How MAGA got trumped

A Location-based Sentiment Analysis of the 2020 US Presidential Elections

Table of contents

1

Introduction

What is our project about?

The Problem

How well Twitter reflects reality

What?

Visualize how people on Twitter think about the candidates.

Conclusions we can draw from sentiment analysis.

Why?

Big impact on society.

Better understanding of events and their impacts.

Assumptions:

Tweets about election.

Sentiment analysis tools correctly detect sentiment.

The Data

US Election 2020 Tweets

- Number of Tweets: 1.72M
- Coverage Period: Oct 15th to Nov 8th 2020
- Presidential elections: Nov 3rd 2020

Filtering:

- Only Tweets from US states
- Only Tweets in English
- Trump: **133K**
- Biden: **101K**

Sentiment Analysis

TextBlob: Library for processing textual data. Used to analyze opinions in social media, reviews, and customer feedback.

Sentiment: Subjectivity and Polarity.

Subjectivity: Quantifies amount of personal opinion in text.

Polarity: Emotional tone behind text.

Polarity range:

- -1: negative
- 0: neutral
- 1: positive

Null Hypotheses

1. A regression model trained on Twitter data cannot predict outcomes more accurately than random chance.

2. The clustering of the red and blue states are no better than random.

3. The clustering of the red and blue states are not distinct.

4. A candidate's sentiment time series does not Granger-cause another candidate's sentiment time series.

2

Implementation

What tools did we use?

Text Preprocessing Pipeline

Step 1

Remove usernames, and URLs

Step 2

Detect language and remove non-English tweets

Step 3

Convert to lowercase and remove HTML tags

Step 4

Demojize text

Step 5

Remove punctuation and numbers

Step 6

Remove stopwords and lemmatize words

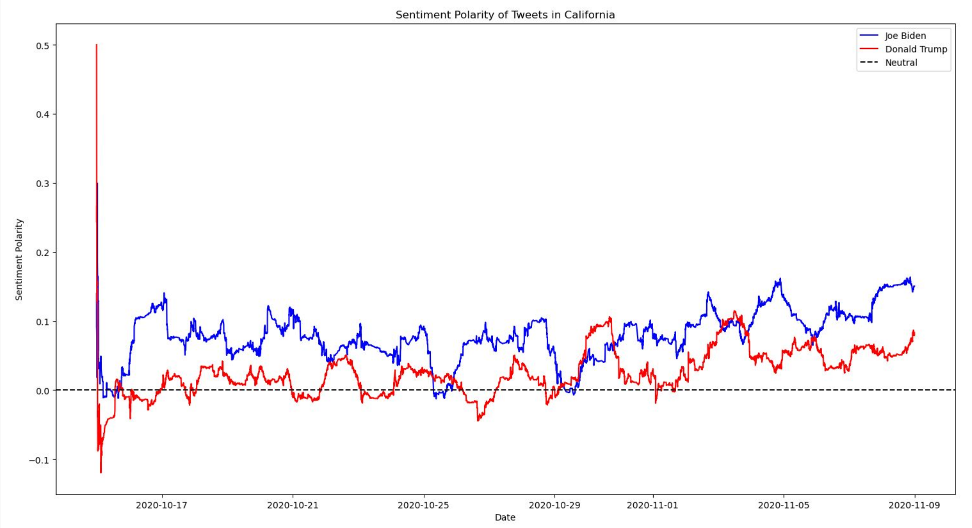
Constructing the sentiment time series

Score sentiment per tweet

Rolling window: Calculate weighted mean in a 24-hour window per tweet

Weighted mean: 2log(#likes) + 5log(#retweets)

Gaussian filter: To remove some noise



‘Vectorizing’ the time series + analysis

Feature extraction: Perform various statistical tests on the time series

Examples: difference in mean, DTW-distance,

Cross-correlation and ks-test value

PCA: In order to analyze the features which contribute most variance

Logistic regression

Use the state features to predict the color of the swing states

KMeans clustering

See if red states and blue states cluster together. Will also be compared to DTW clustering

VAR Analysis

VAR: Vector Autoregression

Multivariate time series (multiple target variables)

Forecast a series based on past values in the series - **lags**

Bivariate VAR(1) with no intercepts:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-2} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \end{bmatrix}$$

