Research Presentation

David Grinberg

Polling Data

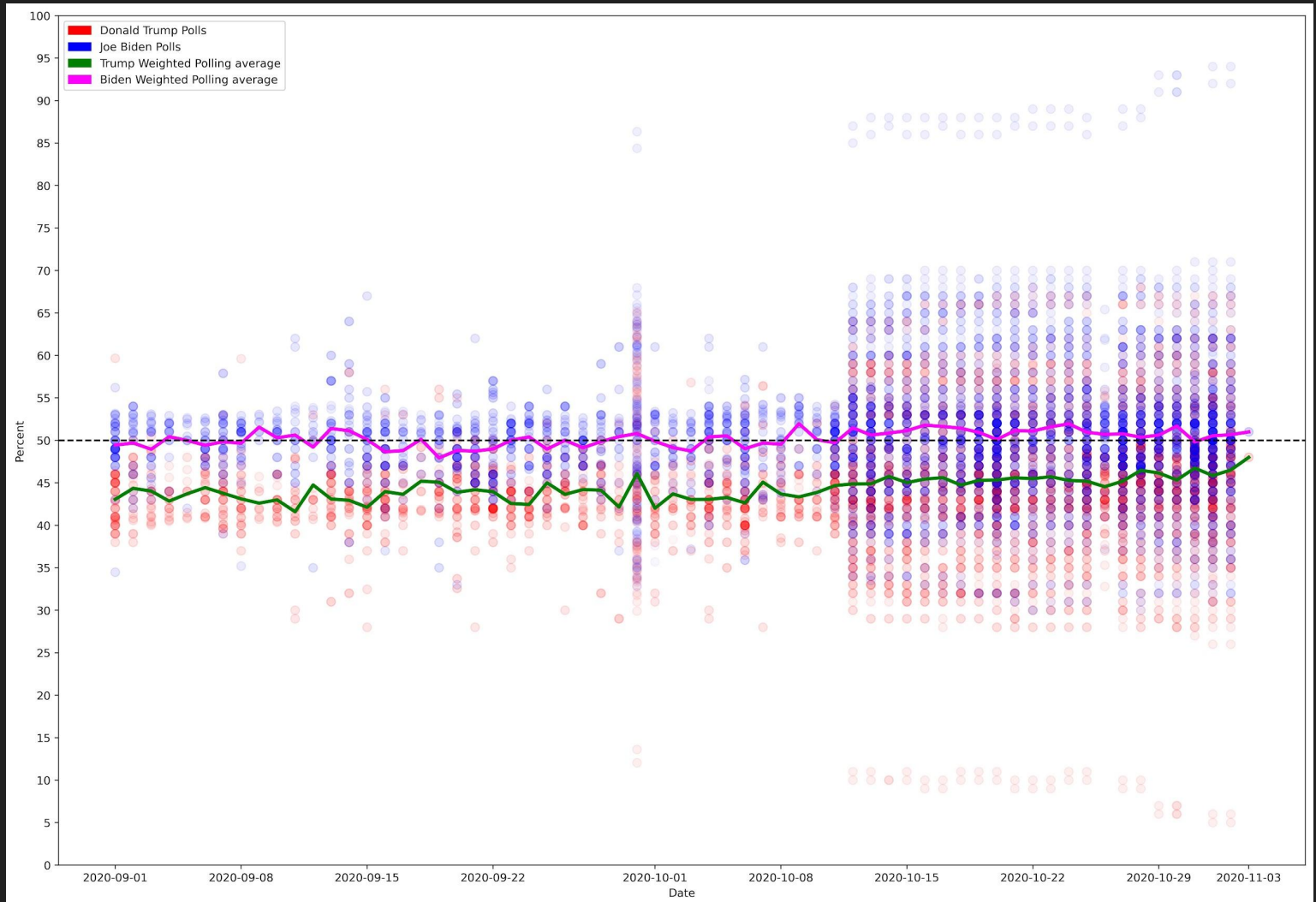
2020 Election Polling Data

- Fivethirtyeight grades polls from F -> A+ based off the methodology, sample size, and previous bias of the pollsters
- I created a numeric weight for each poll based on the letter grade
- A weighted average was calculated using the formula below

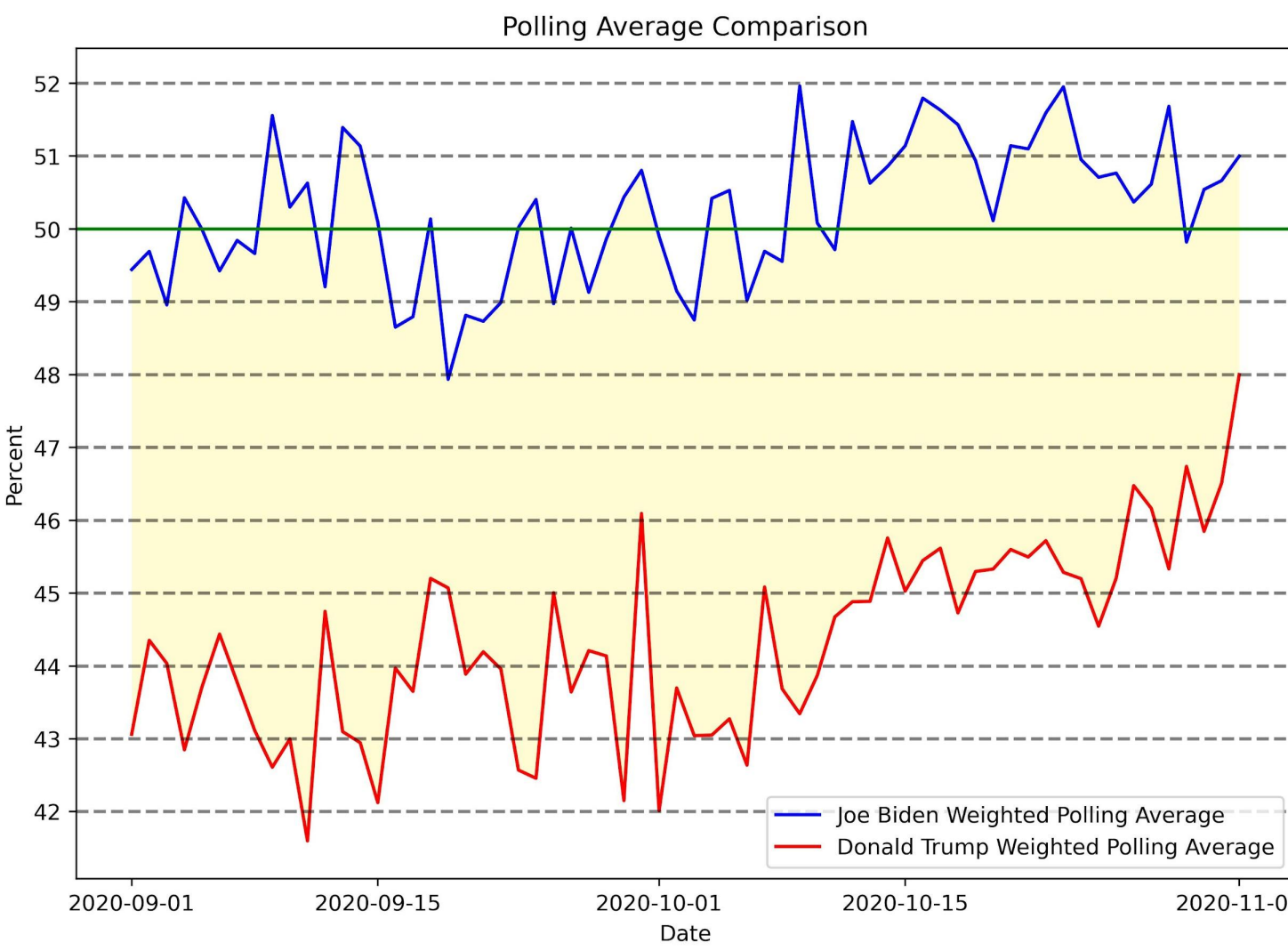
$$\mu_{day} = \frac{\Sigma(pollpercentage * weight)}{\Sigma(weight)}$$

Letter Grade	Weight (fraction)	Weight (Decimal)
A+	13/13	1.00
A	12/13	0.923..
A-	11/13	0.846..
A/B	10/13	0.769..
B+	9/13	0.692..
B	8/13	0.615..
B-	7/13	0.538..
B/C	6/13	0.461..
C+	5/13	0.384..
C	4/13	0.307..
C-	3/13	0.230..
C/D	2/13	0.153..
F	1/13	0.076..
NaN	0/13	0.00

2020 Election Polling Data



Polling Averages



Latent Sentiment Analysis (LSA)

VADER SENTIMENT

VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text

C.J. Hutto **Eric Gilbert**
Georgia Institute of Technology, Atlanta, GA 30032
cjhutto@gatech.edu gilbert@cc.gatech.edu

Abstract

The inherent nature of social media content poses serious challenges to practical applications of sentiment analysis. We present VADER, a simple rule-based model for general

zales, & Booth, 2007). Sociologists, psychologists, linguists, and computer scientists find LIWC appealing because it has been extensively validated. Also, its straightforward dictionary and simple word lists are easily inspect

- Sentiment Analysis works by “reading” the text to find how positive or negative the text is, and assigns a score accordingly:
  - “I really like the new design of your website!” → Positive → (+1)
  - “I’m not sure if I like the new design” → Neutral → (+0)
  - “The new design is awful!” → Negative → (-1)
- The VADER method works by using a Lexicon of words, emojis, and phrases, which all have positive, neutral, or negative values assigned to them. These values are then plugged into an algorithm which calculates a sentiment score.

Subjects of Sentiment Analysis

Politicians Analyzed

- Analyzed the VADER Sentiment Analysis scores for a total of 25 politicians, as well as the two political parties (and their nicknames). 16 were democrats, while 9 were republicans.
- Of the 25 politicians:
  - 7 were women and all 7 of those women were democrats
    - Nancy Pelosi, AOC, Ilhan Omar, Amy Klobuchar, Elizabeth Warren, Tulsi Gabbard, and Kamala Harris.
  - 7 of those politicians in the data-set were either up for election, or part of an electoral ticket (in the case of vice presidents)
    - Mitch McConnell (R), Lindsey Graham (R), Cory Booker (D), Joe Biden (D), Kamala Harris (D), Donald Trump (R), Mike Pence (R),
  - 16 were democrats while 9 were republicans

Politicians Analyzed

Democrats

- Bernie Sanders
- Nancy Pelosi
- Alexandria Ocasio-Cortez
- Ilhan Omar
- Amy Klobuchar
- Gavin Newsom
- Chuck schumer
- Elizabeth Warren
- Cory Booker
- Beto O'Rourke
- Michael Bloomberg
- Andrew Yang
- Tulsi Gabbard
- Pete Buttiegieg
- Kamala Harris
- Joe Biden

Republicans

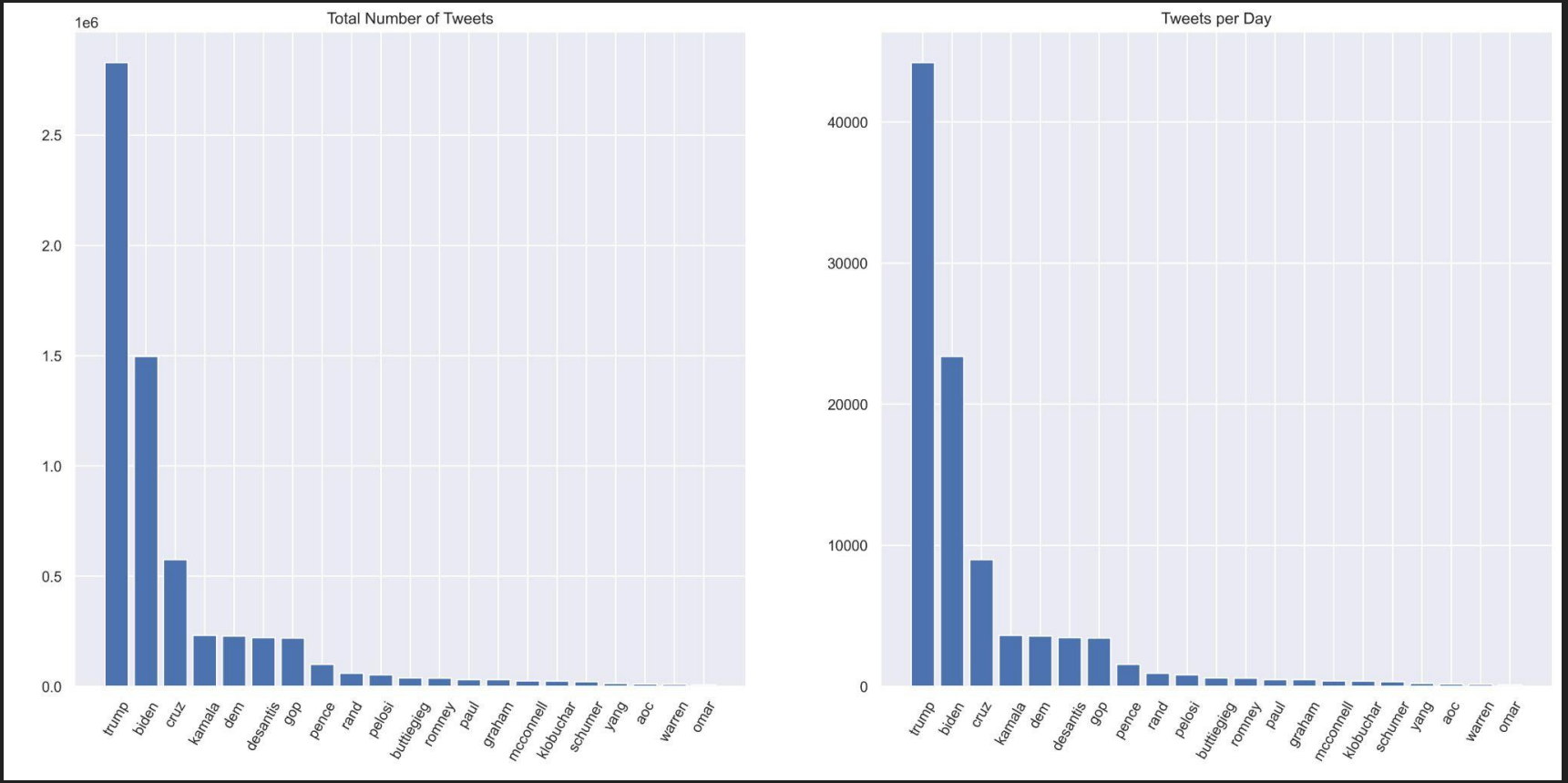
- Rand Paul
- Paul Ryan
- Marco Rubio
- Mitt Romney
- Mitch McConnell
- Ted Cruz
- Lindsey Graham
- Mike Pence
- Donald Trump

Frequency of Tweets

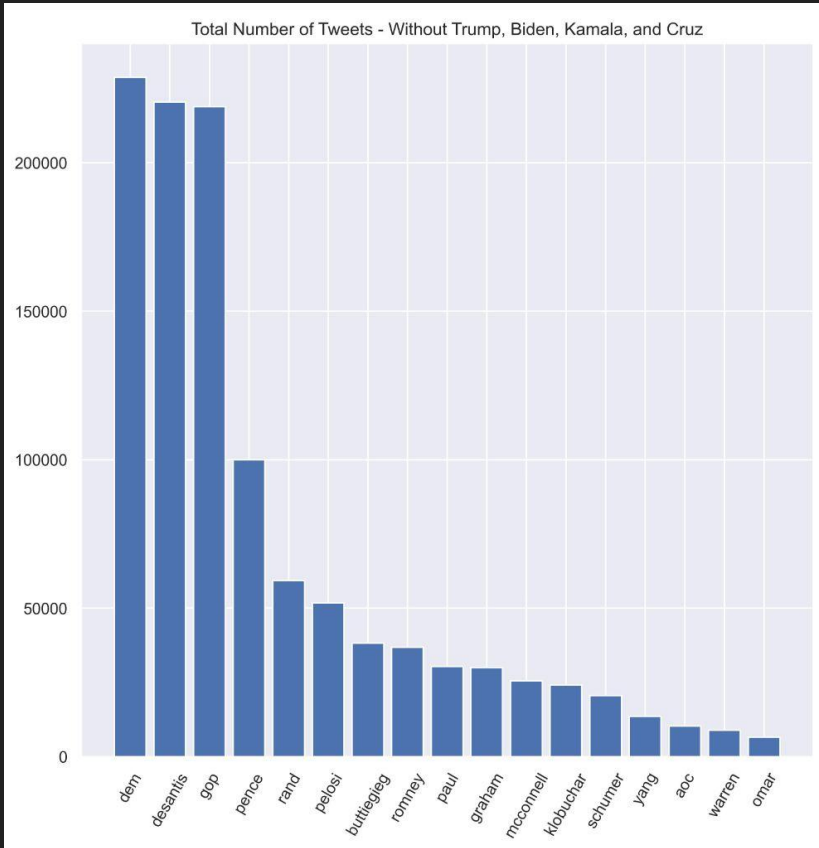
- 6/25 of the politicians analyzed were mentioned less than 100 times per day in our data set
    - Michael Bloomberg (D), Cory Booker (D), Gavin Newsom (D), Tulsi Gabbard (D) (was a democrat at the time), Beto O' Rourke (D), & Marco Rubio (R)

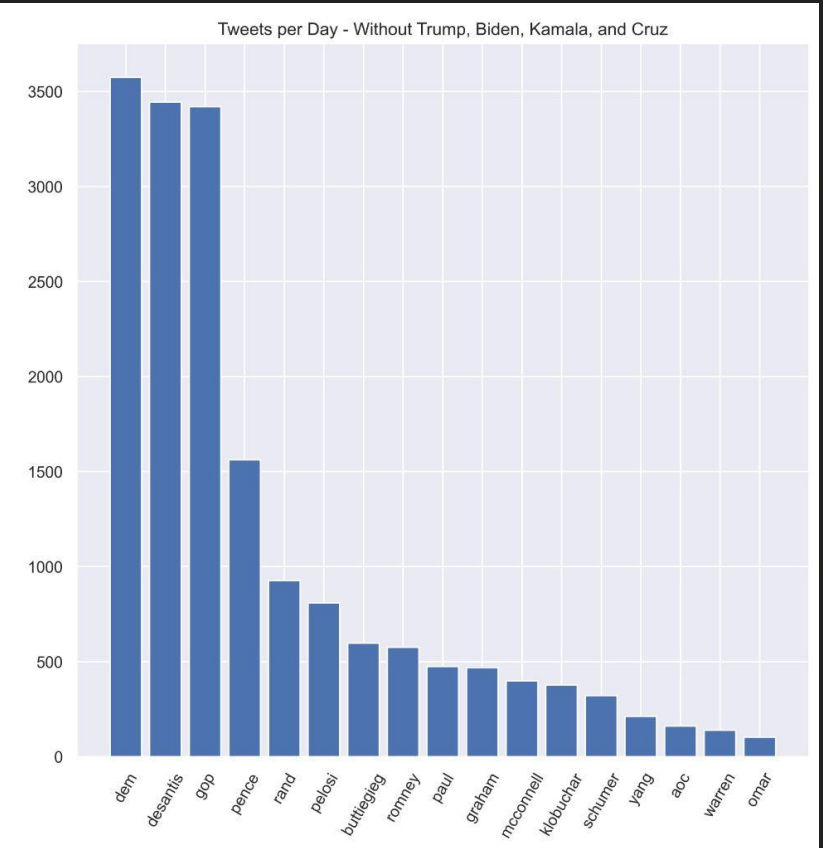
	total	tweets per day
trump	2828995	44203.046875
biden	1496137	23377.140625
cruz	575188	8987.312500
kamala	230698	3604.656250
dem	228775	3574.609375
desantis	220483	3445.046875
gop	218891	3420.171875
pence	99994	1562.406250
rand	59317	926.828125
pelosi	51749	808.578125
buttiegieg	38176	596.500000
romney	36791	574.859375
paul	30366	474.468750
graham	30016	469.000000
mcconnell	25487	398.234375
klobuchar	24138	377.156250
schumer	20496	320.250000
yang	13539	211.546875
aoc	10332	161.437500
warren	8860	138.437500
omar	6546	102.281250
bloomberg	5752	89.875000
rubio	5204	81.312500
booker	4418	69.031250
newsom	4069	63.578125
tulsi	2172	33.937500
beto	2058	32.156250

