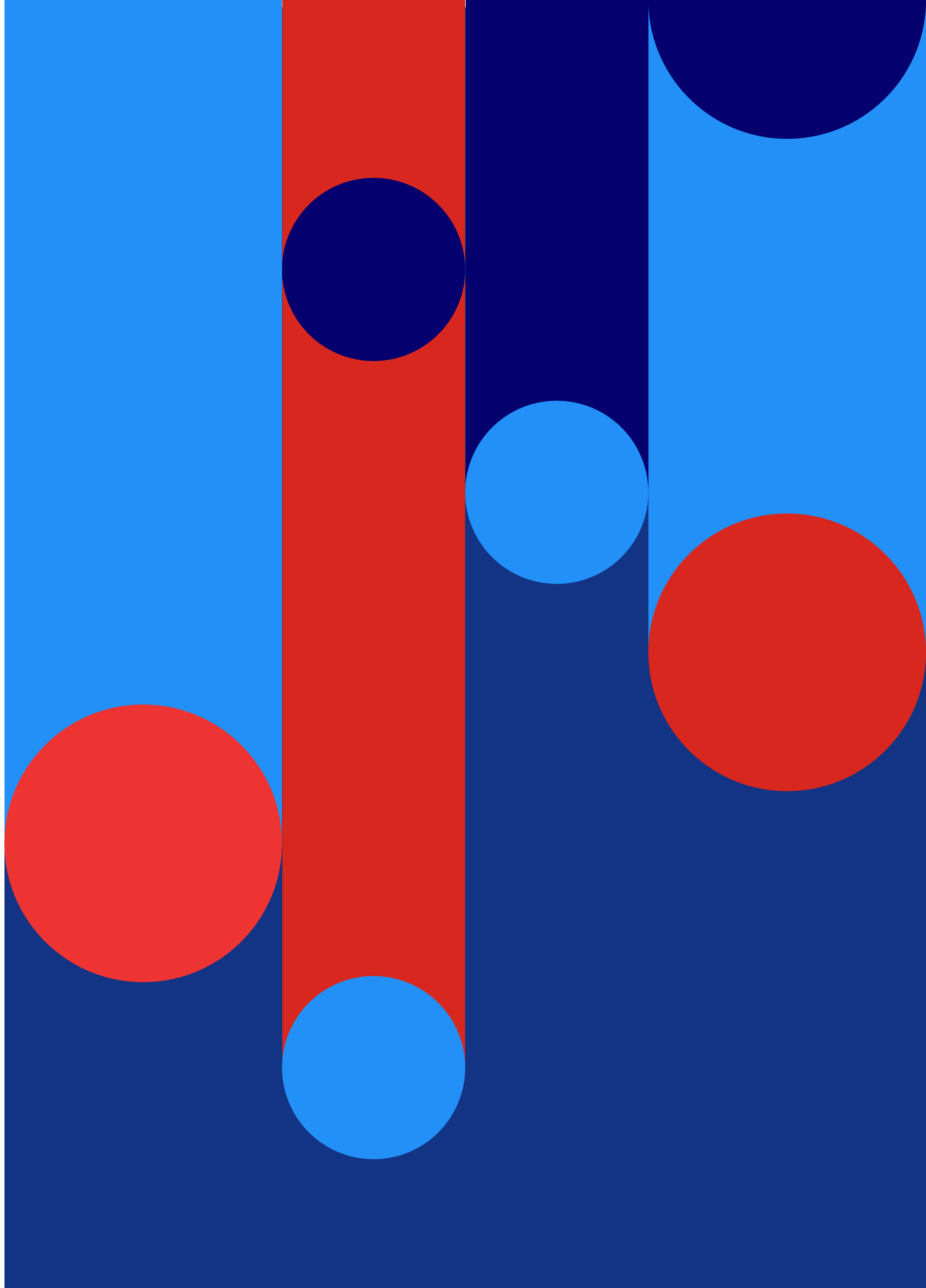
2 Model: Logistic Regression
Accuracy on the Test Set: 0.93
Precision on the Test Set: 0.92
Recall on the Test Set: 0.91
F1 Score on the Test Set: 0.92

3 The model demonstrates strong performance on the test set, indicating that it is able to accurately classify tweets as either positive or negative based on their content.

Top 20 Most Relevant Words for both Positive and Negative sentiment



It's important to note that the labels used in this project were created using sentiment analysis, as we did not have a pre-labeled dataset. However, this means that our labels may not be 100% accurate, and there may be cases where a tweet is labeled as positive when it is actually negative (or vice versa).



Take home messages

1

Users usually tweeted about both candidates. And Biden had more positive tweets overtime than Trump, in spite that there were more tweets related to Trump.

2

Before Election Day, Trump had more tweets, but after Election Day, Biden was more tweeted (although it can be seen that not all his tweets were as positive as before).

3

By using WordEmbedding and Topic Modeling, we could discovered the words and topics more discussed by users regarding both candidates.

4

With the classifier we could classify other tweets regarding the 2020 US Election either as positive or negative. Although we can have a better accuracy if we've had a pre-labeled dataset.