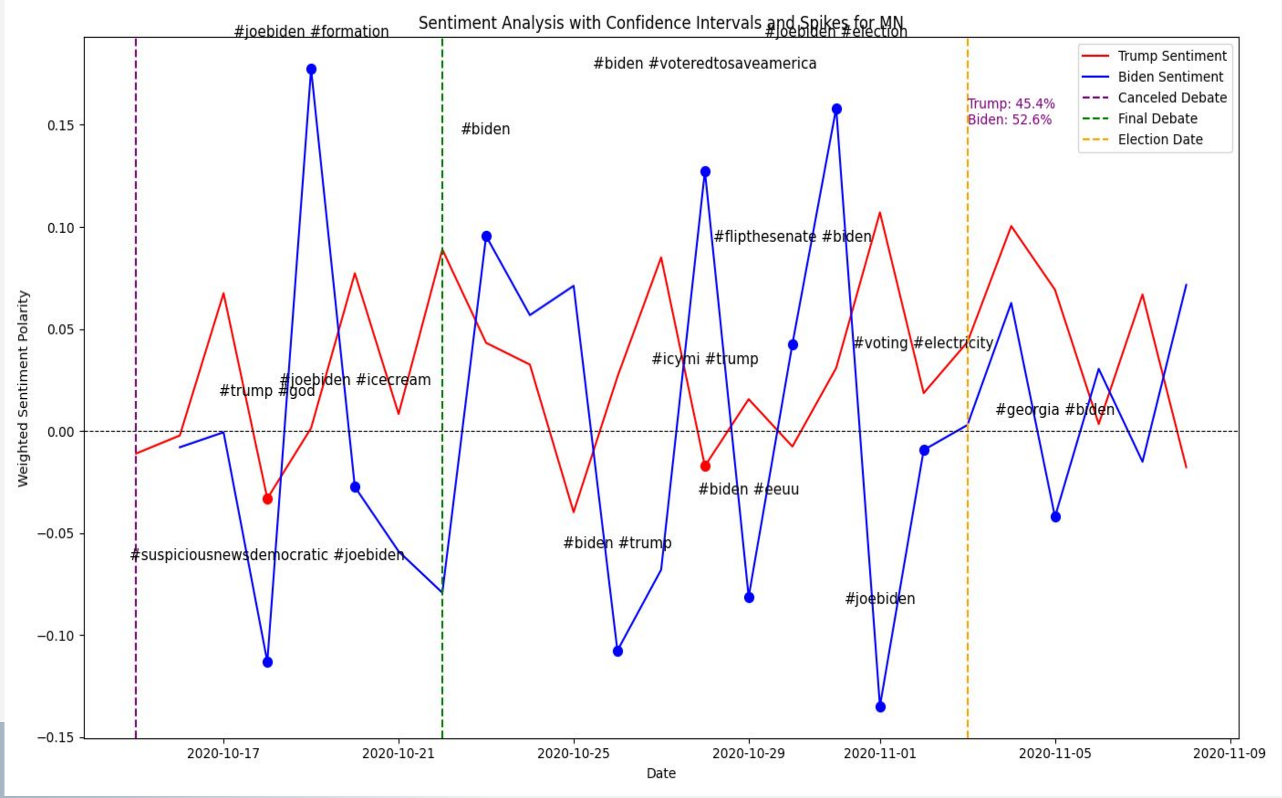
ADF: Augmented Dickey Fuller test

Applied **differencing** to attempt stationarity

VAR is applied only if both series are stationary

Hashtags purpose: To understand and explain trends and peaks in the sentiment time series