Frequency of Tweets

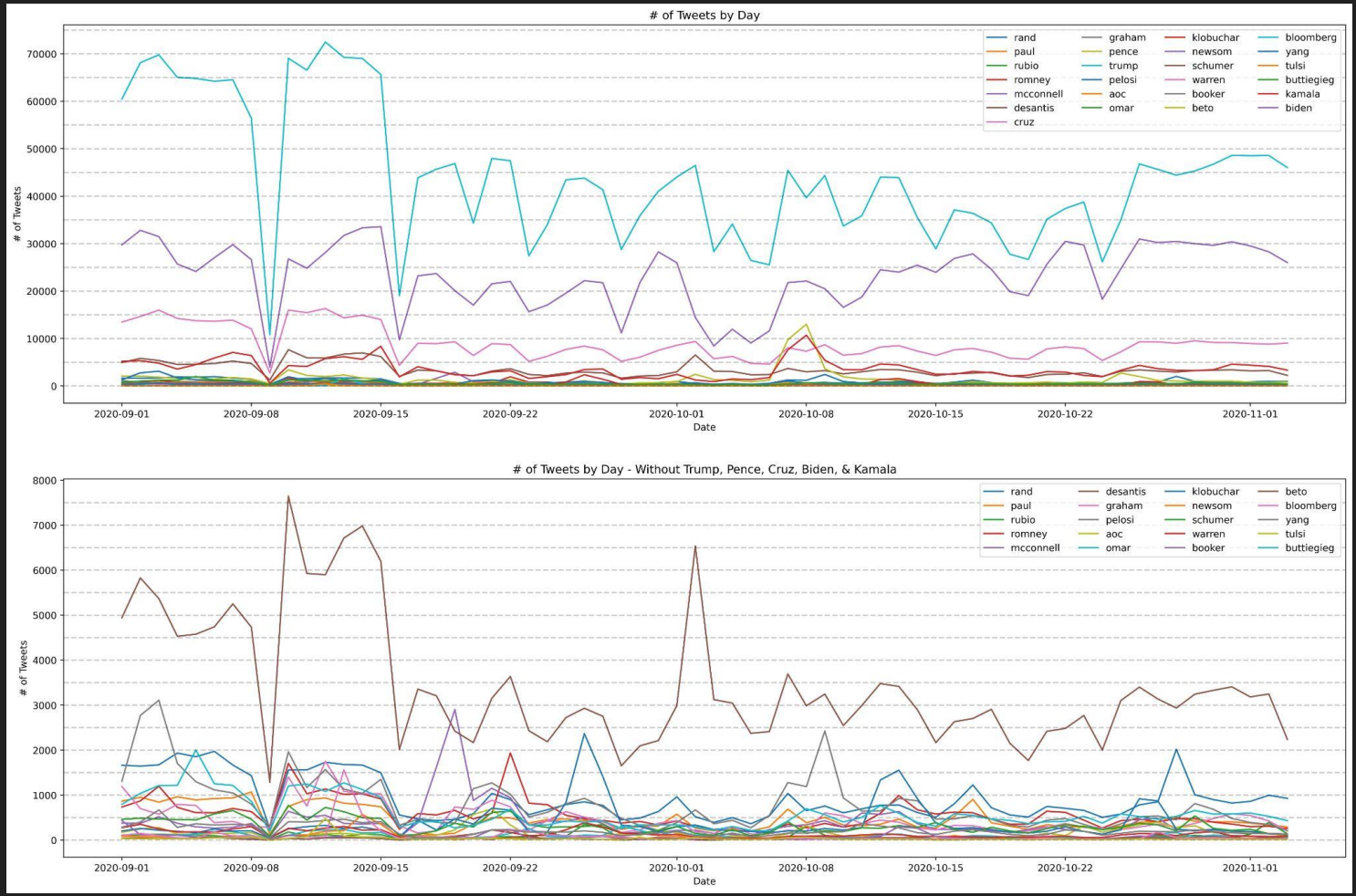


Frequency of Tweets - Excluding most popular politicians





Frequency of Tweets



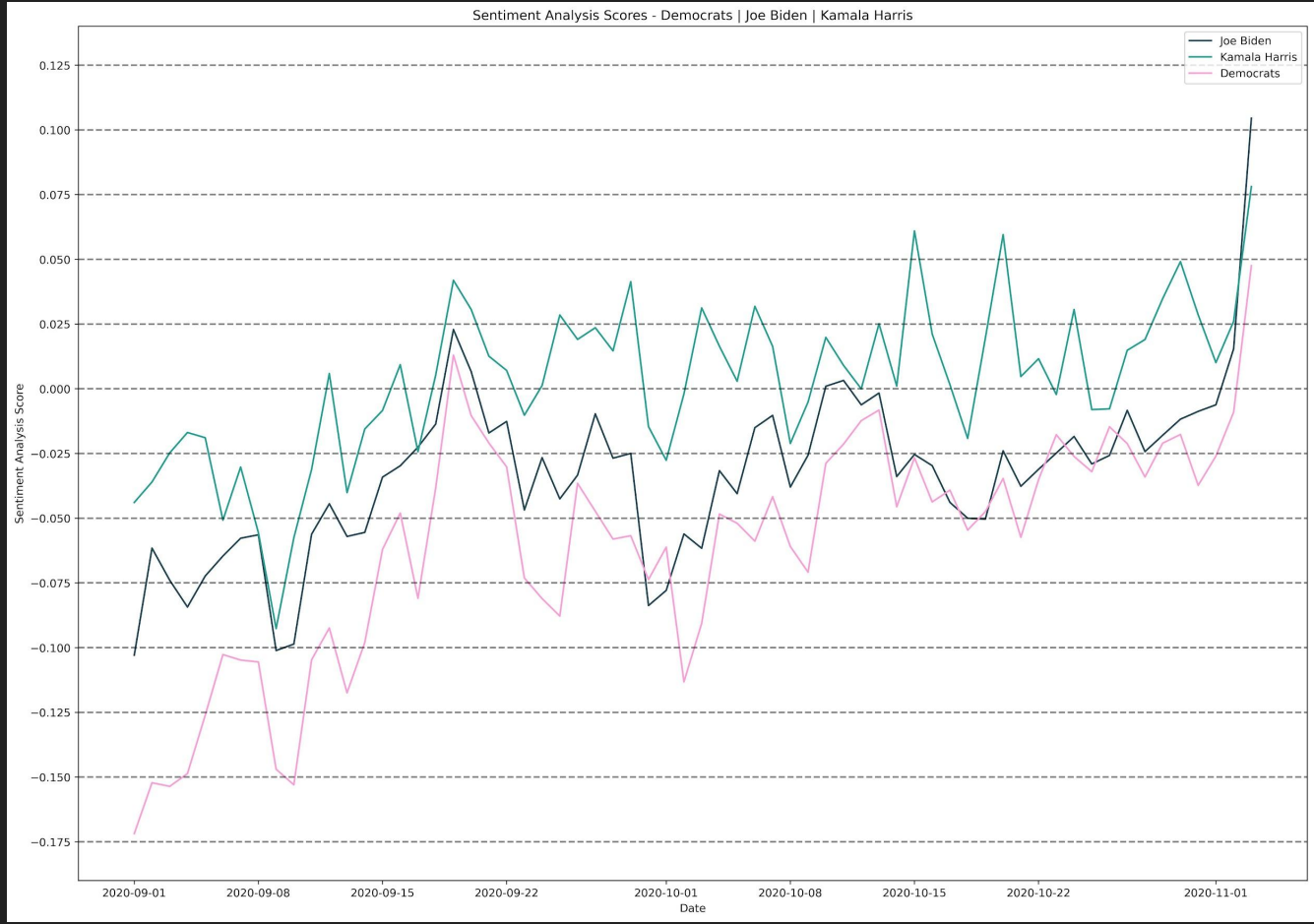
Simple Correlation tests

Trump/Biden - Polling and Sentiment Correlation Matrix

item 1	item 2	correlation
Biden Polling	Trump Sentiment	0.0481
Biden Polling	Biden Sentiment	-0.003
Trump Polling	Biden Sentiment	*0.5381*
Trump Polling	Trump Sentiment	*0.5725*
Biden Sentiment	Trump Sentiment	**0.8636**
Biden Polling	Trump Polling	0.2458

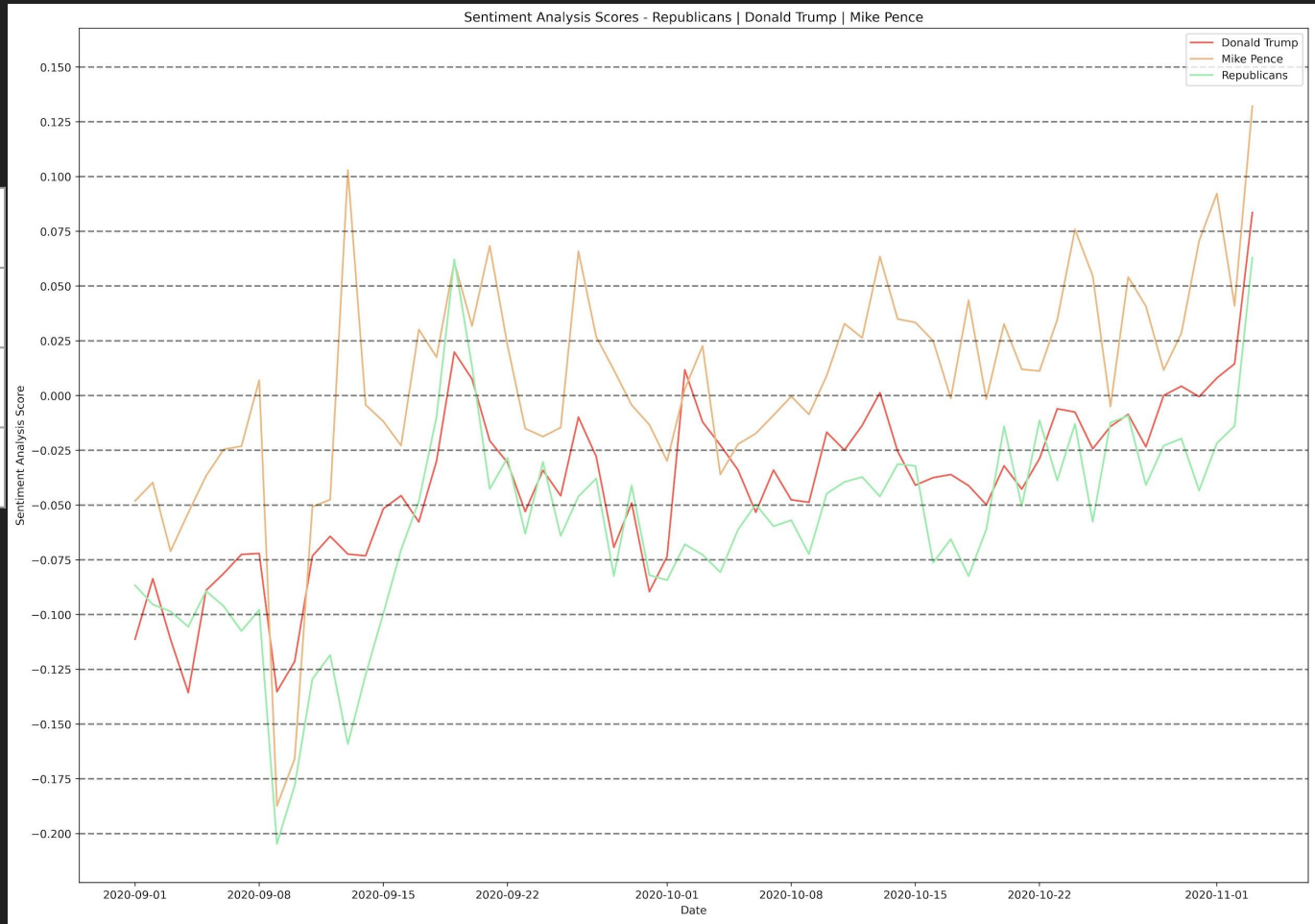
LSA Correlation: Democrats, Biden, & Kamala

Item 1	Item 2	Correlation
Biden	Kamala	0.7505
Biden	Dem	0.8612
Kamala	Dem	0.7502

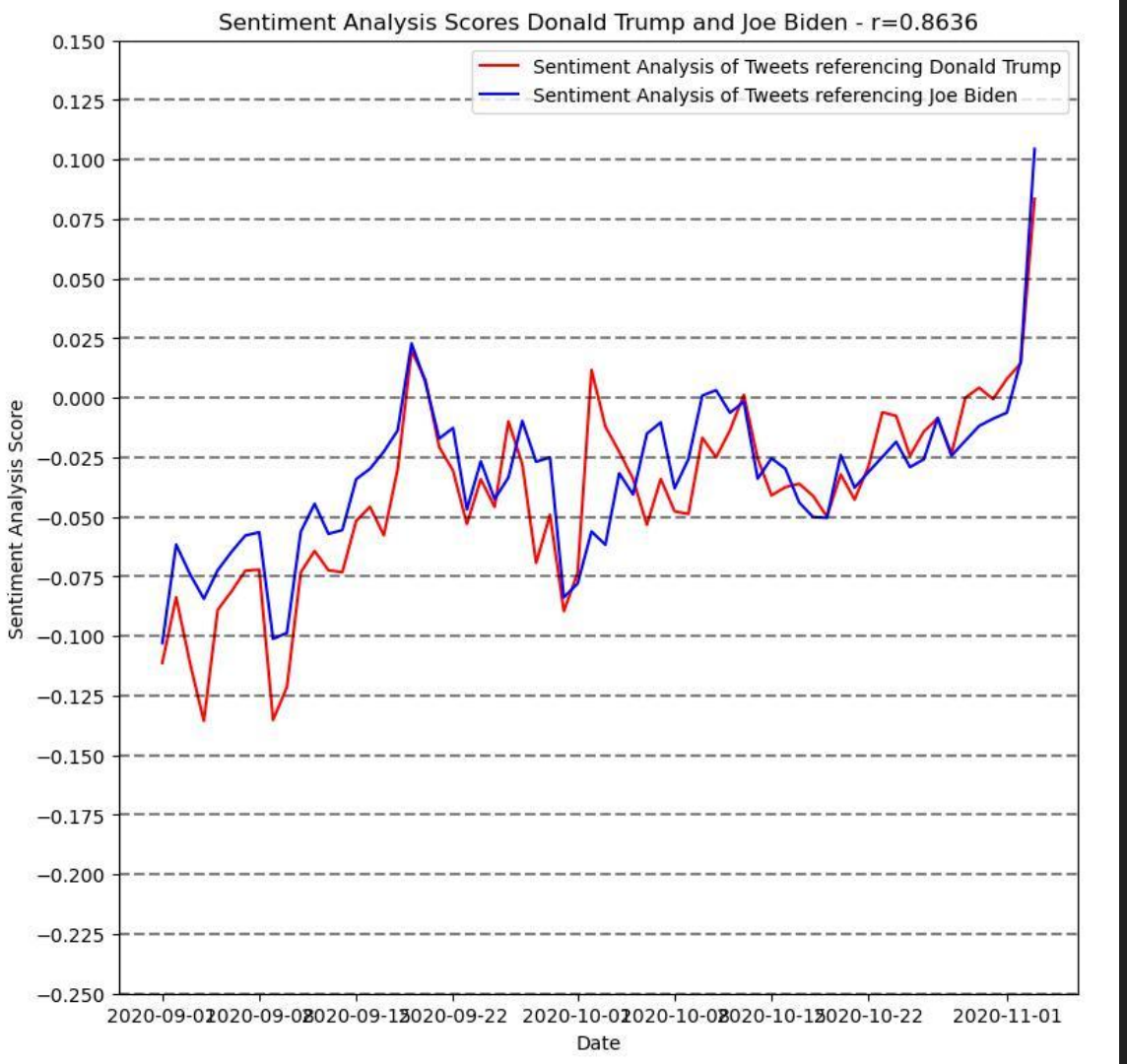


LSA Correlation: Republicans, Trump, & Pence

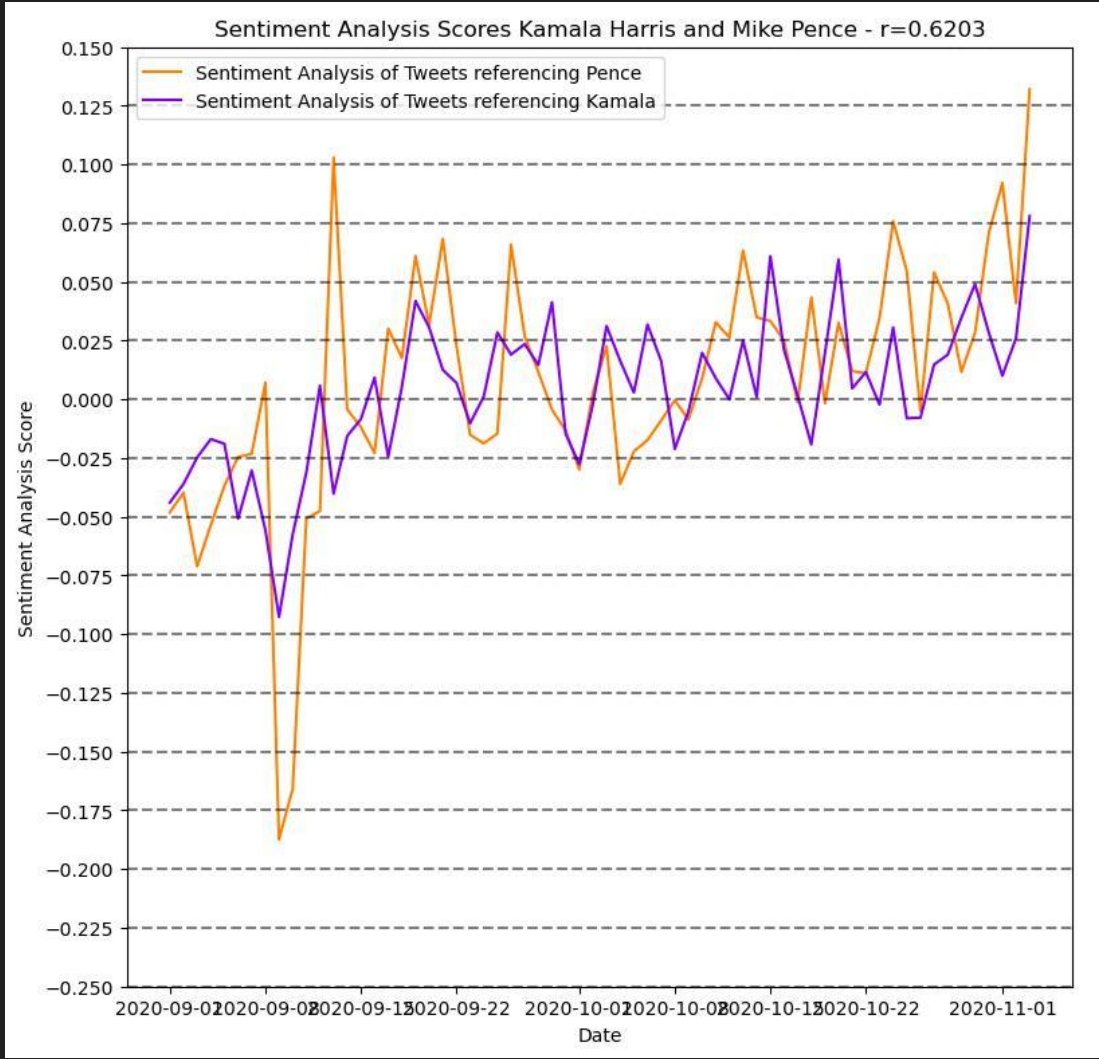
Item 1	Item 2	Correlation
Trump	Pence	0.7704
Trump	GOP	0.8222
Pence	GOP	0.6933



