Sentiment analysis on 2020 US elections

Pietro Campanella - 3080098
Roberto Ceraolo - 3068890
Vittorio Costa - 3065410
Alessandro Fedel - 3069732

December 2020

Abstract

This is the research project for the course of Big Data and Databases done by four attending students enrolled in BEMACS currently at their third year of studies. At a broad level, our study is concerned with the sentiment analysis of twitter data related to the US election. As a matter of fact, this analysis could be useful to understand the general sentiment of Americans when it came to the two main candidates. More in detail, we found two datasets, where each one of them contained every tweet with the name of the runner for presidency in the hashtag form. After a process of data cleaning and preparation, we started with a preliminary analysis of the dataset and some data visualization. Then, we applied the lexicon method and a formula to label the tweets. Furthermore, we deployed three different models on the labeled dataset in order to analyse the general sentiment related to the two candidates.

Goal of the Analysis

This analysis has two major goals. Firstly, understanding whether tweets can provide us information about the tendency of US citizens to like or dislike the two candidates. This purpose is satisfied through the application of the lexicon method, which allows us to construct the ”sentiment” label that will be used in the machine learning part of the analysis. We believe that sentiment analysis on twitter can really show how people express their political opinions on social media platforms. More ambitiously, we also want to check whether the sentiment of the tweets can be informative of the final result of the elections.

Dataset used and descriptive statistics

0.1 Dataset

The U.S. election is one of the most anticipated and covered political event in the whole world. On November 3rd 2020 the elections took place and the U.S. citizens voted for their next president. These years' elections saw the democratic former vice president Joe Biden running against the current president of the United States of America, the republican Donald J. Trump.

In this analysis, we will use data from the Twitter API about tweets published in the time frame between October 15th and November 8th. In particular, we have merged the two following datasets:

- The dataset of tweets that had one or more hashtags mentioning Joe Biden in the text
- The dataset of tweets that had one or more hashtags mentioning Donald Trump in the text

The datasets can be found also on Kaggle [1].
These datasets contain the following features for every tweet:

- created_at: Date and time of tweet creation
- tweet_id: Unique ID of the tweet
- tweet: Full tweet text
- likes: Number of likes
- retweet_count: Number of retweets
- source: Utility used to post tweet
- user_id: User ID of tweet creator
- use_name: Username of tweet creator

- user_screen_name: Screen name of tweet creator
- user_description: Brief self-description by tweet creator
- user_join_date: Join date of tweet creator
- user_followers_count: Followers count on tweet creator
- user_location: Location given on tweet creator's profile
- lat: Latitude parsed from user_location
- long: Longitude parsed from user_location
- city: City parsed from user_location
- country: Country parsed from user_location
- state: State parsed from user_location
- state_code: State code parsed from user_location
- collected_at: Date and time tweet data was mined from twitter

0.2 Descriptive statistics

Coming to descriptive statistics, we will describe the statistical analysis and data visualizations that we deployed in our workflow. Before starting with descriptive analysis (which in the workflow has its dedicated metanode), we decided to create a new variable:

- popularity_group: it tells you, based on user_followers_count, 5 different groups of popularity. Actually, this node has been created through a numeric binner in Knime, and then through the "One to many" node we created 5 dummies, one for each popularity group. It will allow us to recognize patterns, if any, among popular profiles and influencers. The distribution of the levels of popularity is shown in Fig. 2 in the Annex.

Hence, considering the new variables created, we made some statistical considerations. The three main analyses we performed are:

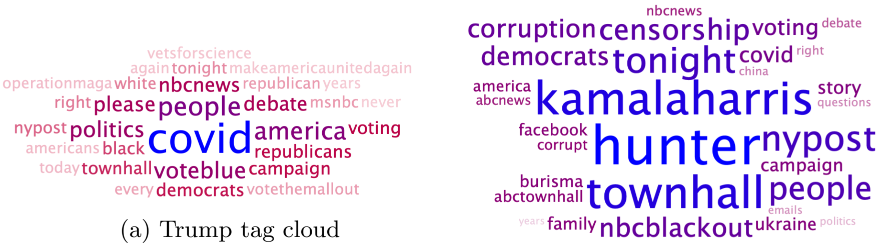
- Crosstab
- One-way Anova

The One-way Anova test allows us to compare mean values of different populations. More specifically, through an F-test, it checks whether the difference is significant or not. The crosstab analysis comprises a Chi-square test to understand the relation of two columns with categorical data. The Knime node displays also the frequency distribution of the categorical data. There are two different statistical analyses that can be done in this case: the first one can be done by working with the sole mention frequency. In other words, checking whether the difference of the frequency with which a certain candidate is mentioned among the various categories is significant or not. To compare the mention frequency among groups, we used the crosstab command on Knime. What we obtained is a Chi square value of 189.722 of a Chi Square distribution with 5 degrees of freedom. The statistic tests the null hypothesis of no association between the candidate mentioned and the popularity group to which the tweeter belongs. We reject the null hypothesis with a p-value less than 0.01., hence there is an association.

The second kind of analysis is more informative. Let's suppose to build a descriptive statistics analysis only once we labeled the dataset through the lexicon method. This way we can verify the statistical significance of the differences in the sentiment scores among the groups. For this matter, we used a One-way ANOVA for a multivariate analysis of variance. We obtained a between-groups p-value of $2.6 * 10^{-14}$ so the result is significant and there are differences of sentiments between the categories.

0.3 Tag clouds

An informative visualisation when dealing with sentiment analysis are tag clouds. They are graphical representation of the most occurring words, with coloring and size depending on the term frequency.



(a) Trump tag cloud

(b) Biden tag cloud

Figure 1: Tag clouds of the two candidates

As we could expect, there are many difference between the most occurring words in the tweets regarding the two candidates. Interestingly enough, the name of the candidate vice-president for Trump, Pence, never appears, while Kamala Harris is frequently mentioned in the tweets regarding Biden. This is probably due to the fact that she was the first woman to run for vice-presidency. Another peculiar character is the significant presence of "voteblue" in the Trump tag cloud. This means that many Biden supporters tweeted using the hashtag Trump. Similarly, from the Biden's cloud we can notice that many Trump supporters tweeted about the conspiracy theories on their political opponent, from corruption, Burisma, Ukraine. In both cases, defamatory tweets could also possibly show attempts of manipulating the elections (e.g. through the use of bots). Verifying such assertion or proving it wrong can be kept in consideration for further research

1 Data preparation and features selection

First of all, the data extraction was done by deploying two python nodes using pandas because with the standard CSV reader implented in Knime we had trouble reading the formatting of our dataset. Moreover, we created a dummy variable called ”candidate” that accepted ”Trump” and ”Biden” as possible values and concatenated the two datasets into one.

As of now, we have almost 2 million tweets and too many columns, so the data cleaning process can start.

- We extracted the two datasets through pandas using a python node
- We created a variable candidate for the two datasets with ”Trump” or ”Biden” as possible values
- We filtered out just US based tweets, since our analysis will be US-focused (also to be able to verify the predicting power)
- We used the Tiki language detector node in order to exclude all the tweets written in languages different from English
- For the sake of clarity, and since we had a very large dataset, we removed the tweets having both hashtags mentioning Trump and Biden
- We've done some text reprocessing, removed punctuation, hashtags, links, twitter stems, we applied stemming to words (to have just roots of a word and be able to ignore slight modifications),
- We eliminated low frequency words, since they are not informative of the general sentiment but are mostly ”outliers”
- We transformed date and times from string to DateTime format

2 Data Modeling

2.1 Lexicon tagging

Having a much cleaner dataset, we started working on the proper sentiment analysis. Dealing with an unlabeled dataset the possibilities are quite constrained. A quite intuitive and commonly used method for basic Natural Language Processing is the Lexicon method[1]. Basically, given lists of positive and negative words [2], it looks for those words in the tweets and apply sentiment tags. On

[1] We are aware that there are many more advanced and performing procedures for NLP, but they fall outside the scope of this paper

top of that, we were able to construct the sentiment label, applying the following formula to the tagged tweets:

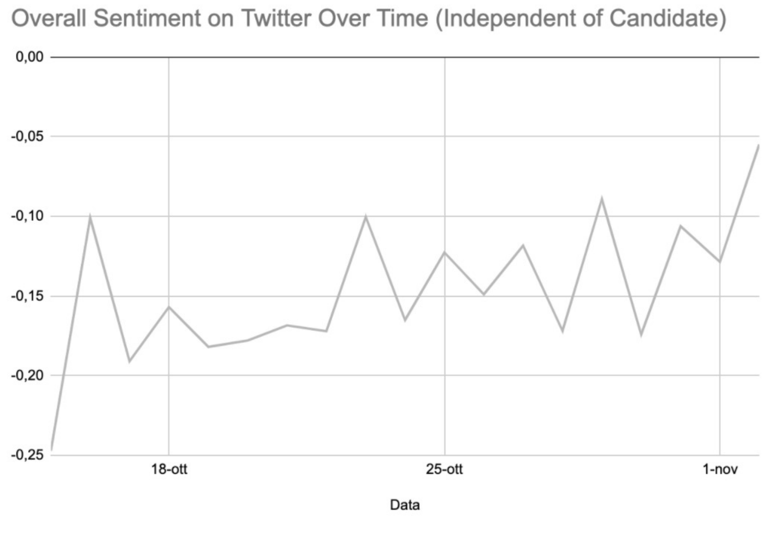
$$Sentiment\ score = \frac{Number\ of\ positive\ words\ -\ Number\ of\ negative\ words}{Total\ number\ of\ words}$$

The rationale we used is this one: when a Twitter user mentions the hashtag Trump or Biden, it could either be a supporter of the mentioned candidate (hence he/she will use positive words) or an opponent (so he/she will use negative words). Depending on the values of the sentiment score, we assigned labels "Very negative" for values less than -0.5, "Negative" for values ranging from -0.5 and 0, "Neutral" for 0-scored tweets, "Positive" for values from 0 to 0.5 and "Very Positive" for values greater than 0.5.

2.2 Visualisation and Analysis

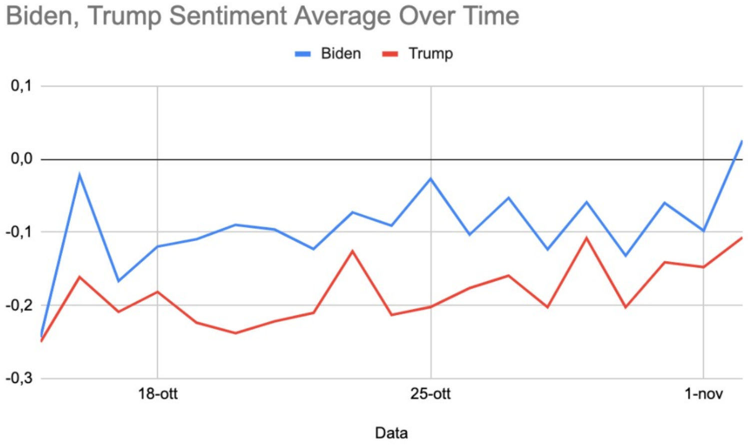
Having calculated a Sentiment Score through the lexicon Methodology we can analyze the overall sentiment for both candidates on twitter and its change over time.

We can start by looking at the Average Sentiment Score of the tweets independent of the candidate to see whether the overall environment of the social network is positive or negative, and whether it's constant or changes over time.



As we can see in the graph above, the overall sentiment is moderately negative, regardless of the candidate taken into consideration. This might be due to many reasons and may have an impact on the result of our predictions, as we will see in the managerial implications. We can also notice a slight improvement in the sentiment of the social network during the last few days leading to election.

Now that we know the platform's overall sentiment, our interest lies of course in investigating separately the sentiment regarding the 2 candidates. We can see here how the average sentiment score of the tweets regarding each candidate changes over time.



As we can see in the Graph, the tweets regarding Biden are less negative than those regarding Trump for the whole period taken into consideration. It's also interesting to notice that there are series of days in which both candidate's sentiment scores follow really similar patterns especially between the 26th of October and the 2nd of November. You can also notice that Biden's sentiment becomes positive on the last day before the election, which is the only day in which either one of the candidates has a positive overall sentiment score (even if only really slightly).

We can also look at the Average Sentiment Score of a specific candidate grouped by state.