Construction of VAR

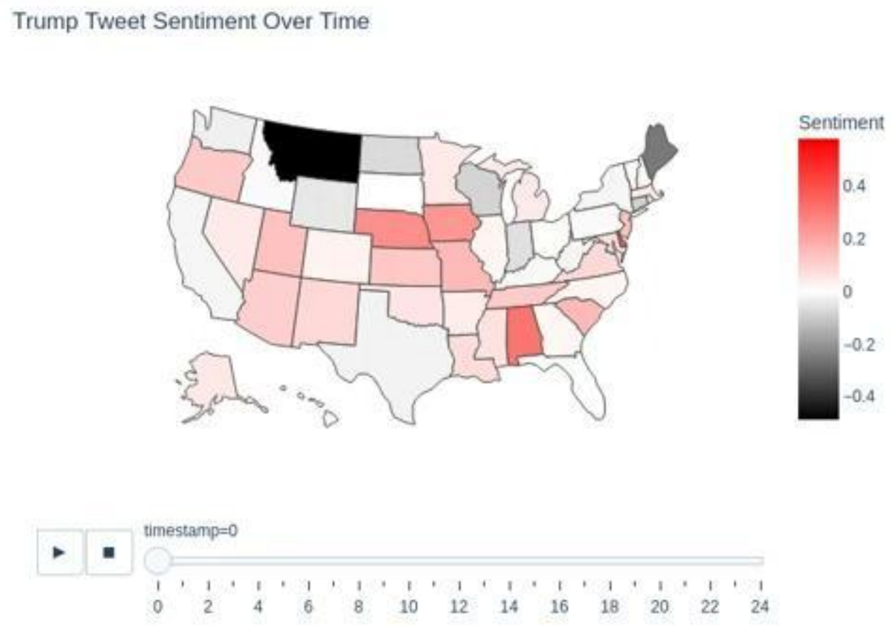


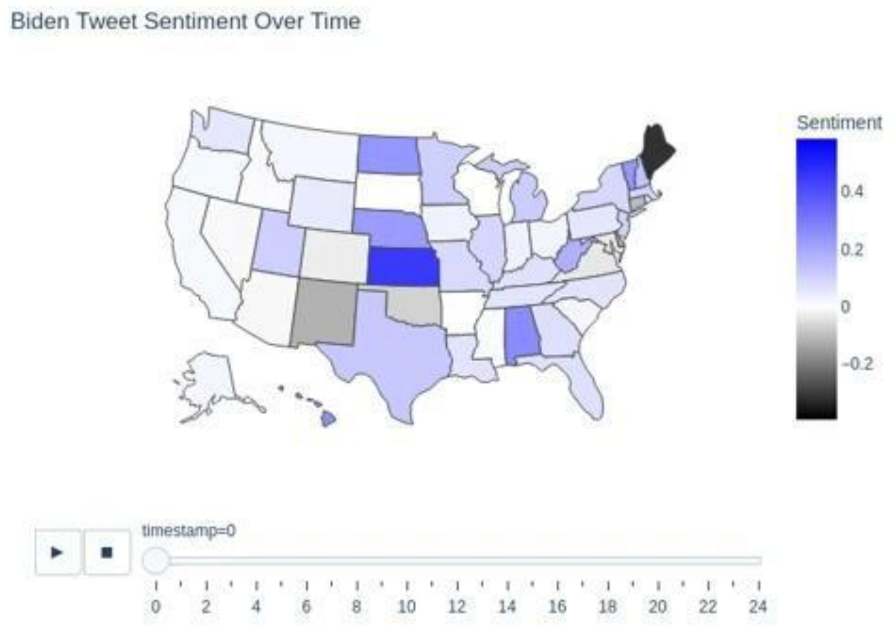
3

Results

What have we achieved?

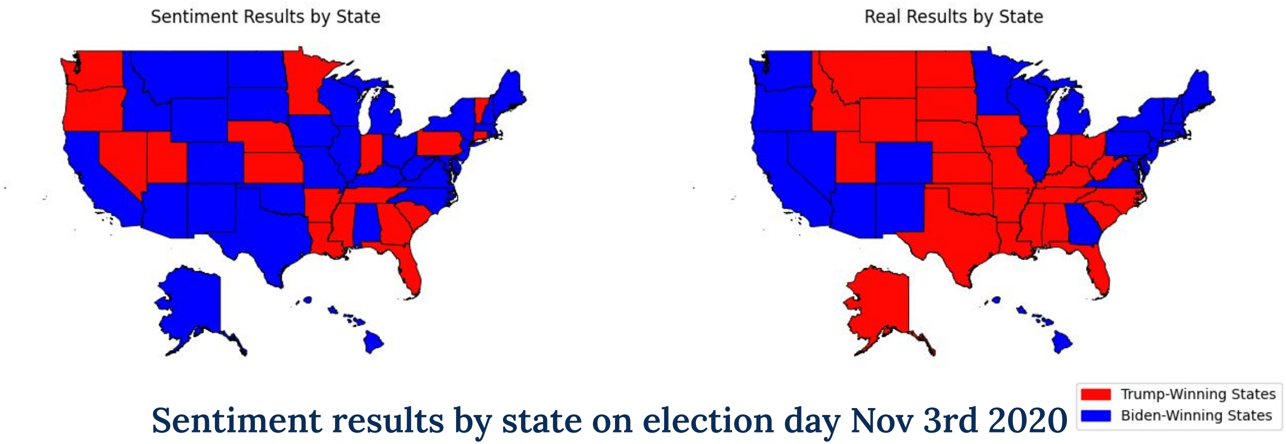
Visualizing the time series on geomap





Comparing Sentiment with Actual Election Outcomes

How well did sentiment analysis predict the results?