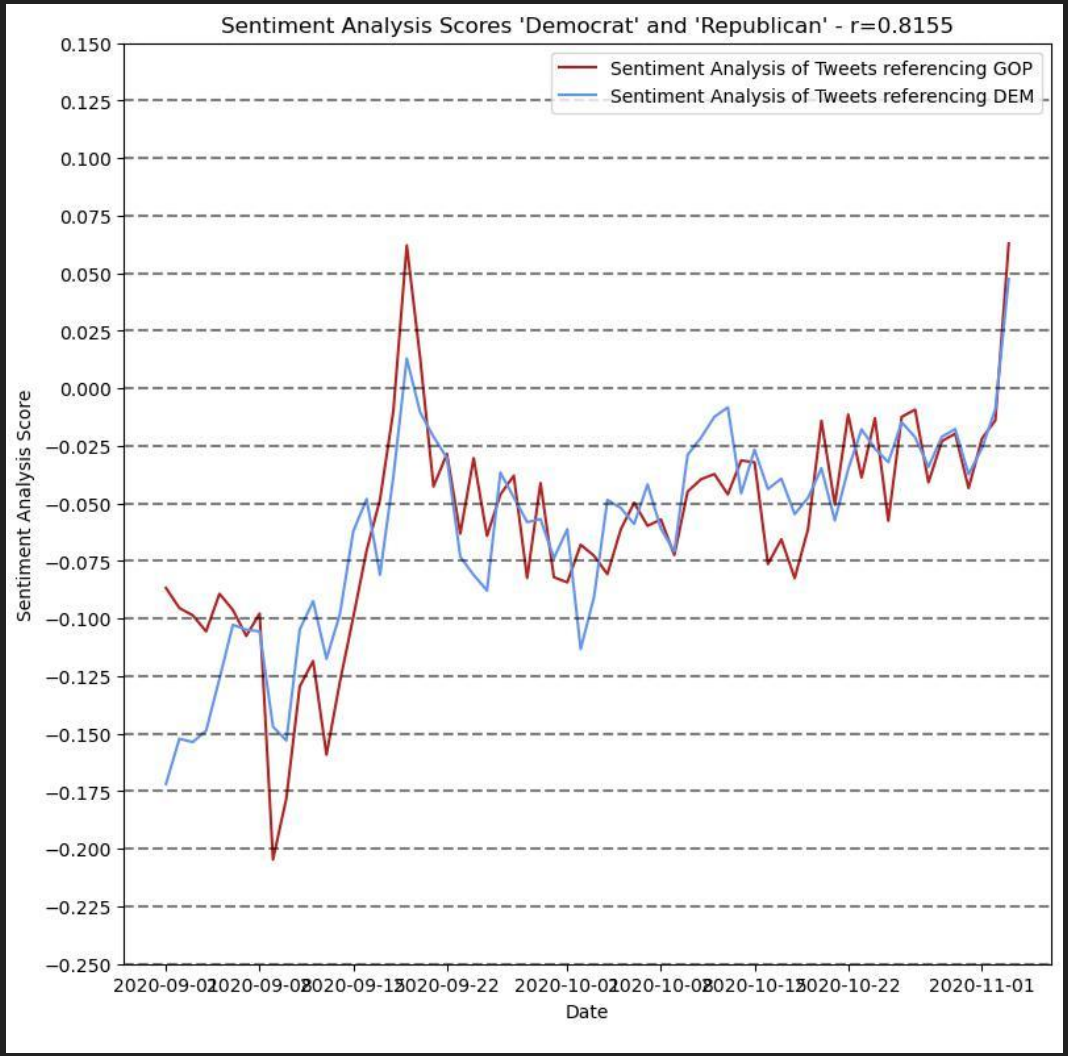
Presidential Candidates



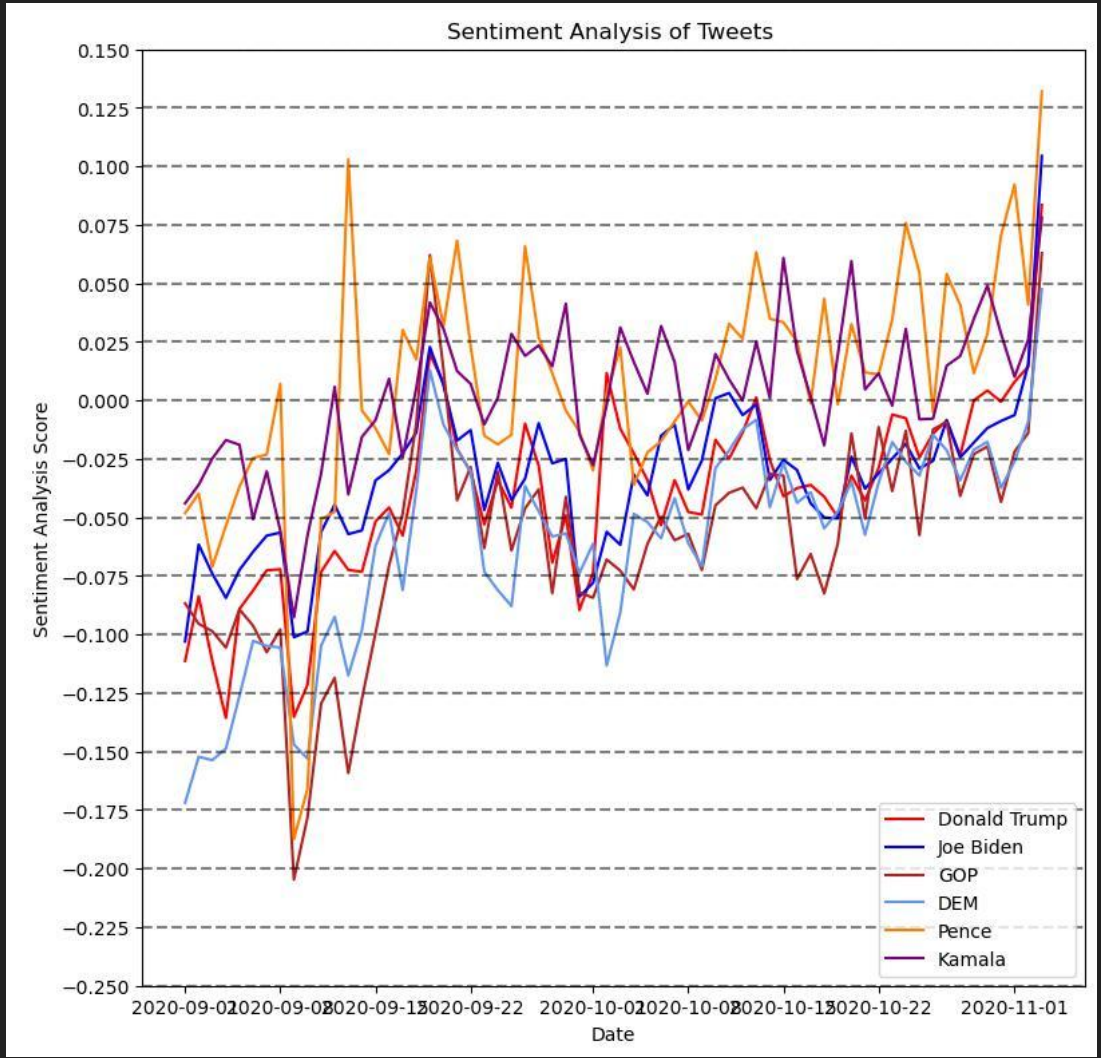
Vice Presidential Candidates



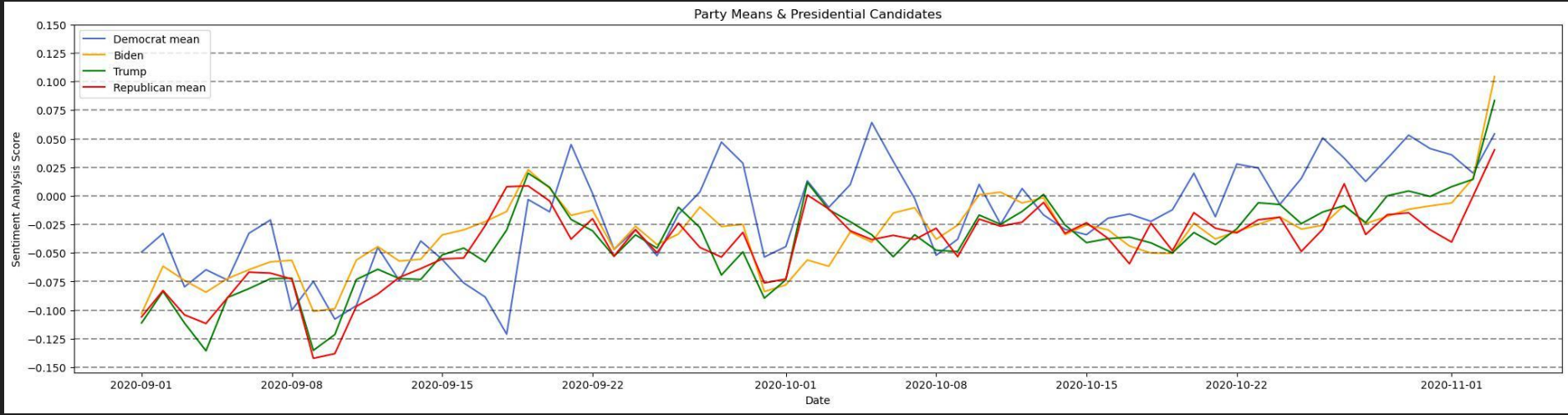
Political Parties



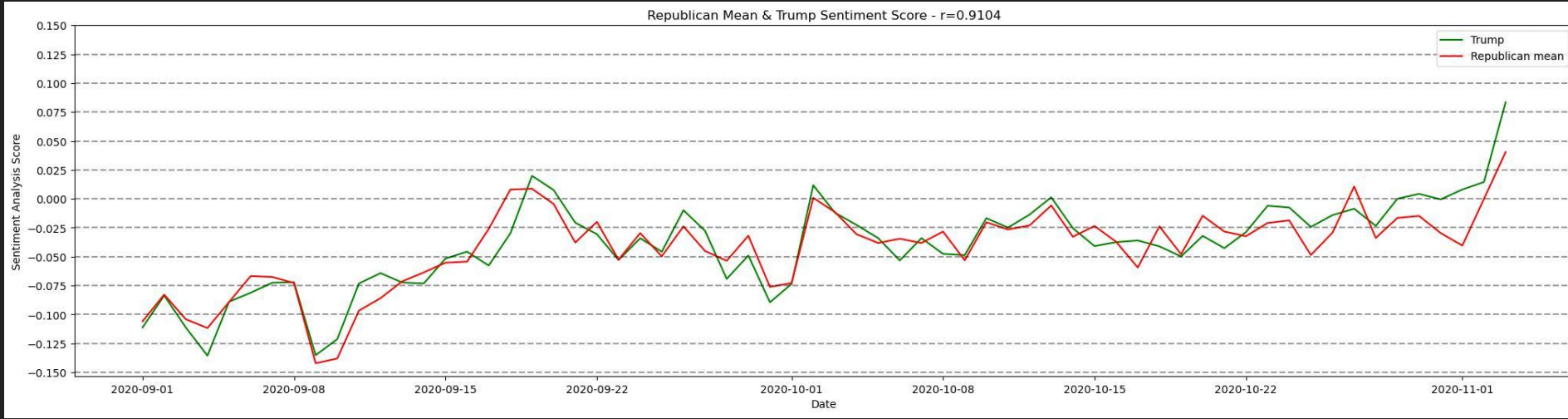
President, Political Parties, & Vice President



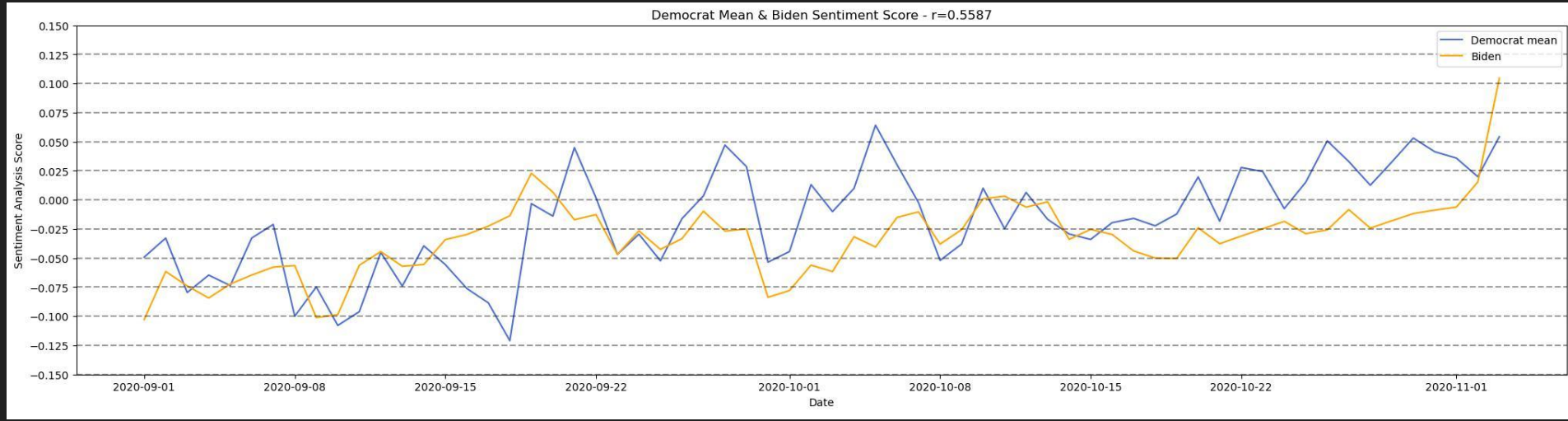
Political Party & Presidential Candidate