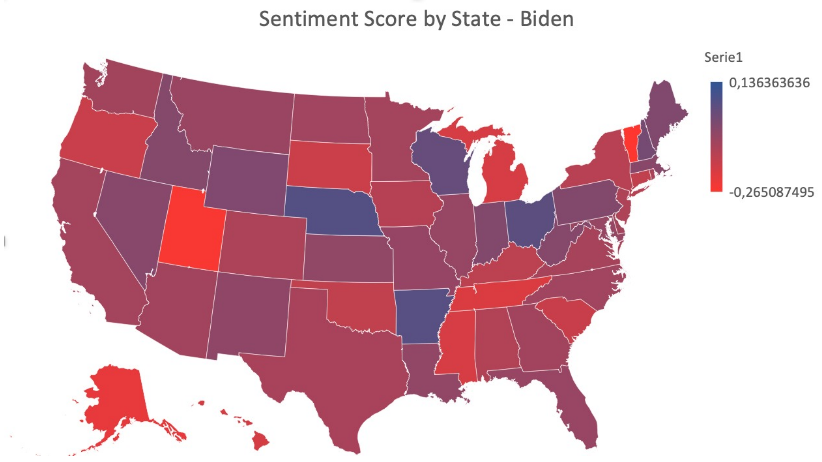


Figure 2: Here is the map of the United states with Red states being the most critic of Biden and Blue being the most supportive of his campaign.

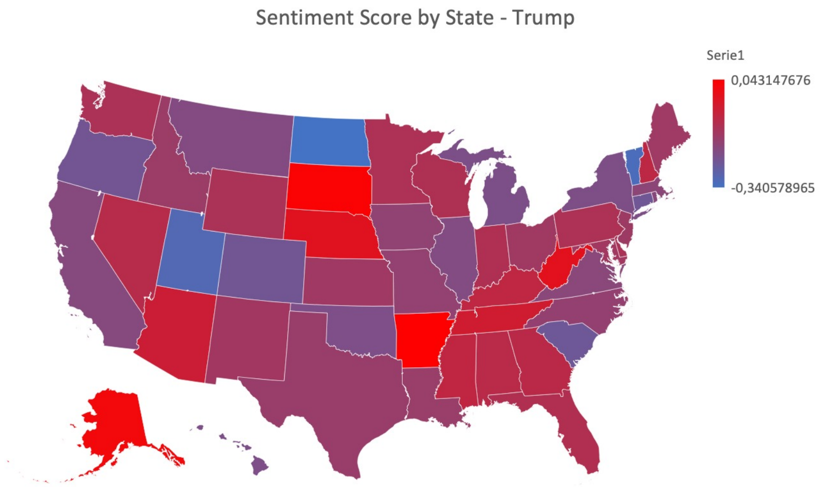


Figure 3: This second map is instead about Trump and the colors have been inverted (Blue critics, red supporters).

2.3 Predicting the elections

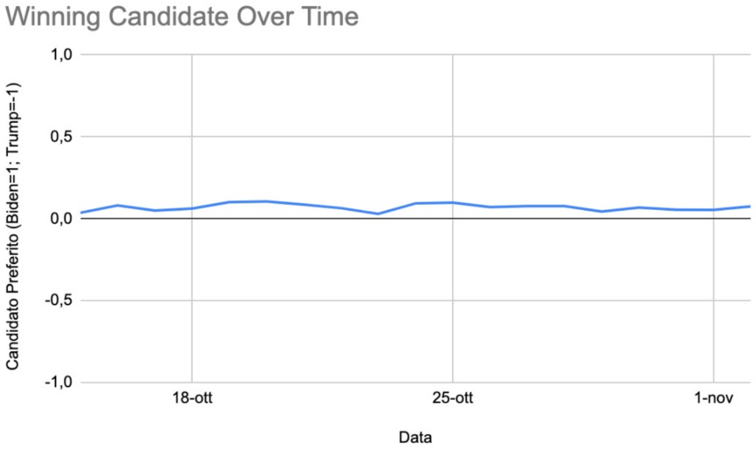
To predict the result of the election we have to define a theoretical assumption about the relationship of the sentiment score of a tweet and the preferred candidate of the user. Applied on the whole dataset this assumption can help us provide a prediction for who the preferred candidate is for the whole nation and for each state.

To define candidate preference we assume that a positive sentiment represents a propensity to vote for that candidate while a negative sentiment represents the opposite. Therefore we build a new variable calculated by multiplying a dummy variable and the sentiment score of the tweet.

Candidate multiplier = 1 if Candidate = "Biden"; -1 if Candidate = "Trump"

Candidate preference = Candidate multiplier * Sentiment score

This new variable will have values between -1 and 1, with -1 representing absolute preference for Trump and 1 representing absolute preference for Biden. Therefore, the average value of the candidate preference variable in the US will be our predictor of the winning candidate. In particular, this value is 0,066, which represents a slight preference for Biden. We can look at the change of the variable over time in the following graph.