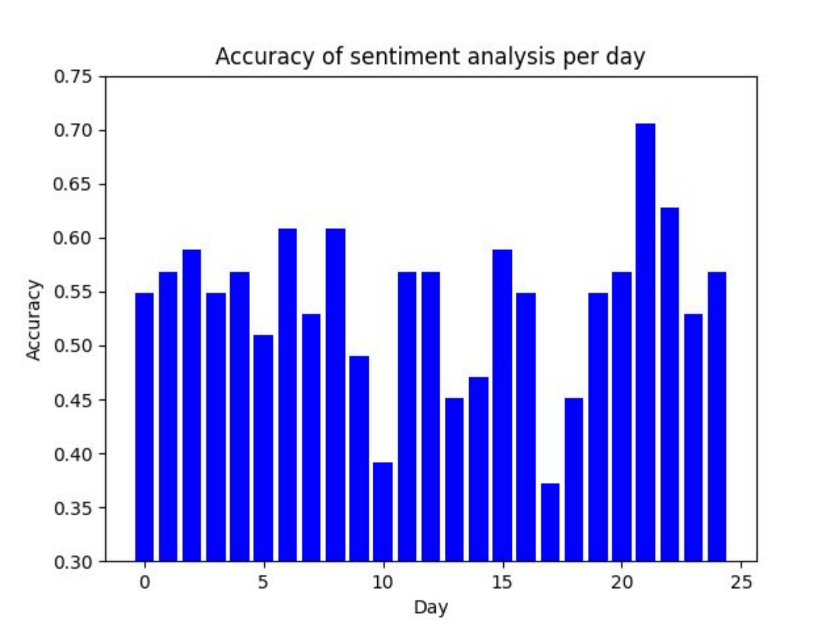
Sentiment results by state on election day Nov 3rd 2020

27 out of 51 states correct

53%

Comparing Sentiment with Actual Election Outcomes

How well did sentiment analysis predict the results?



Accuracy between **40%** and **70%**

Highest accuracy 2 days after election

Logistic Regression Results

Swing States

Test robustness: Add noise to the features

Performance:

- Predicts 11/14 swing states correctly
- Cross-validation accuracy of 64%
- F1-score of 77%

Misclassification: Very often Iowa, Wisconsin and Pennsylvania

95% Confidence intervals:

- Accuracy: [0.57, 0.86]
- Cross-validation accuracy: [0.44, 0.73]
- F1-score: [0.57, 0.90]

H_0: A regression model trained on Twitter data cannot predict outcomes more accurately than random chance

One-sample t-test: Reject the null hypothesis! All statistics have a p-value of 0 when the t-test pop-mean is 50%

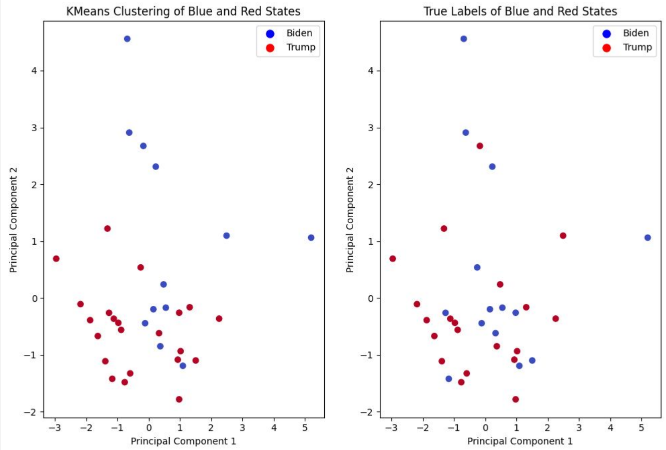
Conclusion: Can complement traditional methods, but is unlikely to replace it entirely

Clustering Results

KMeans clustering

Accuracy: 71%
Silhouette score: 0.227
P-value: 0.03

Fail to reject H0: Clustering of red and blue are not distinct.
Reject H0: Clustering is no better than random.



DTW clustering

Accuracy: 59%
Silhouette score: 0.494
P-value: 0.41

Fail to reject H0: Clustering of red and blue are not distinct.
Fail to reject H0: Clustering is no better than random.

PCA: biden_kurotisis, skew_biden, cross-correlation and difference in mean contribute 99.6% of the variance.

Silhouette score: A measure of how well data points in a clustering are assigned to their clusters.
Permutation test: Compares clusters to random assignments

VAR Results

Granger Causality: Determine whether one time series is useful for forecasting another.

Example Trump -> Biden for state MN and after 5 lags:

ssr based F test: **p-value=**0.6492

ssr based chi2 test: **p-value=**0.1817

likelihood ratio test: **p-value=**0.2676

parameter F test: **p_value=**0.6492

H_0: A candidate's sentiment time series does not Granger-cause another candidate's sentiment time series.

Conclusion: Fail to reject H_0 meaning the candidate's sentiment time series is not useful for forecasting the other candidate's time series.

Thank you!

Are there any questions?