Republicans & Trump

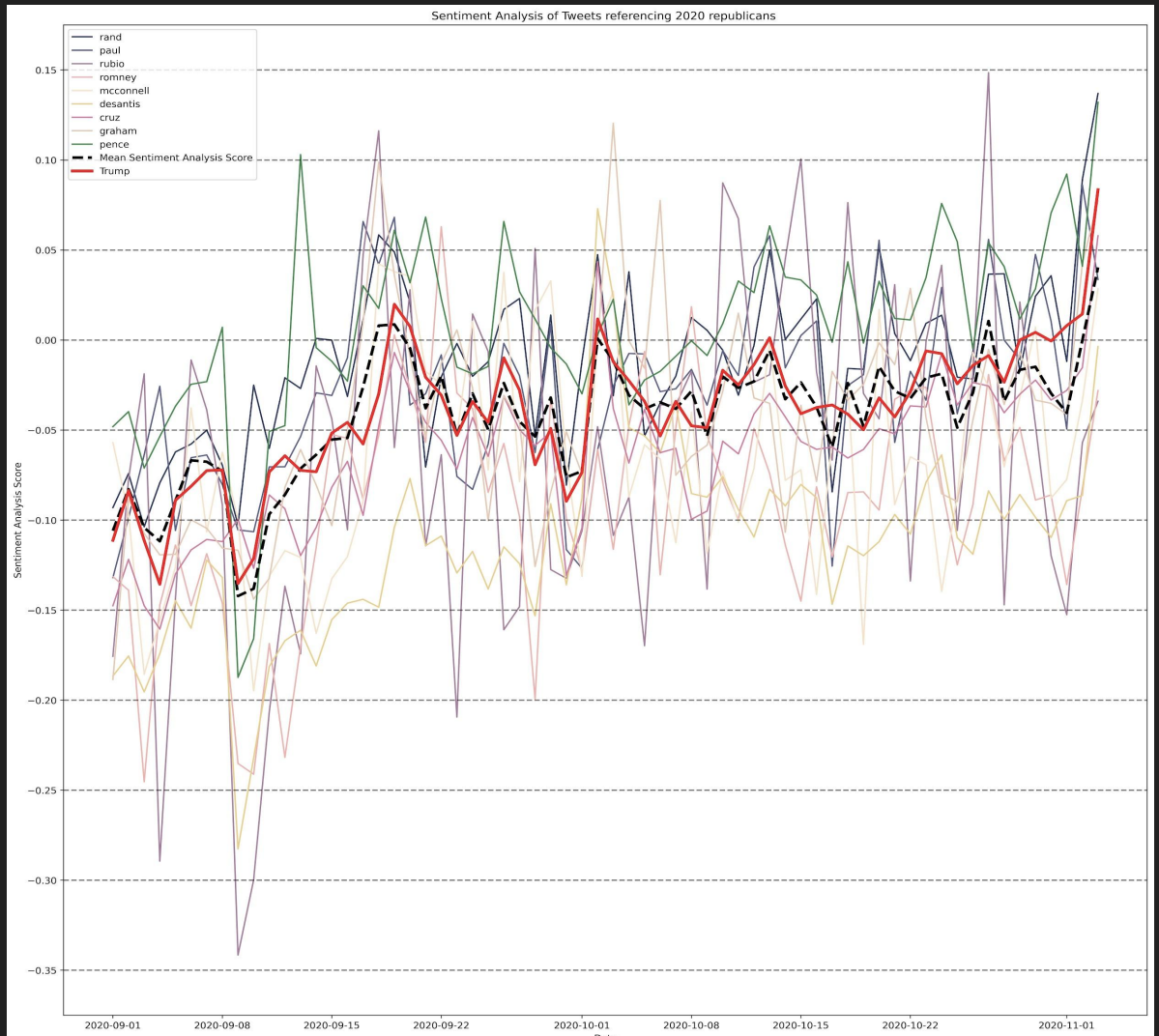


Democrats & Biden



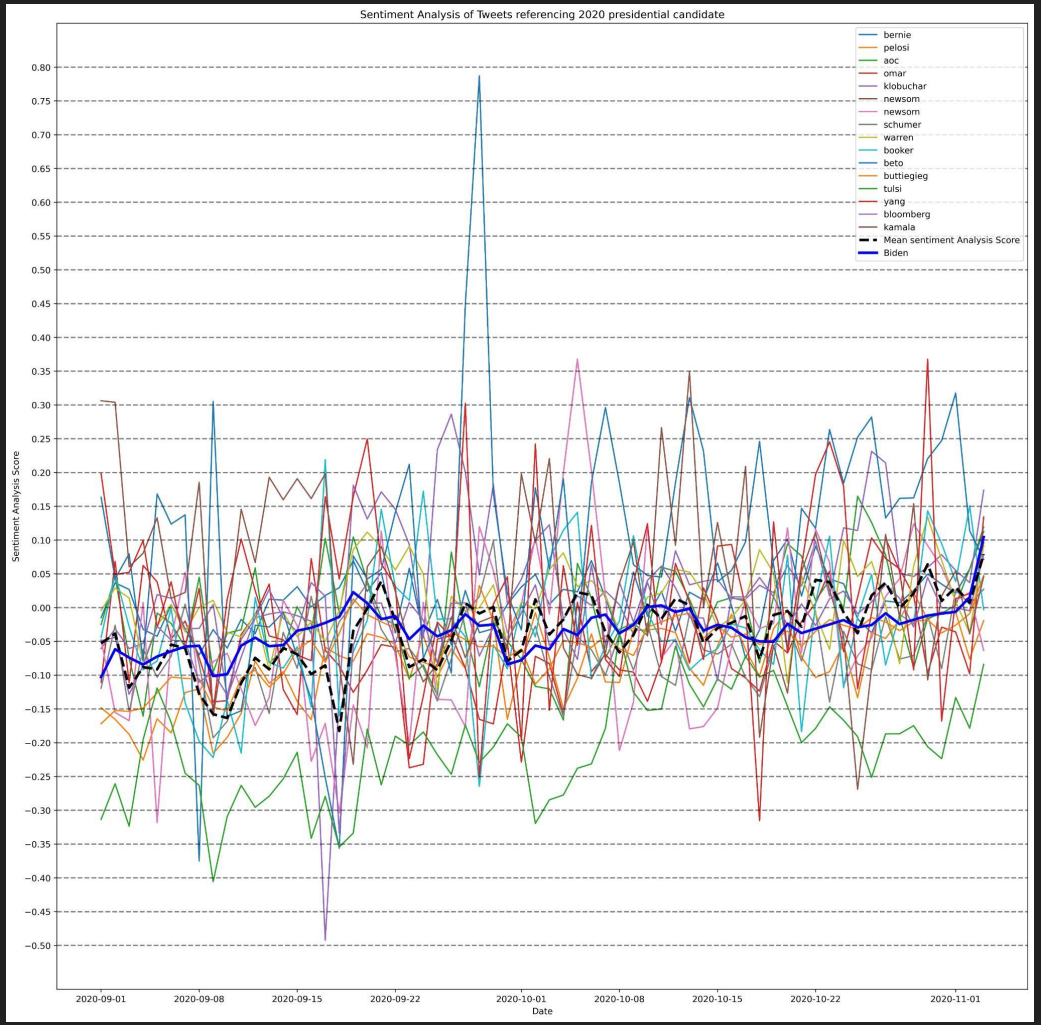
Republican Politicians

1. Rand Paul
2. Paul Ryan
3. Marco Rubio
4. Mitt Romney
5. Mitch McConnell
6. Ted Cruz
7. Lindsey Graham
8. Mike Pence
9. Donald Trump

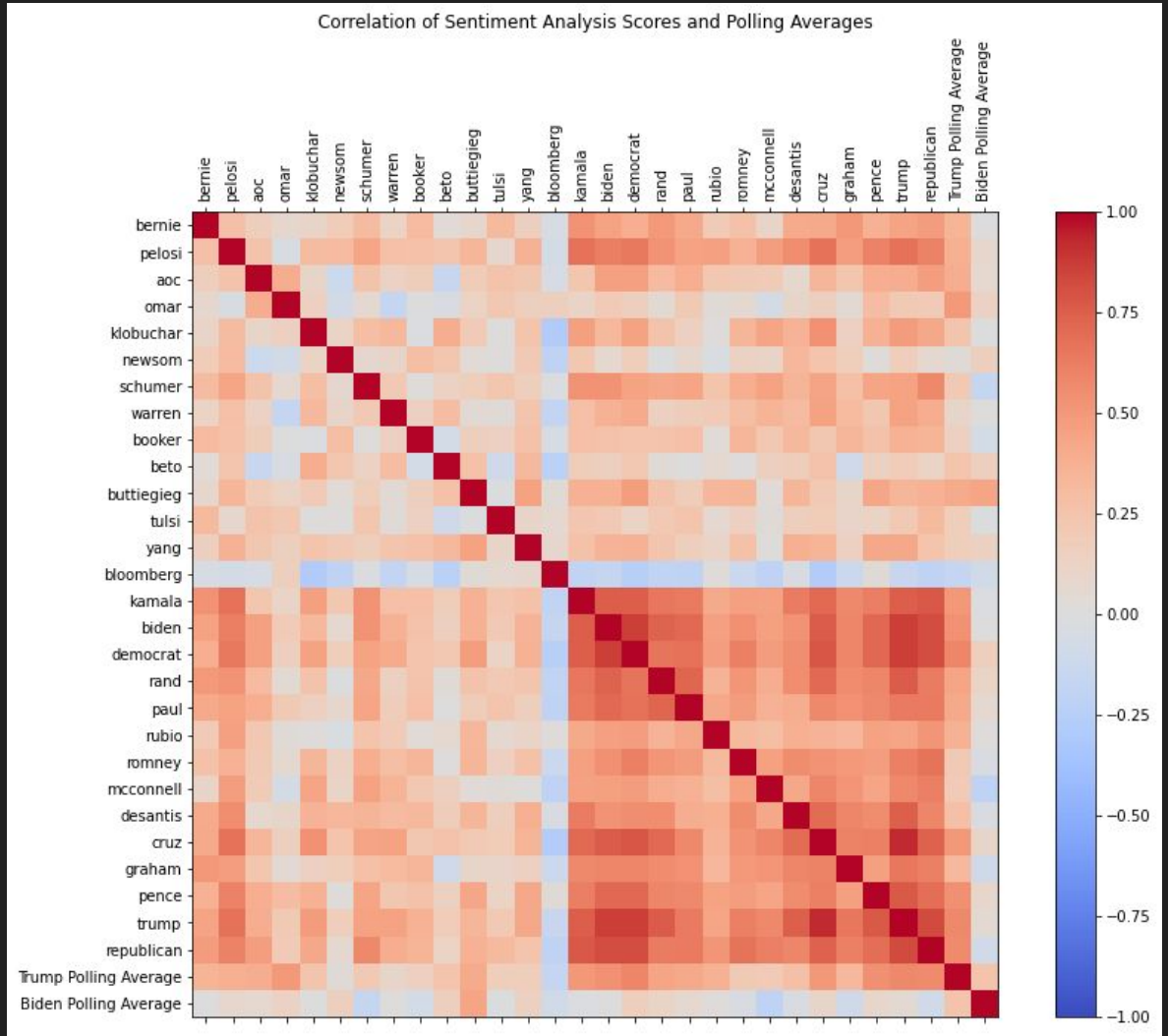


Democrat Politicians

1. Bernie Sanders
2. Nancy Pelosi
3. Alexandria Ocasio-Cortez
4. Ilhan Omar
5. Amy Klobuchar
6. Gavin Newsom
7. Chuck schumer
8. Elizabeth Warren
9. Cory Booker
10. Beto O’Rourke
11. Michael Bloomberg
12. Andrew Yang
13. Tulsi Gabbard
14. Pete Buttiegieg
15. Kamala Harris
16. Joe Biden

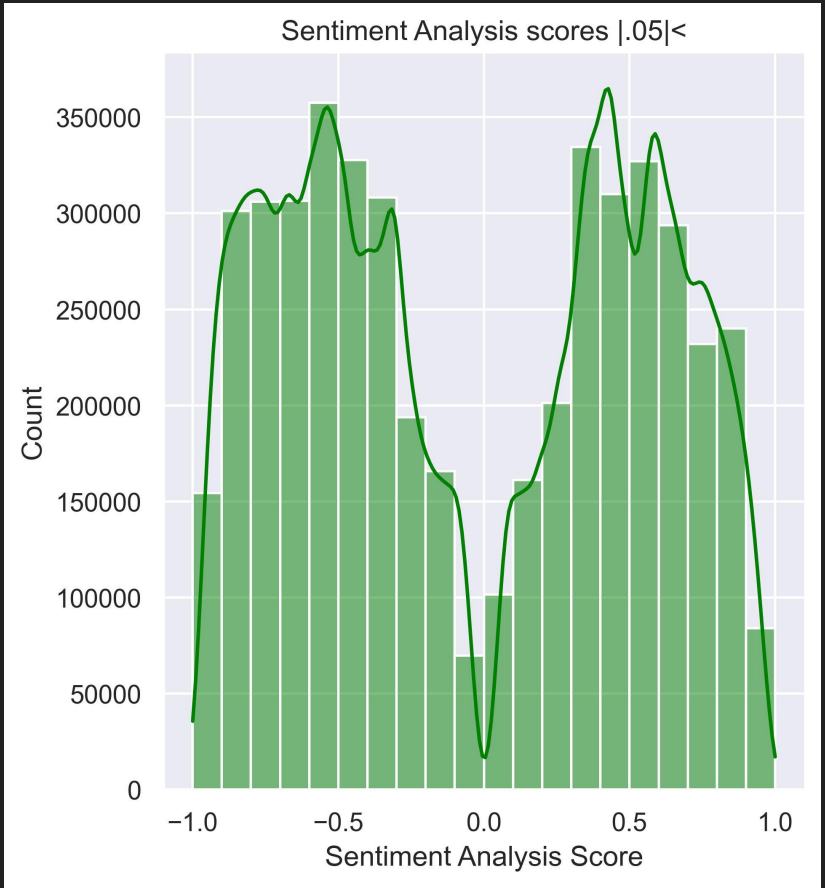


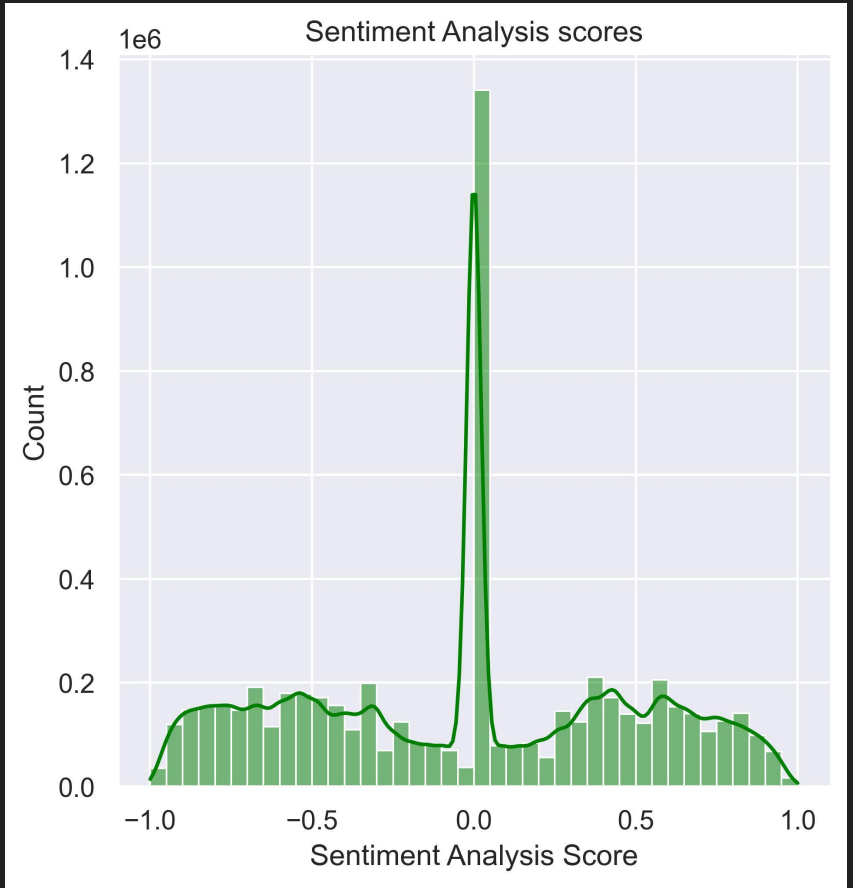
Sentiment Correlation Matrix



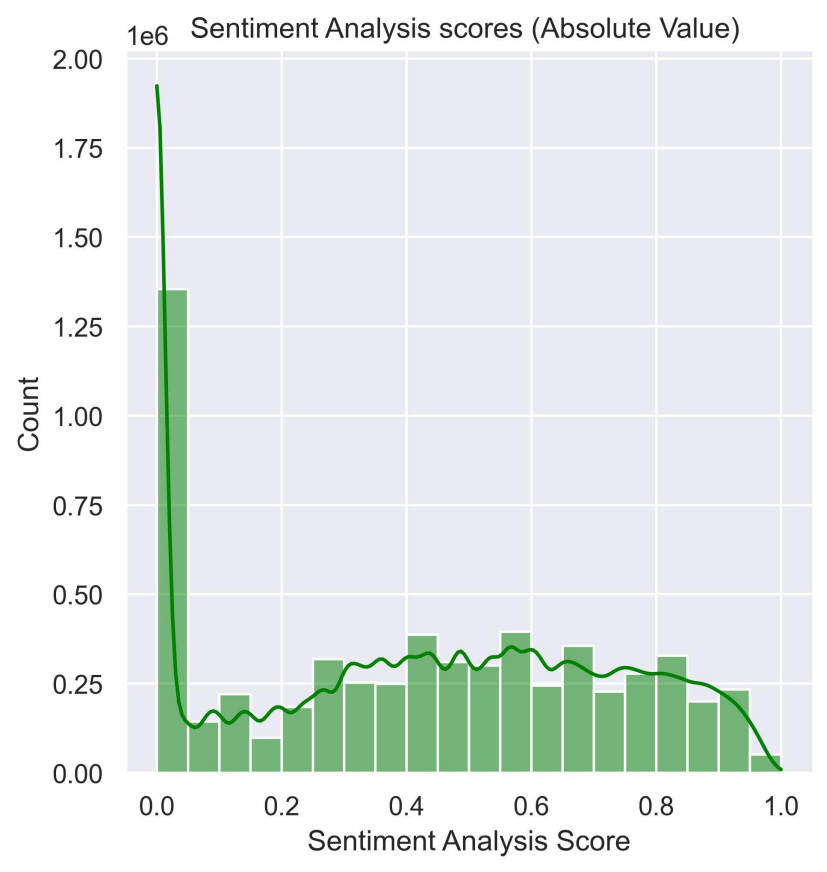
Sentiment Analysis Histograms

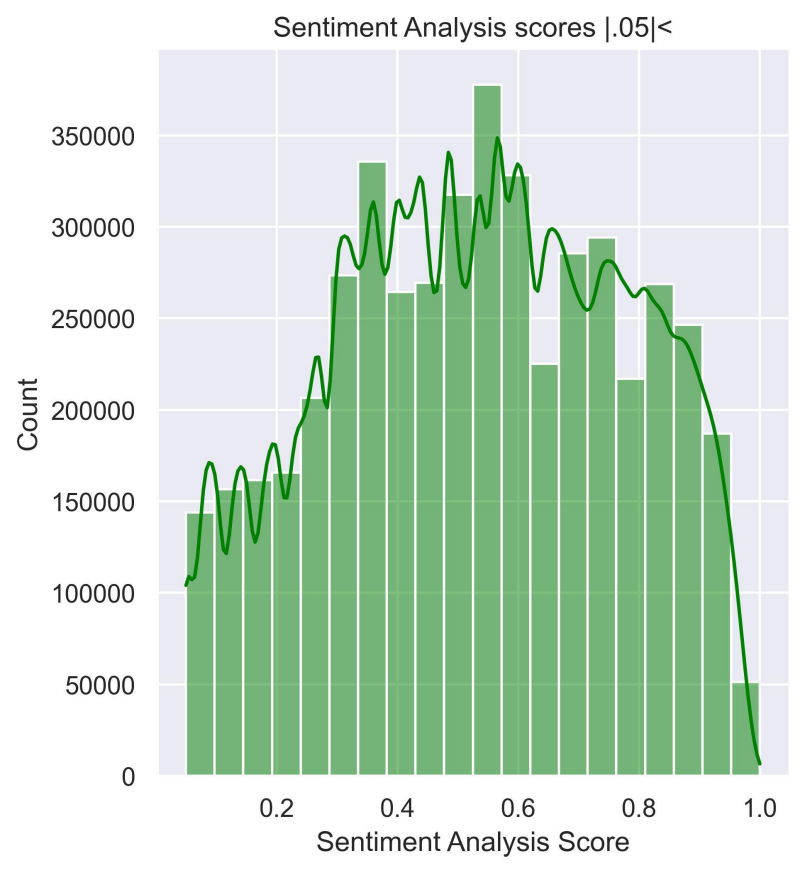
All Tweets


Sentiment Analysis scores |.05|<
Count
350000
300000
250000
200000
150000
100000
50000
0
−1.0
−0.5
0.0
0.5
1.0
Sentiment Analysis Score

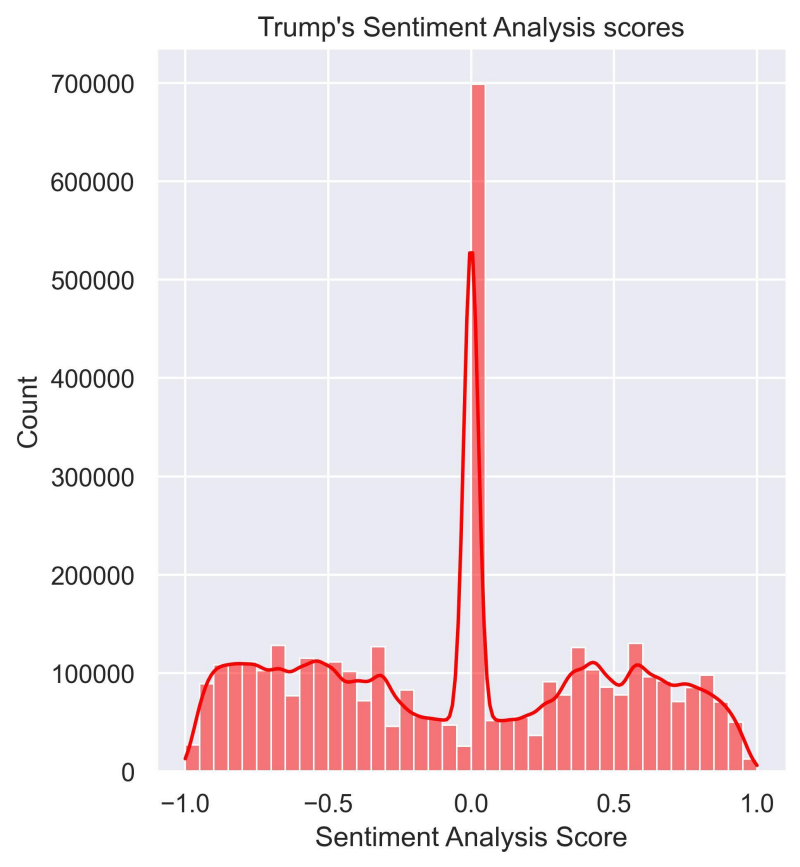

Sentiment Analysis scores
1e6
Count
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Sentiment Analysis Score