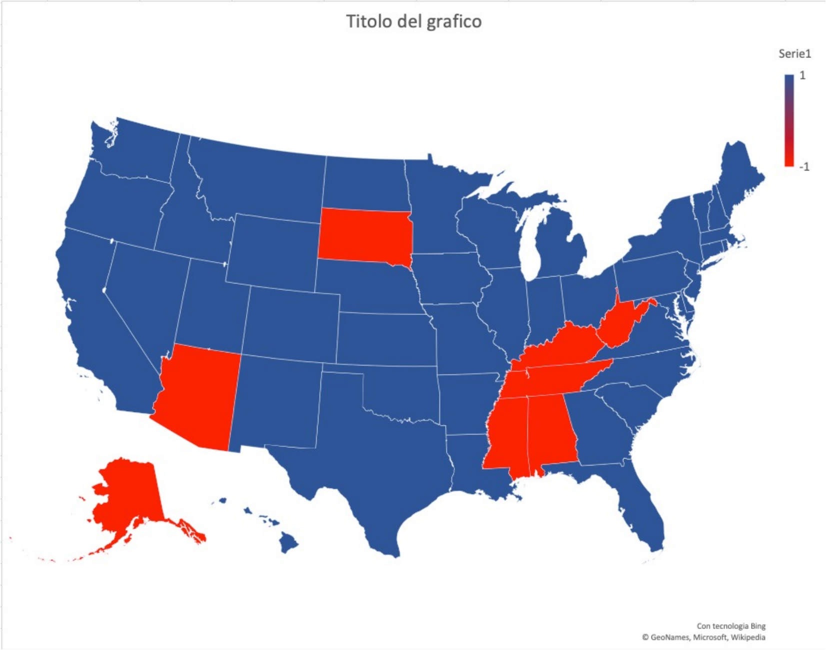


As you can see our method predicts a Biden victory for the whole duration of the period being analysed. If we compare our result with the final result of the elections we can see that the predicted winning candidate is correct. We can now look at candidate preference throughout the states and compare the average candidate preference variable with the actual election results for every state.

In the graph below you can see the states in which we predicted a Biden victory in blue and the ones in which we predicted a Trump victory in red. Even at first sight you can see that the final prediction is heavily overestimating the amount of states with a democratic victory.

For context, you can see in the annex the same map with the actual results of the election.

In particular, we correctly predicted the outcome of 32 states out of 51, missing 19. Therefore we accurately predicted around 63% of the states. Furthermore, we predicted a democratic win in 43 cases, out of which 25 were correct (58% correctness) and we predicted a republican win in 8 cases, out of which 7 were correct (88% correctness).



2.4 Supervised Learning Models

After all of this hard work to get the sentiment variable through the lexicon method, we finally got use this variable as the label of our classification. We decided that for simplicity of such a difficult task like predicting sentiment on tweets, we could have simplified our lives by creating a label that would be positive if the sentiment score was bigger than 0 and negative if the score was lower than 0. For what concerns neutral tweets we filtered them out, not considering them informative for our classification model.

As a striating point we decided to use the single words (the most frequent ones) contained in the tweets as dummies for each tweet that had least one of the "frequent" words selected. The we used a 70-30 portioning with both equal sized sampling and SMOTE deploying different types of classifiers, such as H20 Gradient Boosting and Random Forest. Right away we understood that the models were working really bad in term of accuracy so we decided to shift our analysis to a technique that we saw in the Wine analytics workflow. This procedure is called Ngrams and with that we were able to approach the classification problem from a new perspective; now we can analyze and create dummies for more than one word. So we decided to try to use the combination of two words and accept only 200 combinations, creating then 200 dummies for the classification.

Therefore, we tried with both equal sized sampling and SMOTE the three models mentioned above on the Ngrams version. The accuracy weren't that great, but from the 60% of the one word version analysis we managed to get an accuracy of nearly 70% with the equal sized sampling-Random Forest-Ngrams version of the model.

For what it concerns the models' results we want to assert that they are related to a reduced sample of datapoint to speed up the models computation. This is a metanode and it can be removed easily in future.

3 Managerial implications

By looking at the results of the sentiment analysis of the tweets we can reach a number of learnings that we consider important.