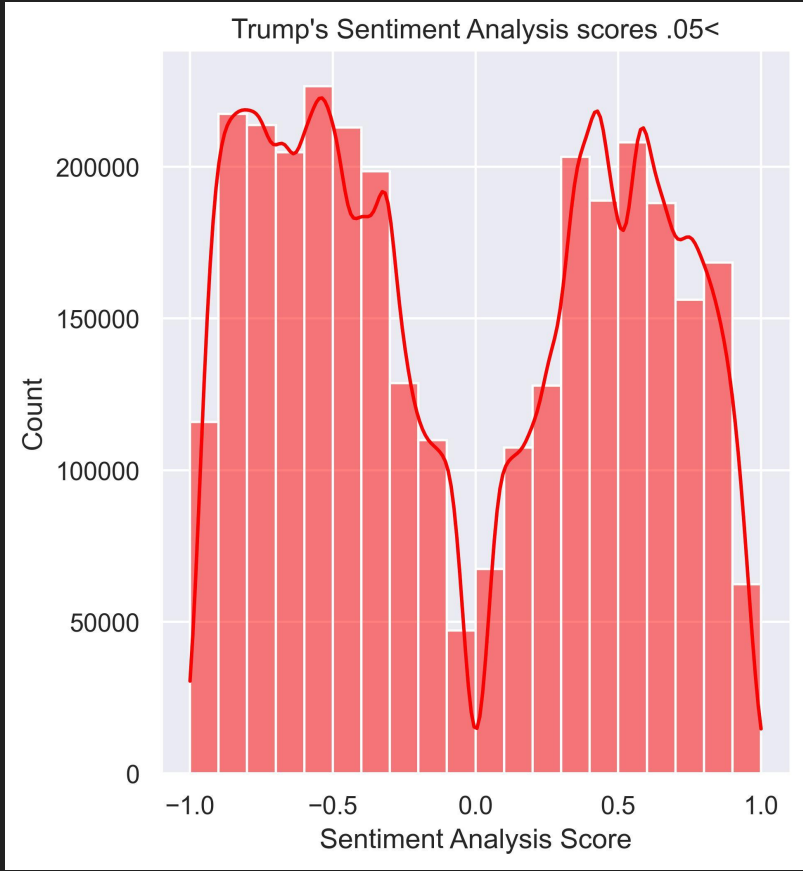
All Tweets (absolute Value)



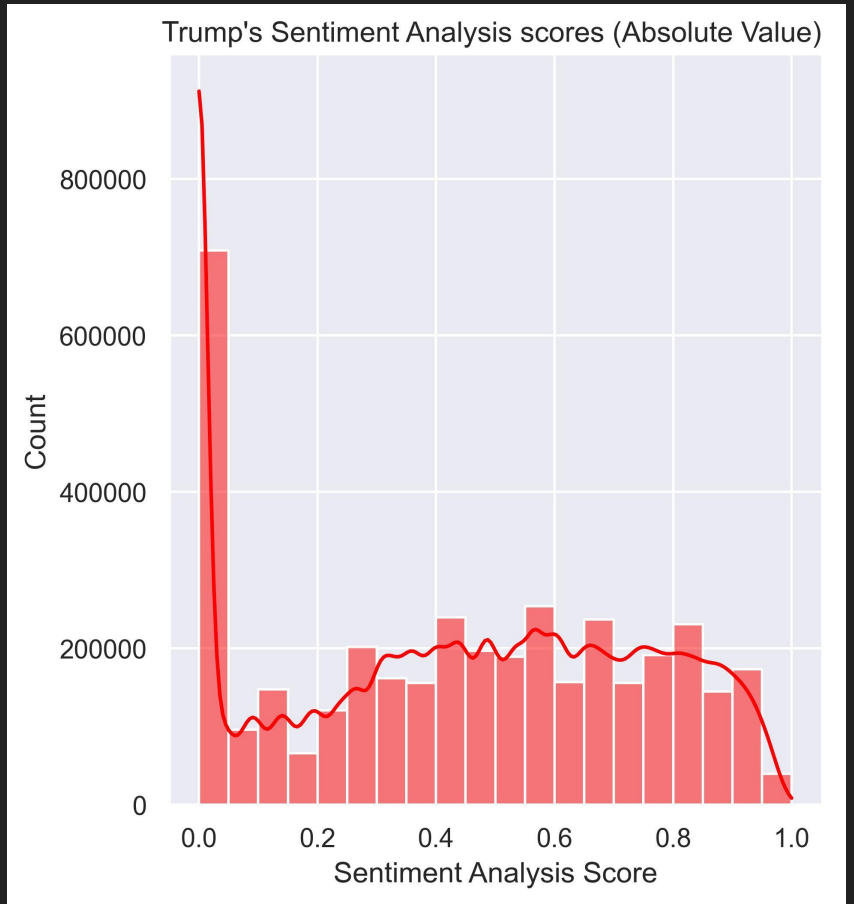


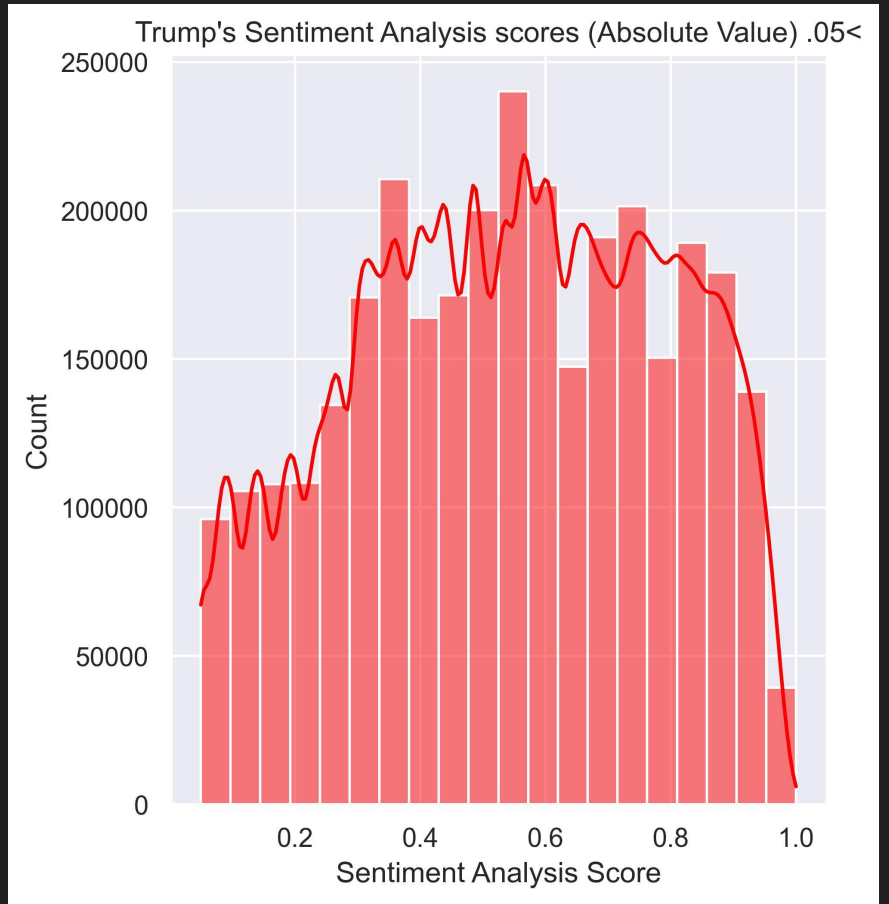
Trump Tweets



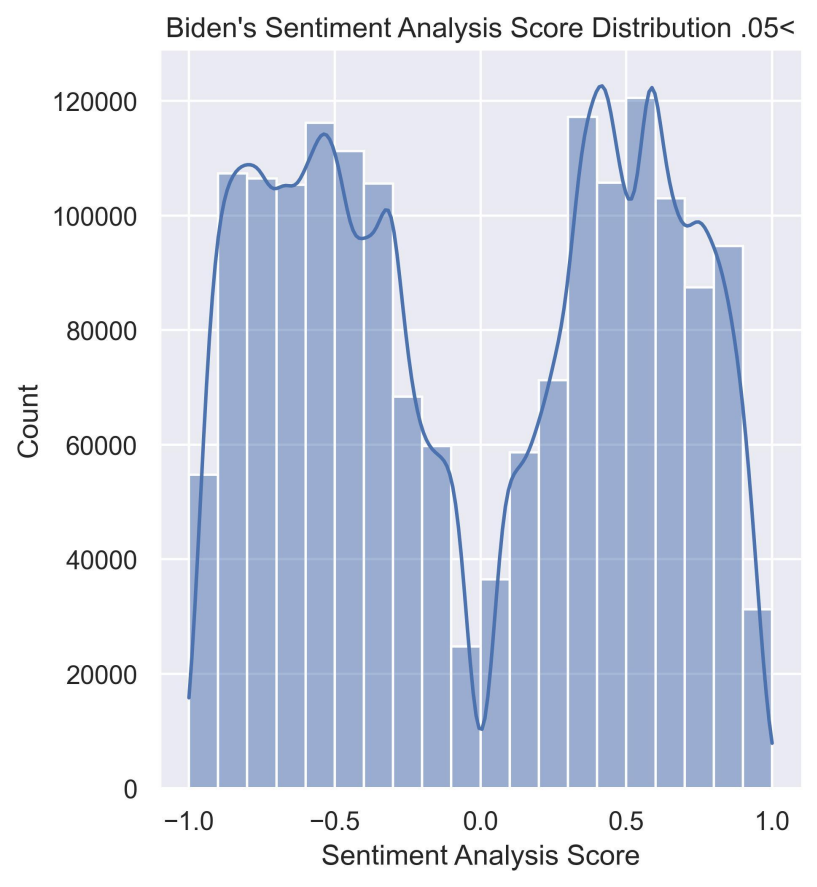


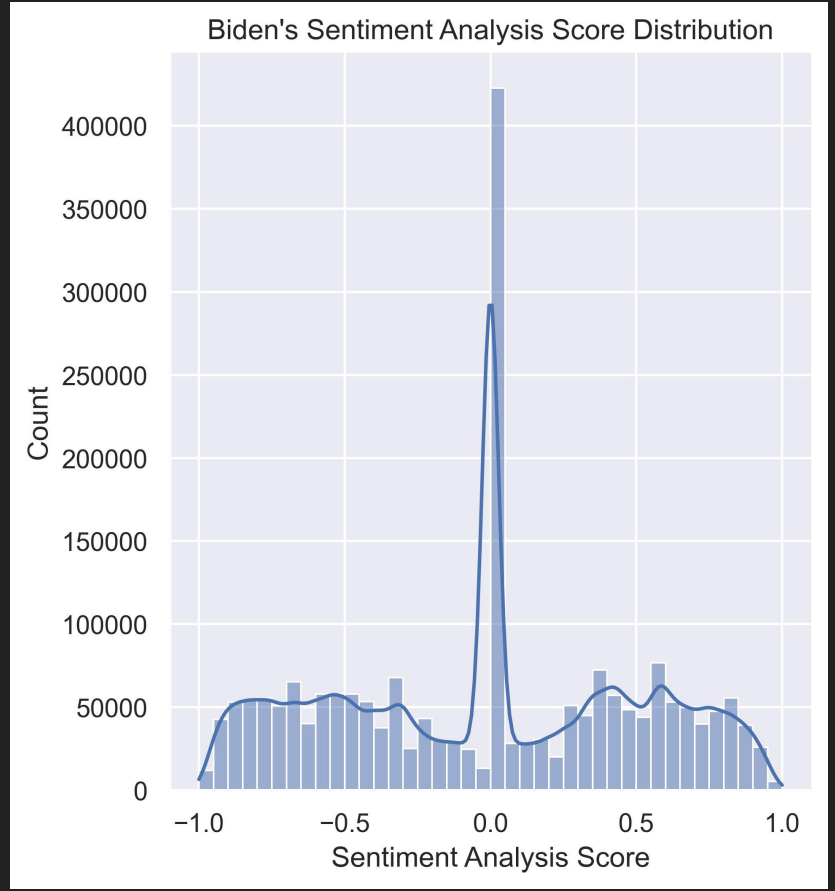
Trump Tweets (absolute Value)



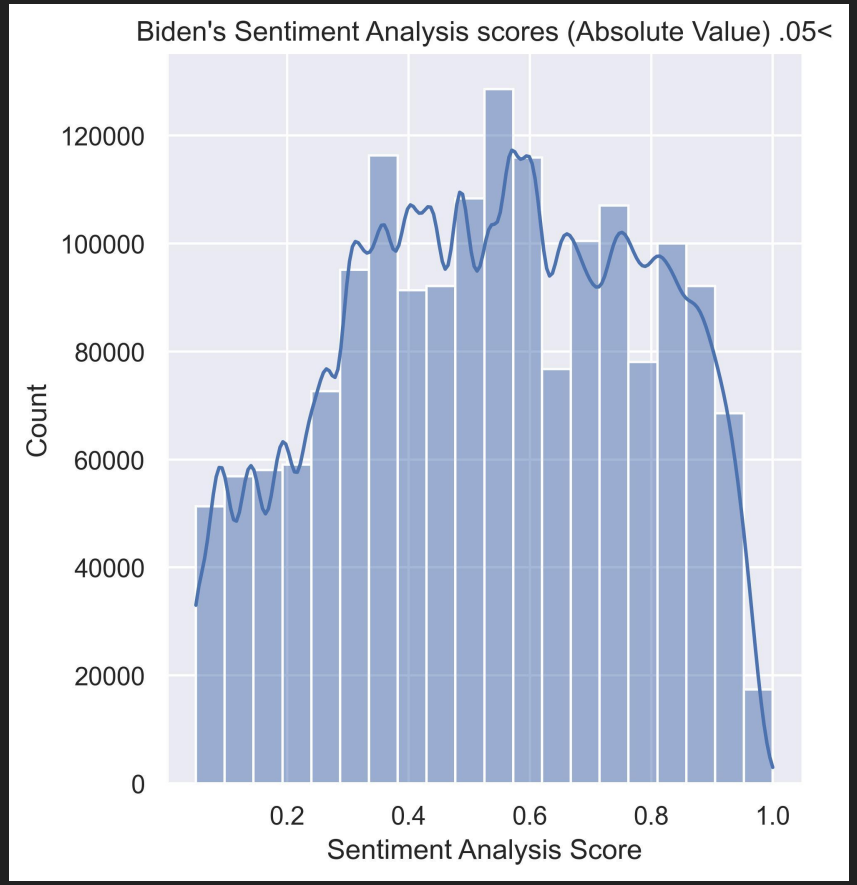


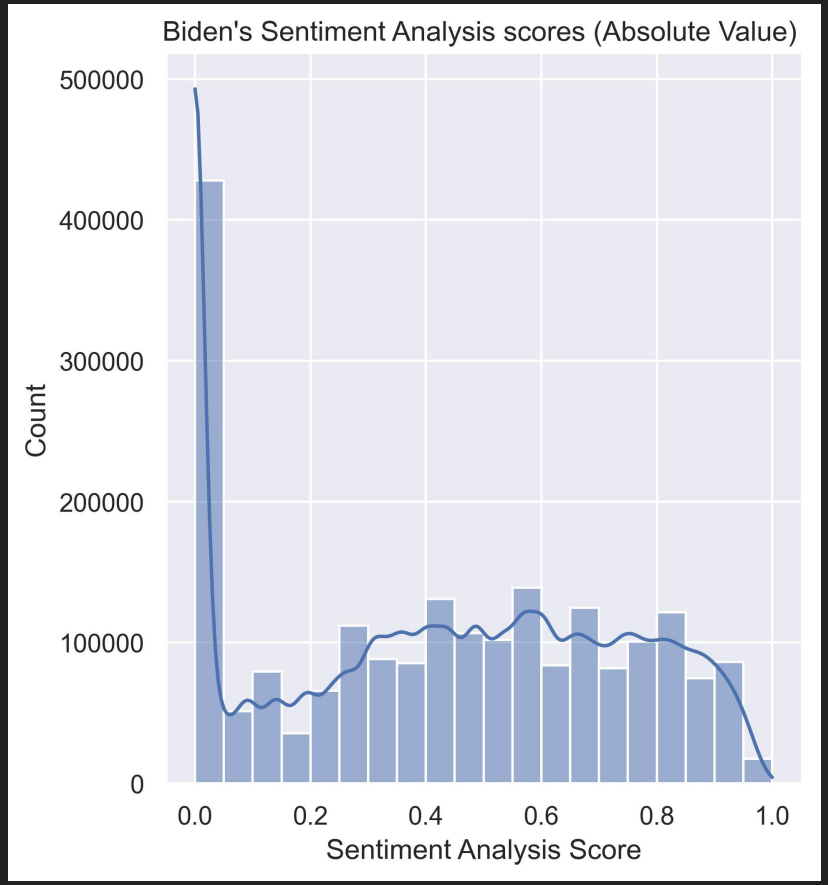
Biden Tweets



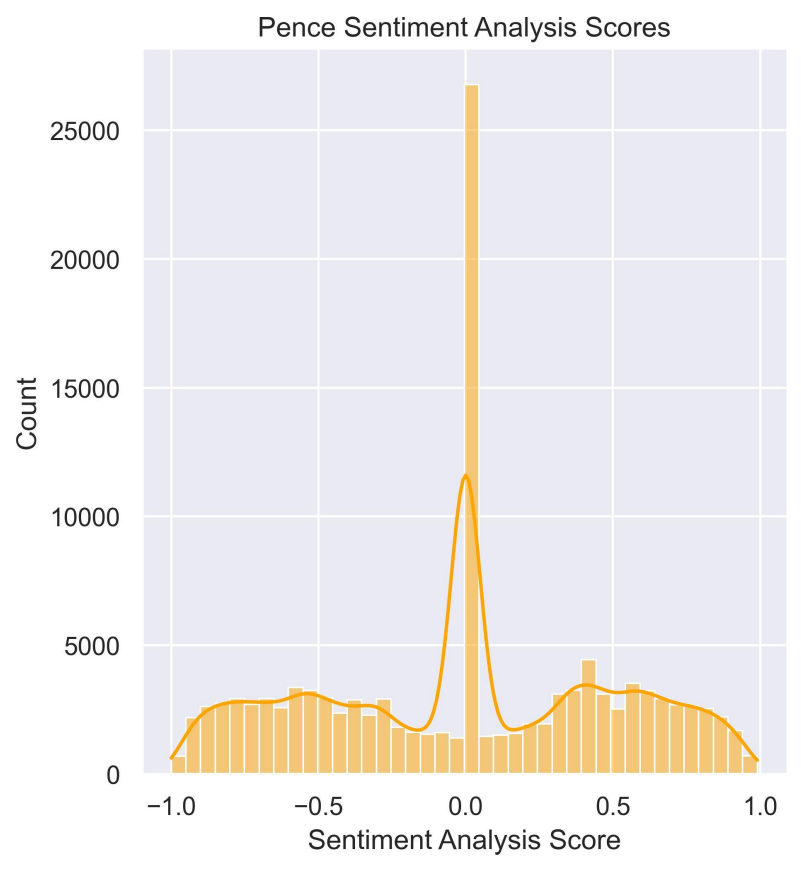


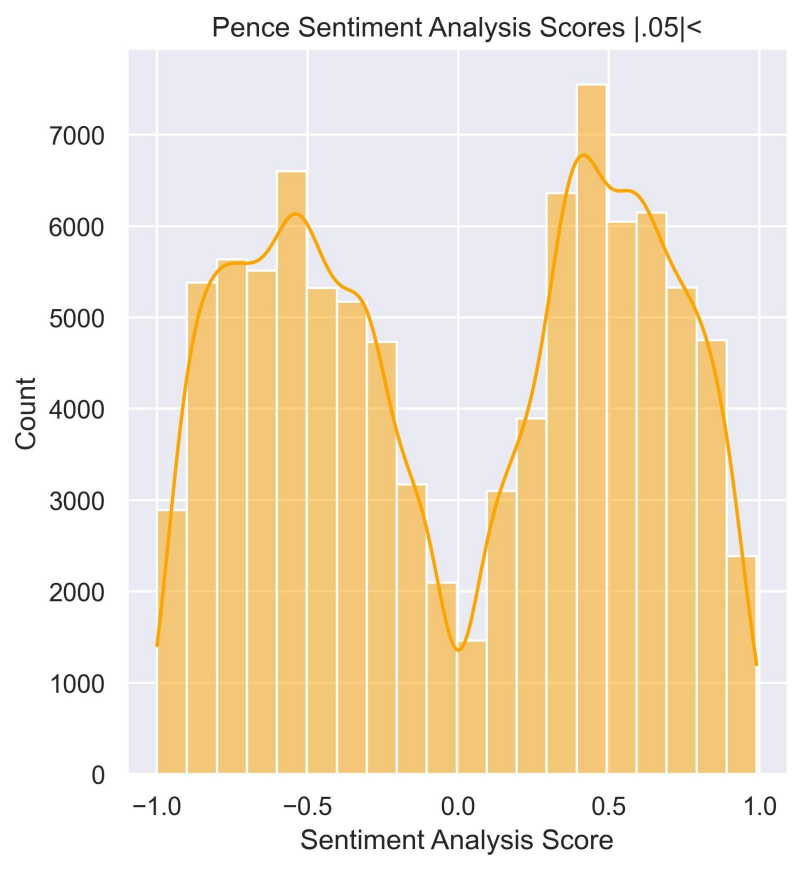
Biden Tweets (absolute Value)



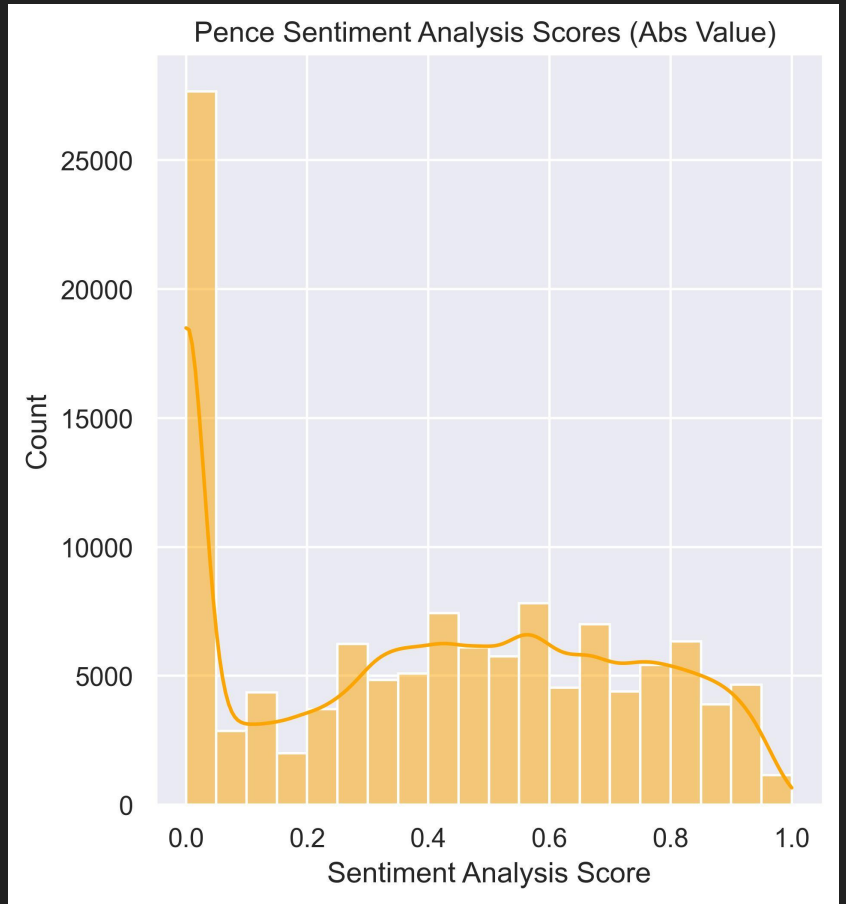


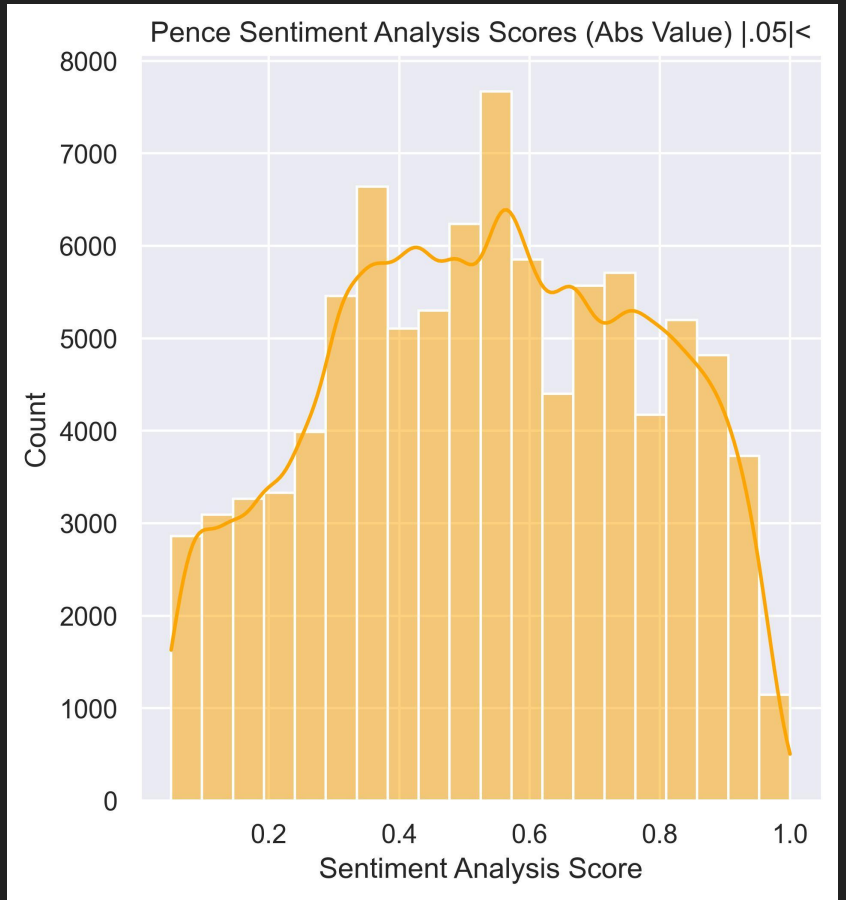
Pence Tweets



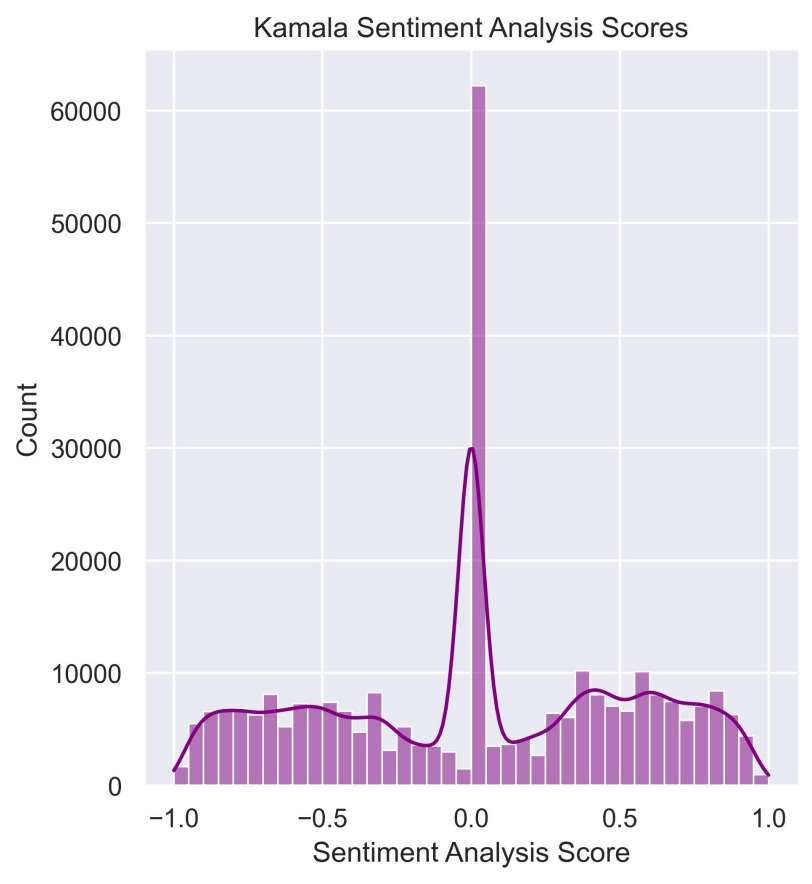


Pence Tweets (absolute Value)

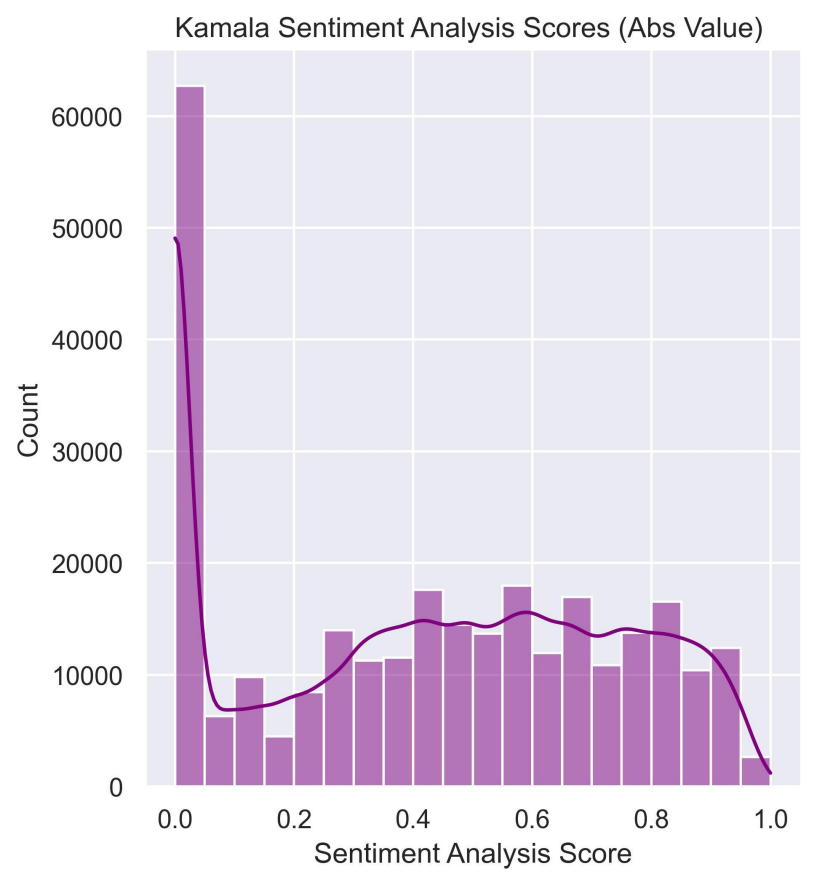


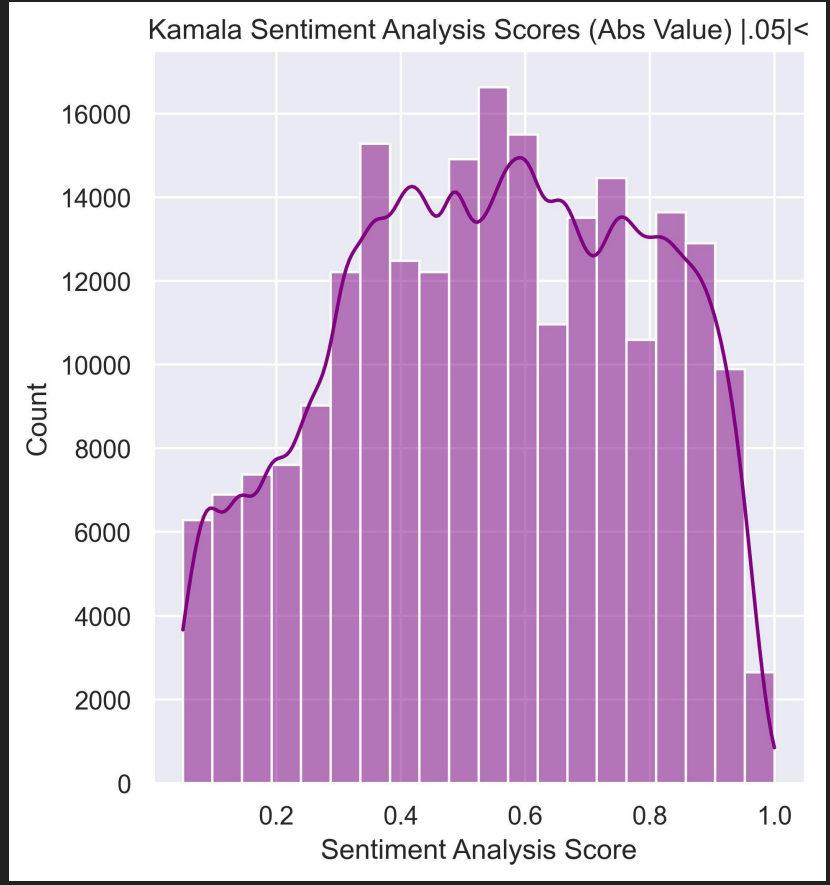


Kamala Tweets



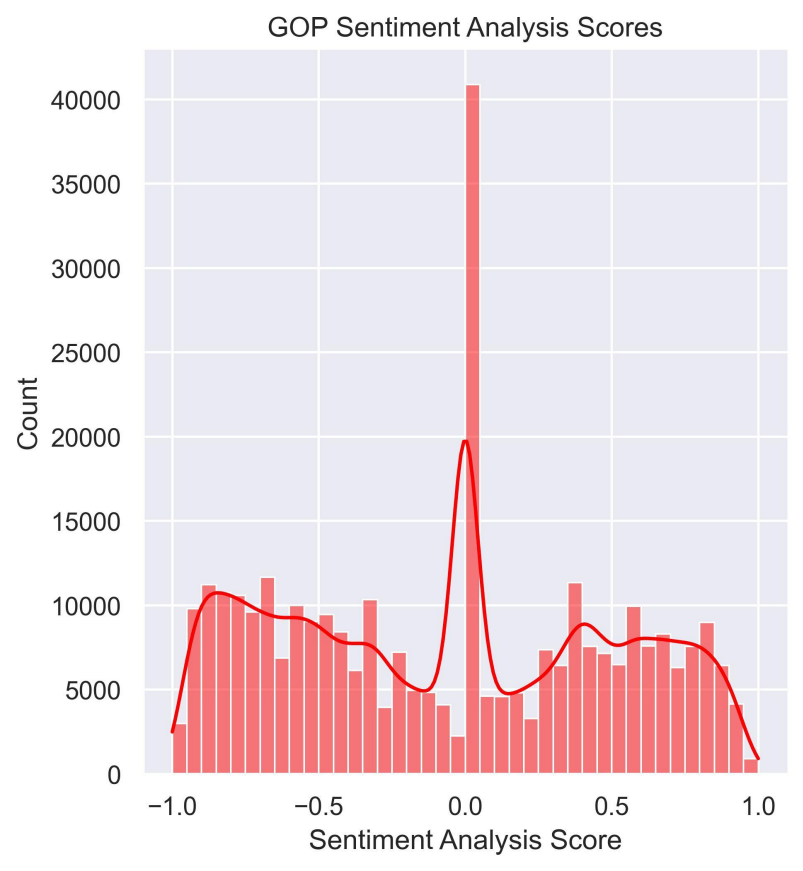
Kamala Tweets (absolute Value)



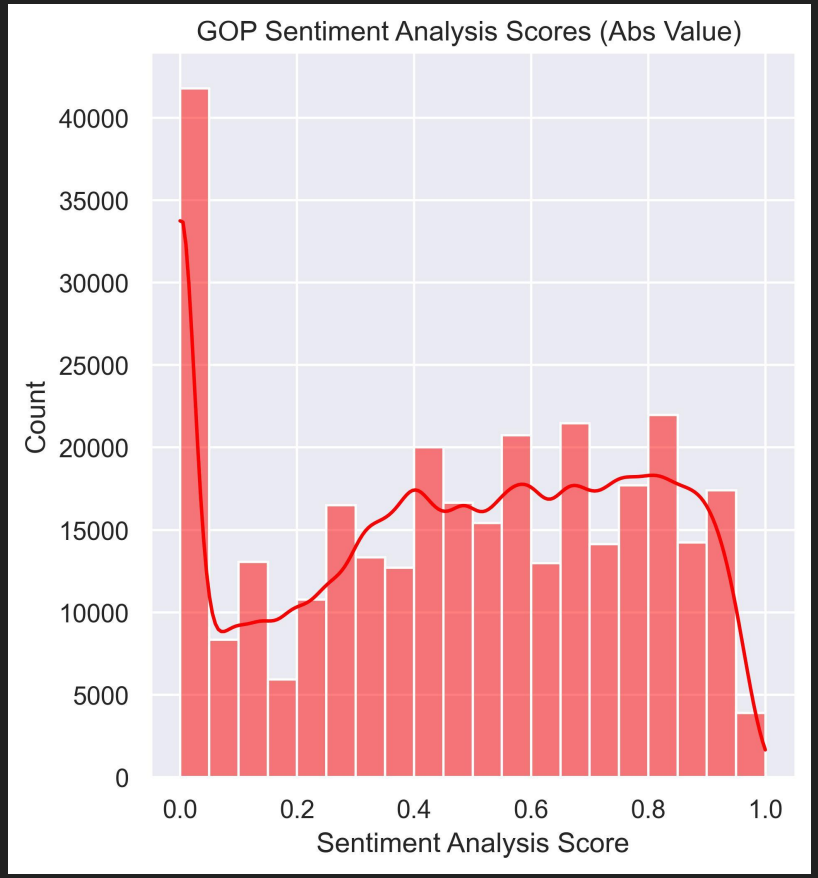


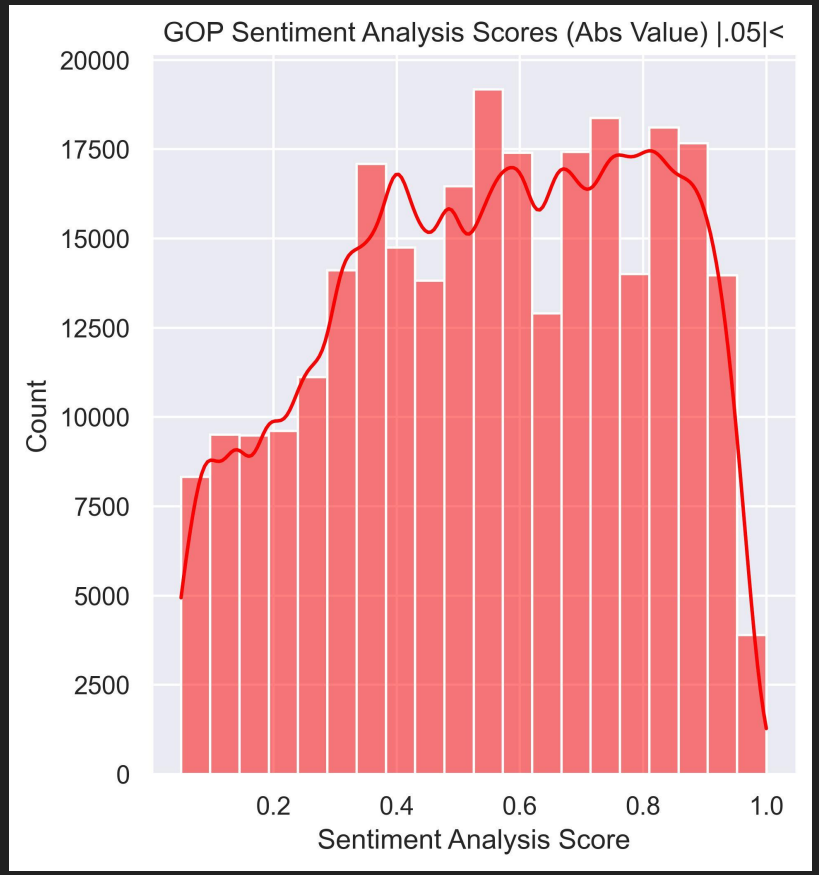
GOP Tweets


GOP Sentiment Analysis Scores |.05|<
Count
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0
Sentiment Analysis Score