First, we have seen the overall sentiment of the platform regardless of the candidates and we have noticed that it is negative. This means that it is more common to be critic than supportive when talking about elections on twitter. This is interesting for 2 reasons: it gives a lot of perspective about the type of communication and content that can be found on twitter about this topic; the overall negativity of the content might have an impact on the precision of our predictions.

Second, the fact that Biden's sentiment score was less negative than Trump's might be rooted in political reasons. As a matter of fact, Biden has been depicted as a candidate which would be difficult both to hate and to love. This characteristic of the candidate might have mitigated the overall sentiment of the platform, while Trump's contrarian views and strong character might have an augmenting effect on the negativity of the social network.

Third, we have to acknowledge the fact that the lexicon approach to sentiment analysis is a relatively poor model that is not able to take into account a lot of the structures and subtleties of the human language. This of course had an impact on our analysis and further studies should be made on changes in the performance of the prediction using more advanced approaches.

Finally, there are many factors intrinsic to the platform taken into consideration (aside from the overall sentiment already stated above) that might have impacted our prediction, in particular: The population of twitter users might not be representative of the population of the states in terms of demographics and political preferences (It might be that democrats are more present on the social network than republicans); The opinions expressed on the platform might not be representative of the true beliefs of the users; A negative sentiment towards one candidate does not necessarily represent positive views on the other candidates.

References

[1] Manch Hui. *US Elections 2020 Tweets.* URL: `https://www.kaggle.com/manchunhui/us-election-2020-tweets`.

[2] Minqing Hu and Bing Liu. “Mining and Summarizing Customer Reviews”. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2004).

Annex

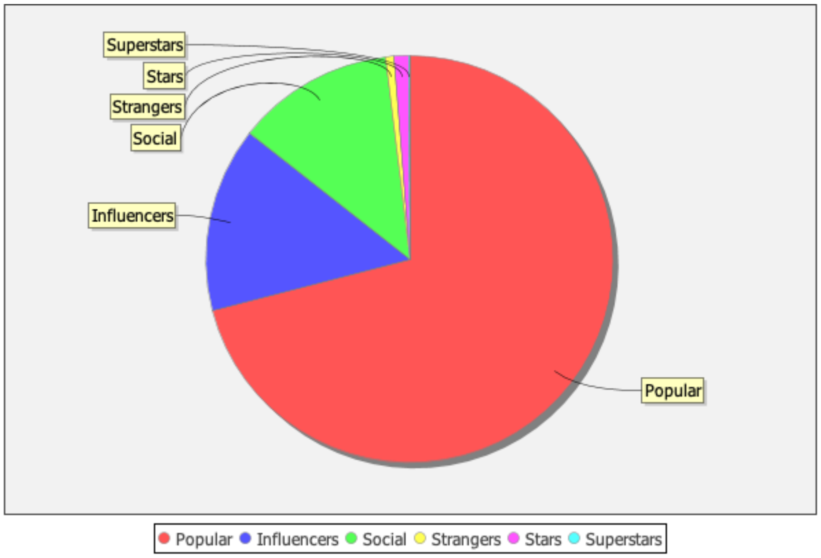


Figure 4: Pie chart of popularity levels distribution

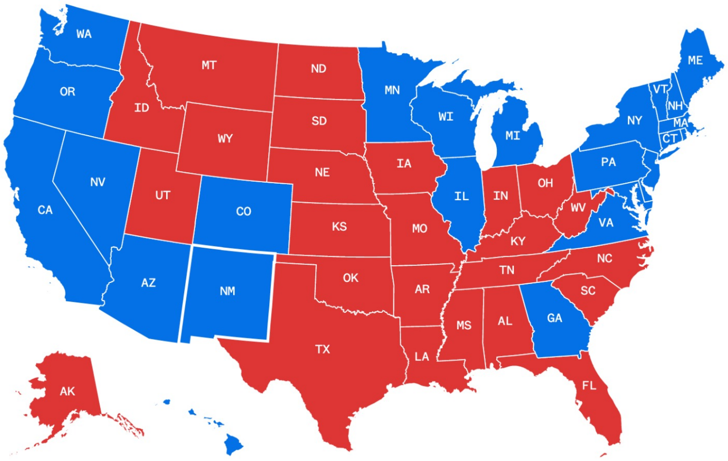


Figure 5: Actual elections outcome