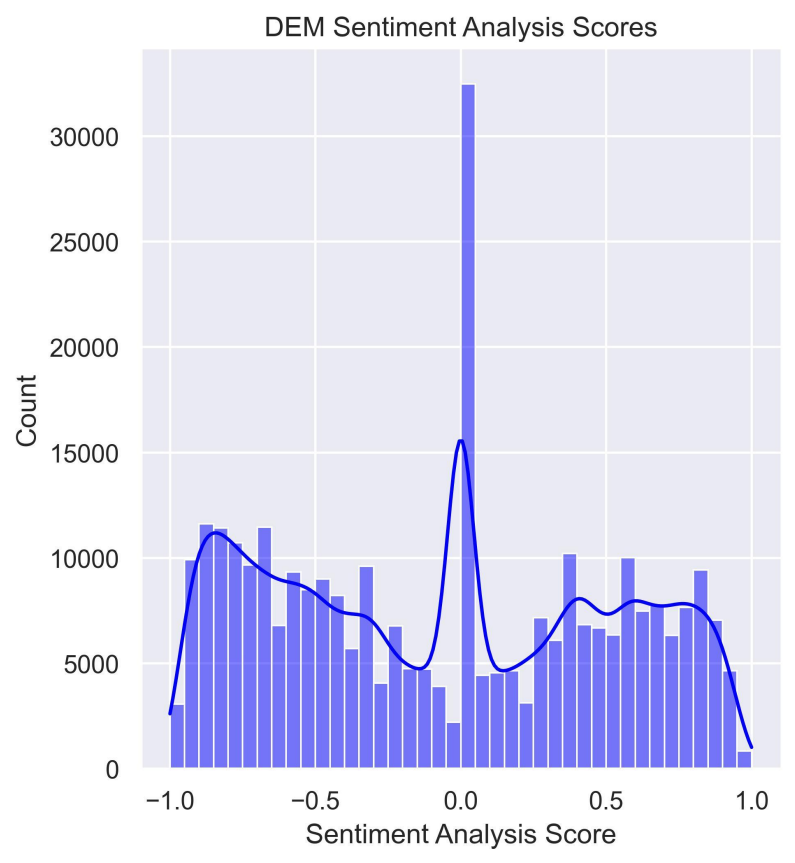

GOP Sentiment Analysis Scores
Count
40000
35000
30000
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0
Sentiment Analysis Score

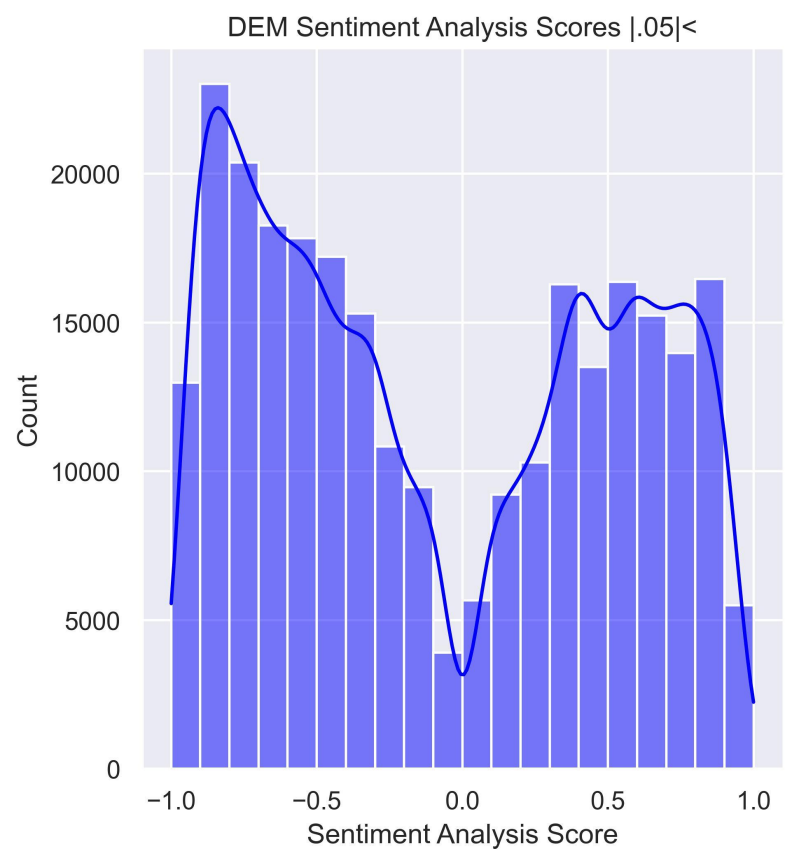
GOP Tweets (absolute Value)



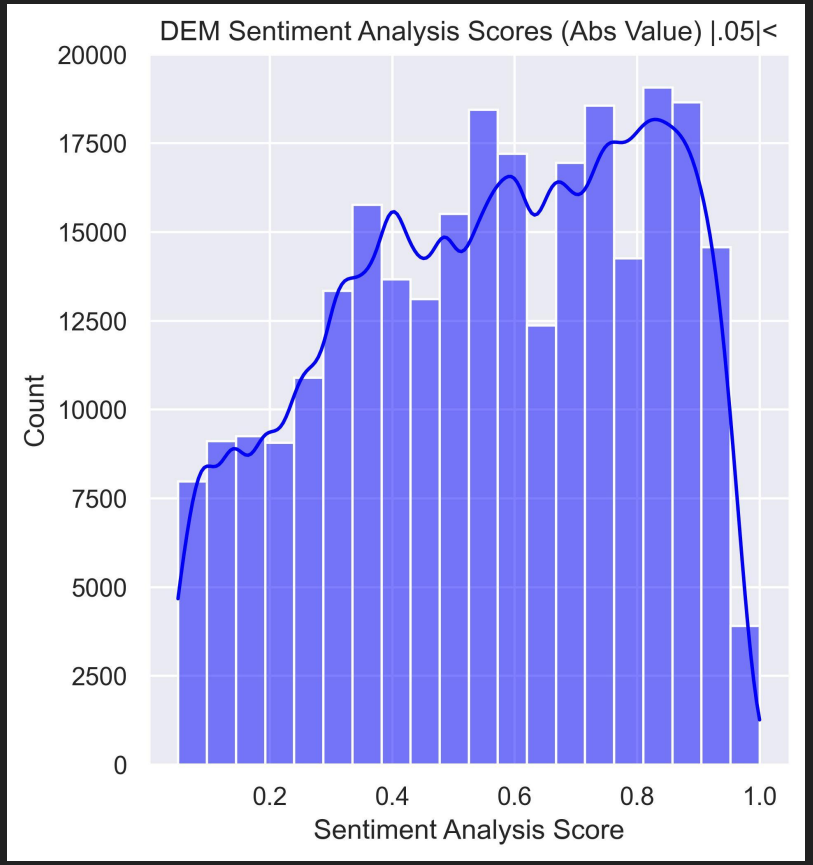


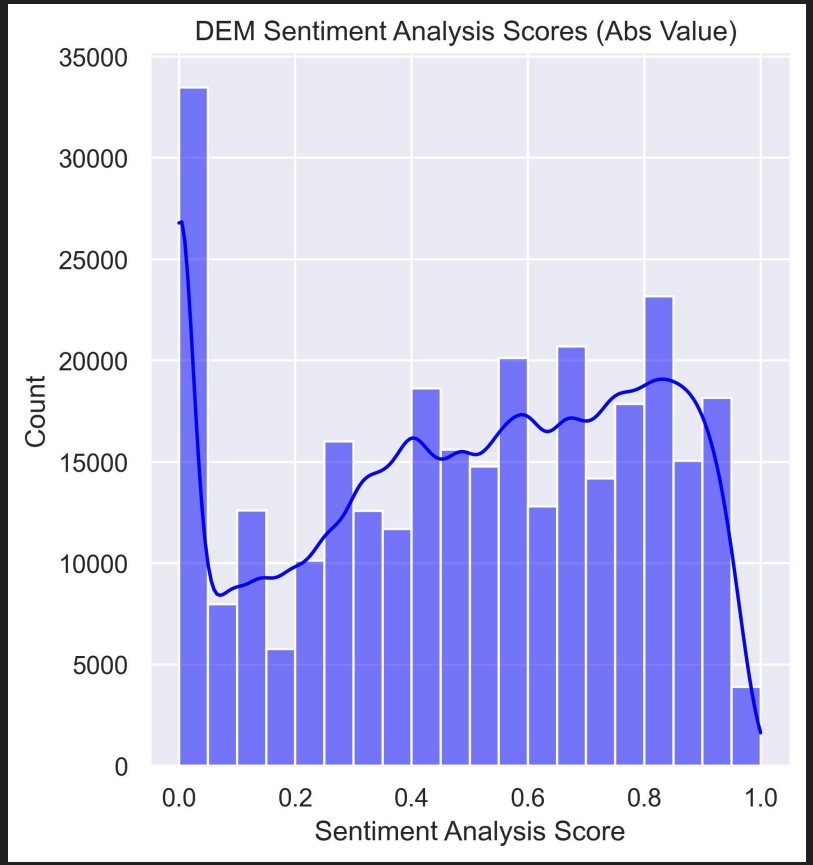
DEM Tweets


DEM Sentiment Analysis Scores
Count
30000
25000
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0
Sentiment Analysis Score


DEM Sentiment Analysis Scores |.05|<
Count
20000
15000
10000
5000
0
−1.0
−0.5
0.0
0.5
1.0
Sentiment Analysis Score

DEM Tweets (absolute Value)



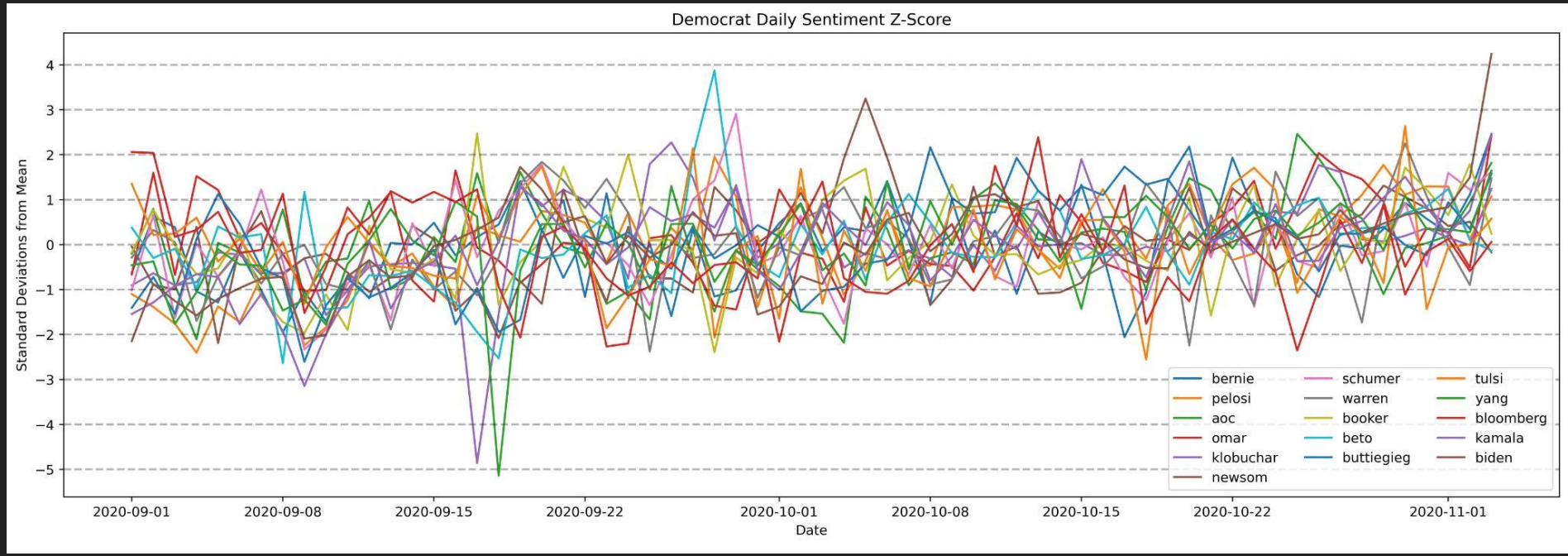


Sentiment Analysis Volatility

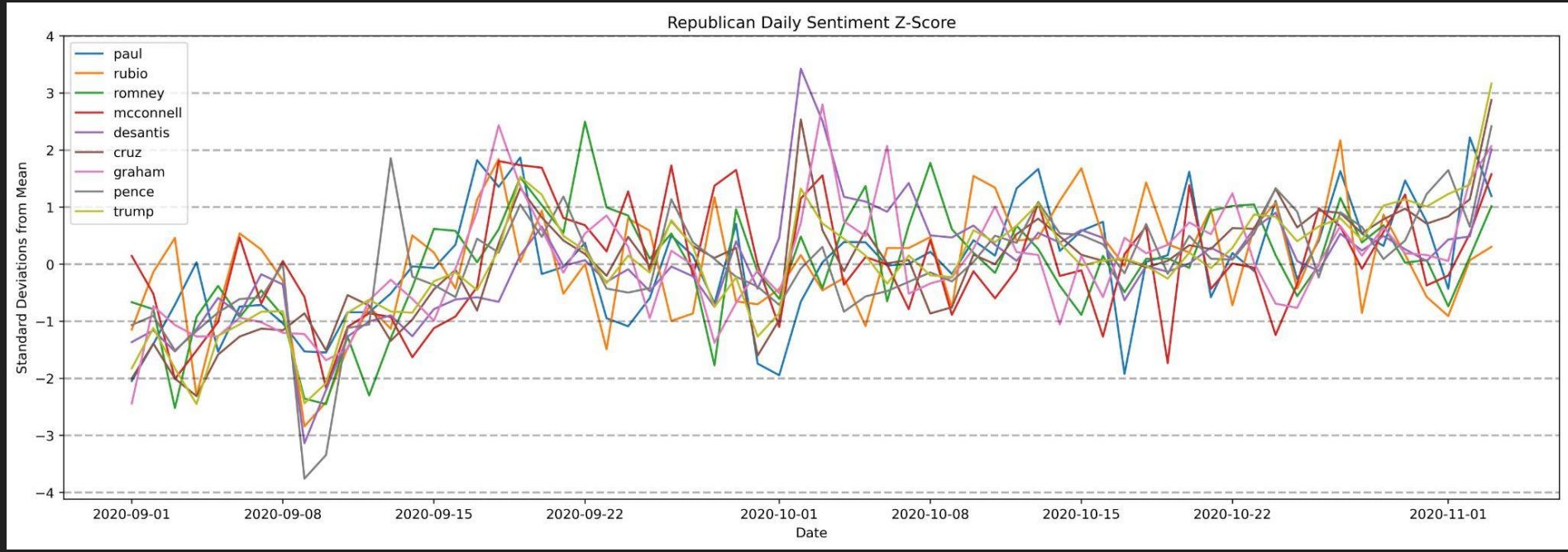
% Change

Large spikes caused by values approaching 0.00

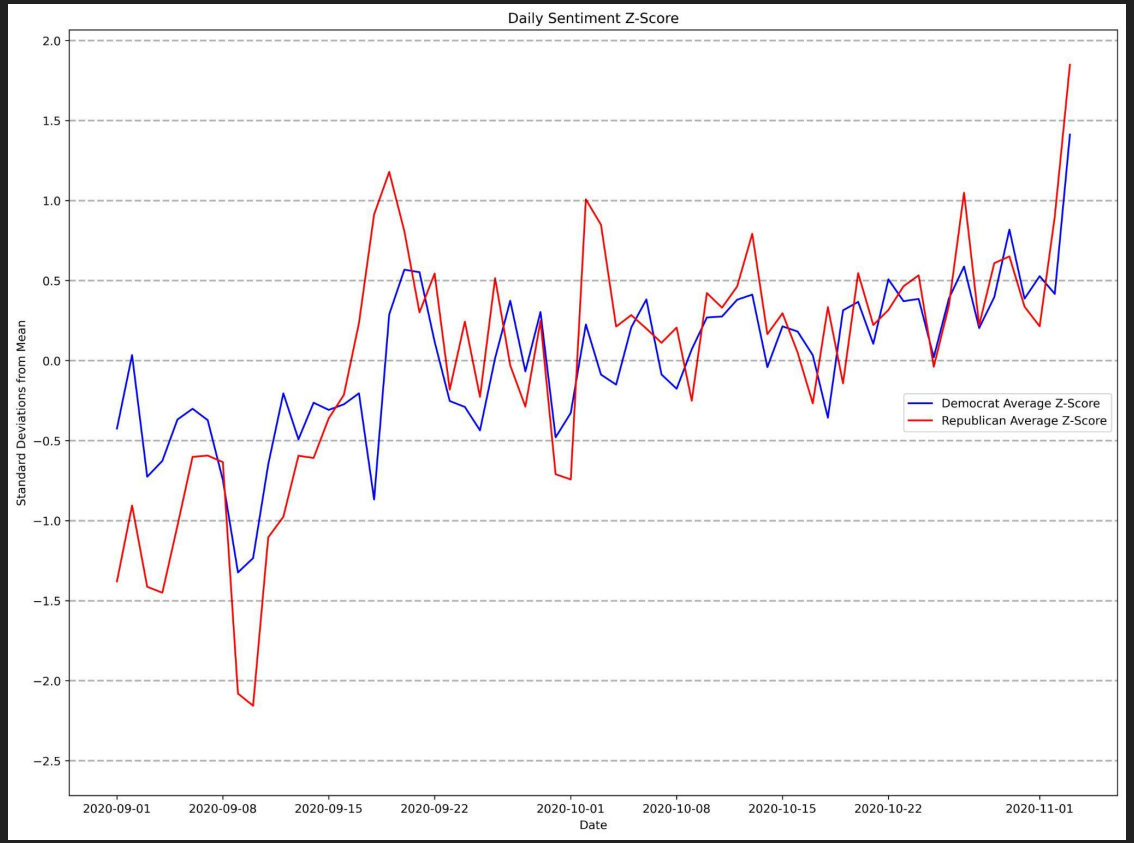
Z-Score Democrats



Z-Score Republicans

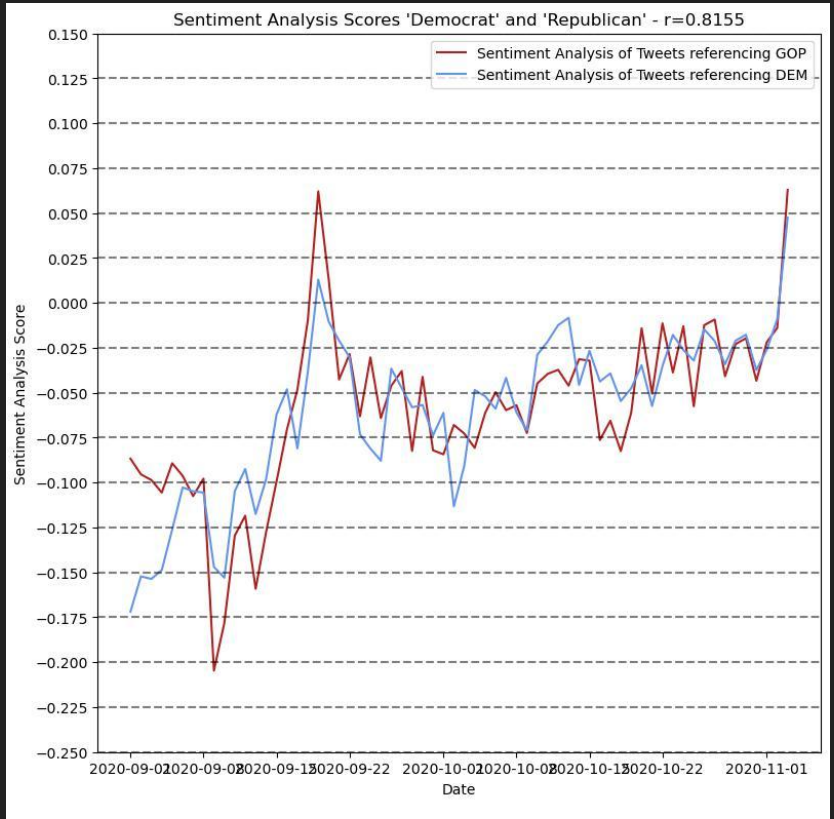


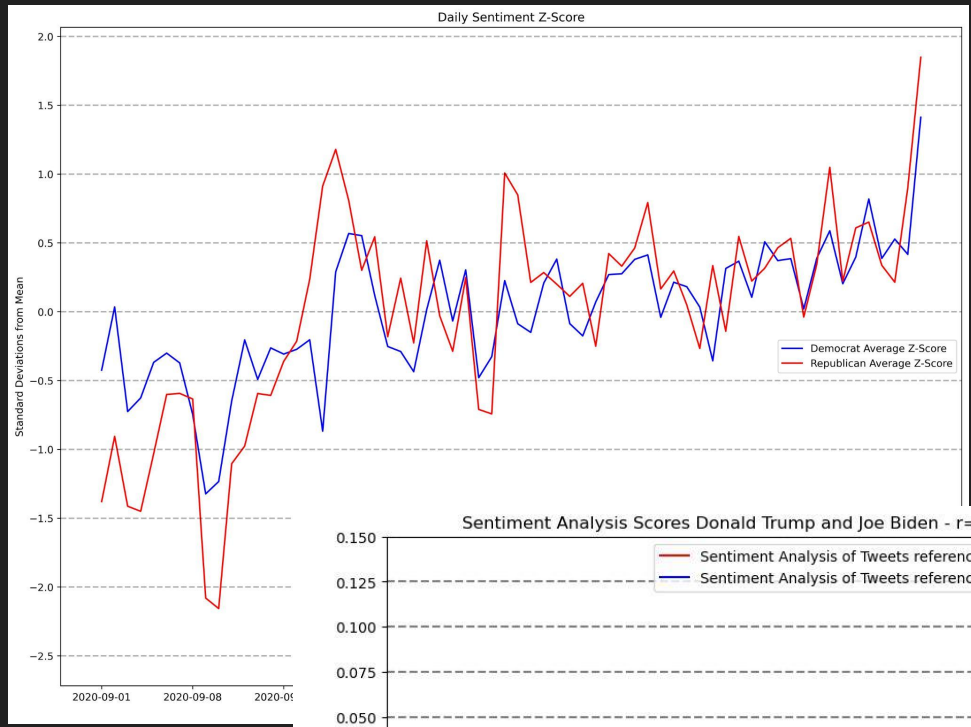
Combined average Z-Score

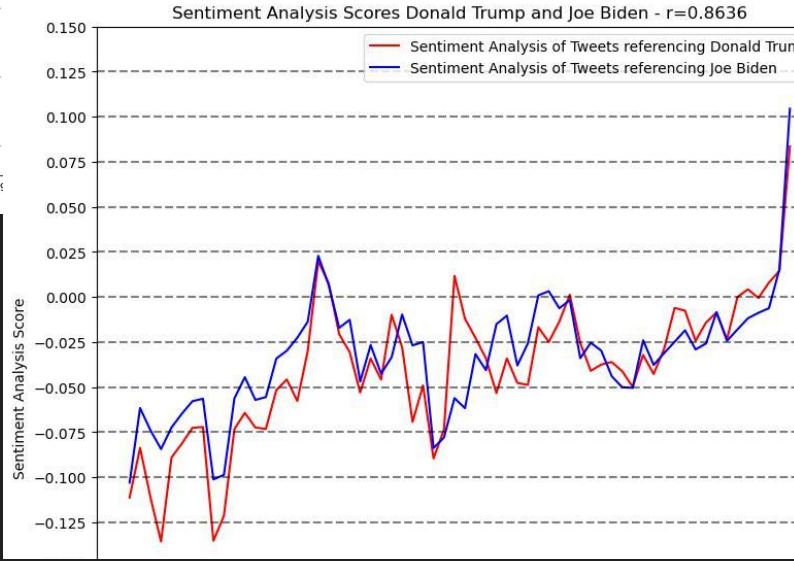


similar shape to other sentiment analysis graphs

Z-score, Party VADER, & Trump/Biden VADER





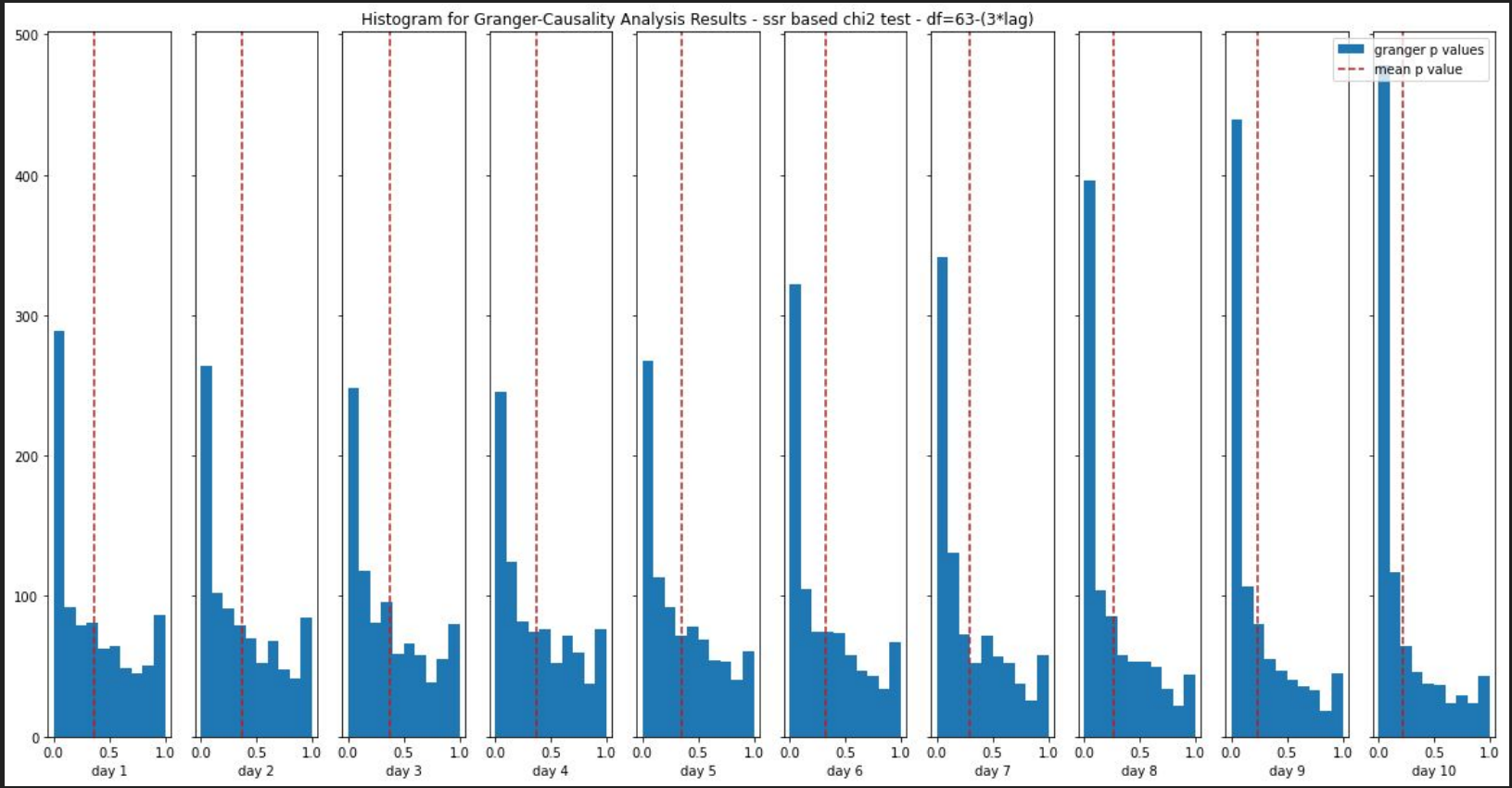


Granger Causality Results

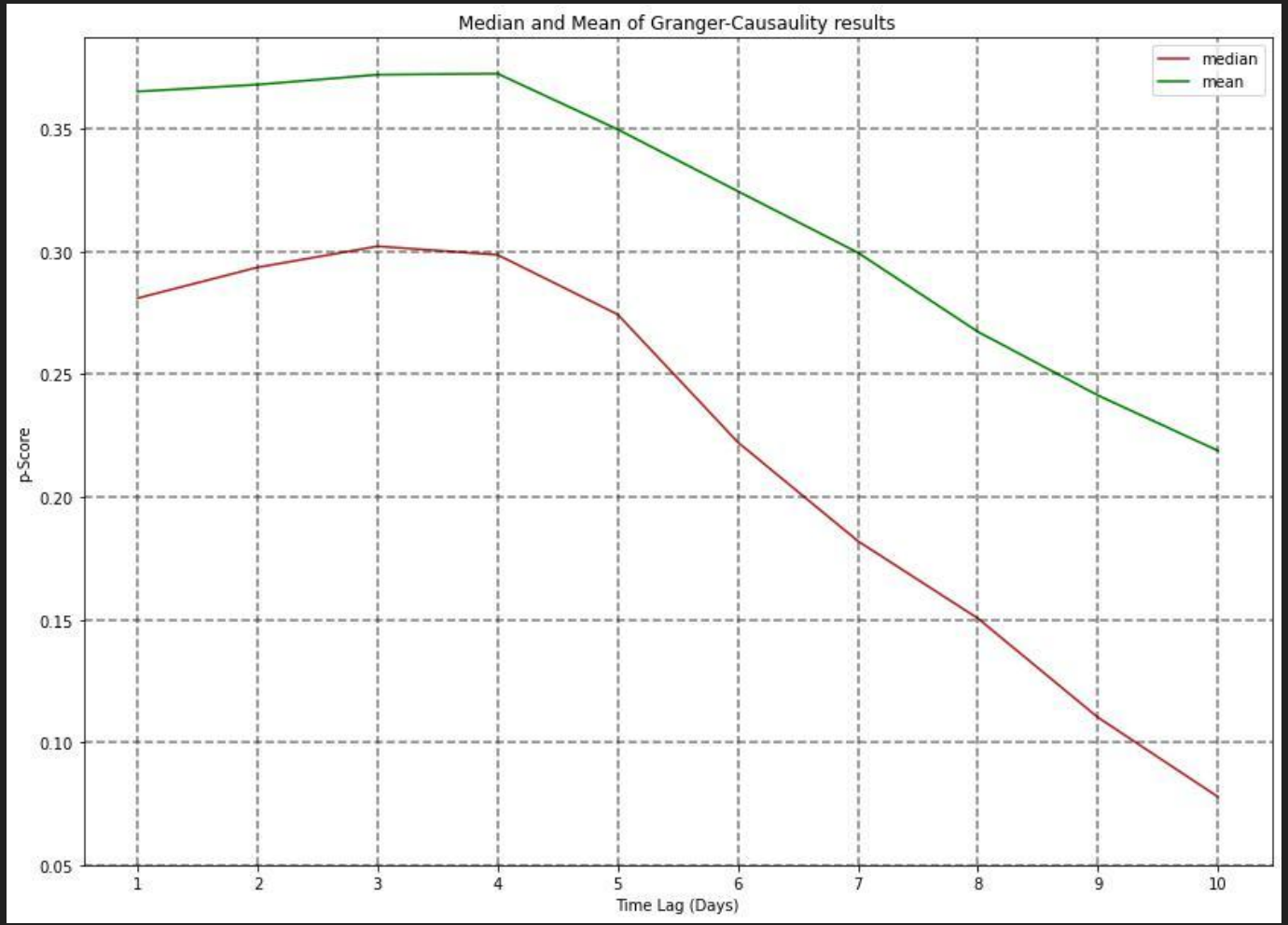
Granger Causality test - 10 days

item 1	item 2	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9	Day 10
Biden Polling	Trump Sentiment	0.6354	***0.0084***	**0.0344**	*0.0739*	0.1466	0.1482	0.183	0.1353	*0.0571*	***0.0077***
Biden Polling	Biden Sentiment	0.4918	**0.0479**	0.1701	**0.0484**	**0.0246**	**0.0155**	**0.0199**	**0.0201**	**0.0162**	***0.0008***
Biden Polling	Trump Polling	***0.0062***	***0.006***	**0.0103**	*0.0726*	*0.0624*	**0.0113**	**0.0179**	**0.0107**	***0.0012***	***0.0032***
Biden Sentiment	Trump Sentiment	0.162	0.2835	0.4718	0.6476	0.5696	0.3126	0.3567	*0.0752*	0.1396	**0.0212**
Biden Sentiment	Biden Polling	0.3402	0.5128	0.8481	0.6722	0.7863	0.2791	0.3552	0.3331	0.6385	*0.0512*
Biden Sentiment	Trump Polling	0.1829	0.2789	0.3244	0.4362	0.6171	*0.0577*	0.1602	0.326	0.4228	*0.0596*
Trump Polling	Biden Sentiment	***0.0002***	***0.0***	***0.0***	***0.0001***	***0.0***	***0.0***	***0.0***	***0.0***	***0.0001***	***0.0***
Trump Polling	Trump Sentiment	***0.0012***	***0.0044***	**0.027**	**0.0397**	**0.0202**	**0.0326**	**0.0333**	*0.0631*	*0.0844*	***0.0007***
Trump Polling	Biden Polling	**0.0321**	0.2876	0.3188	0.2005	0.468	0.5888	0.1559	0.1671	**0.0423**	**0.0256**
Trump Sentiment	Biden Polling	0.1439	0.2774	0.6335	0.7919	0.5839	0.4714	0.719	0.1385	0.2388	*0.0893*
Trump Sentiment	Trump Polling	0.1416	*0.0871*	*0.0847*	*0.0732*	0.3829	0.1451	0.3162	0.2948	0.1028	*0.0874*
Trump Sentiment	Biden Sentiment	0.9037	0.8442	0.5549	0.7891	0.7615	0.6516	0.4672	0.1258	**0.0343**	***0.0044***

Distribution of Granger Causality Results - All Politicians

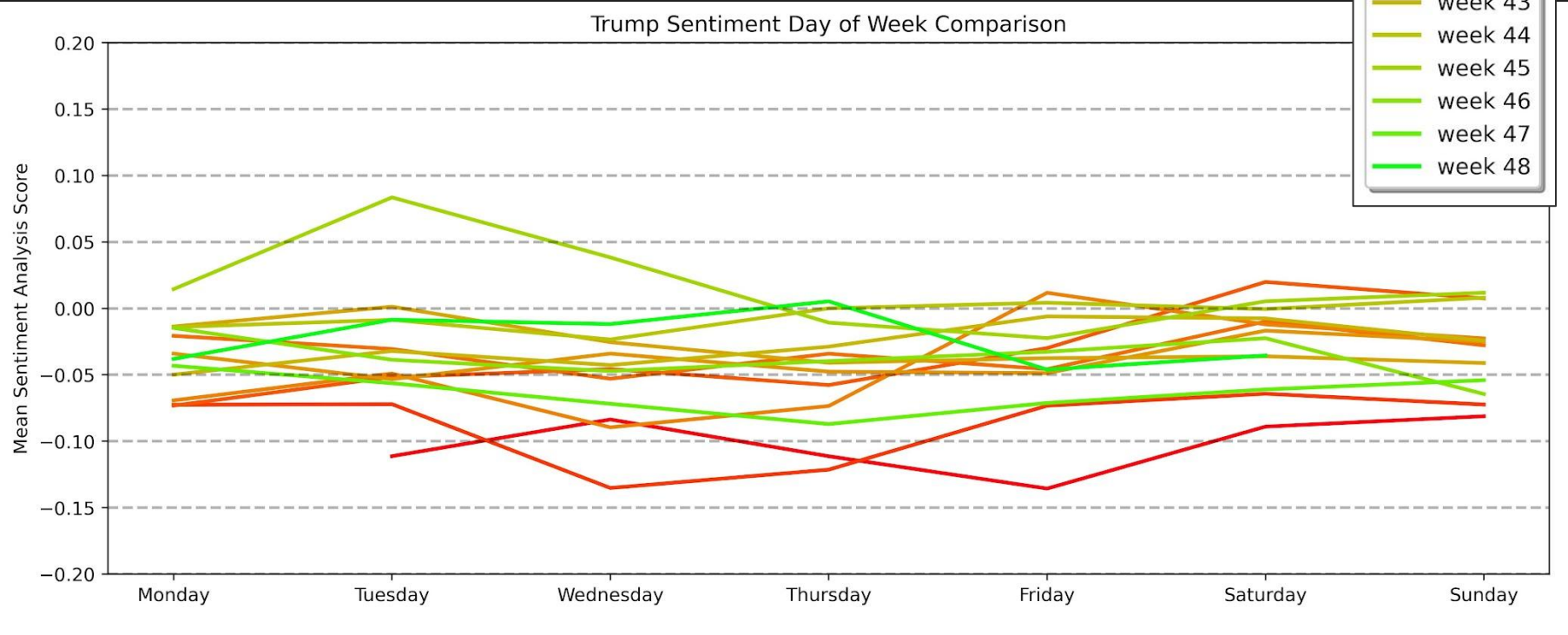


Granger Causality Results - All Politicians

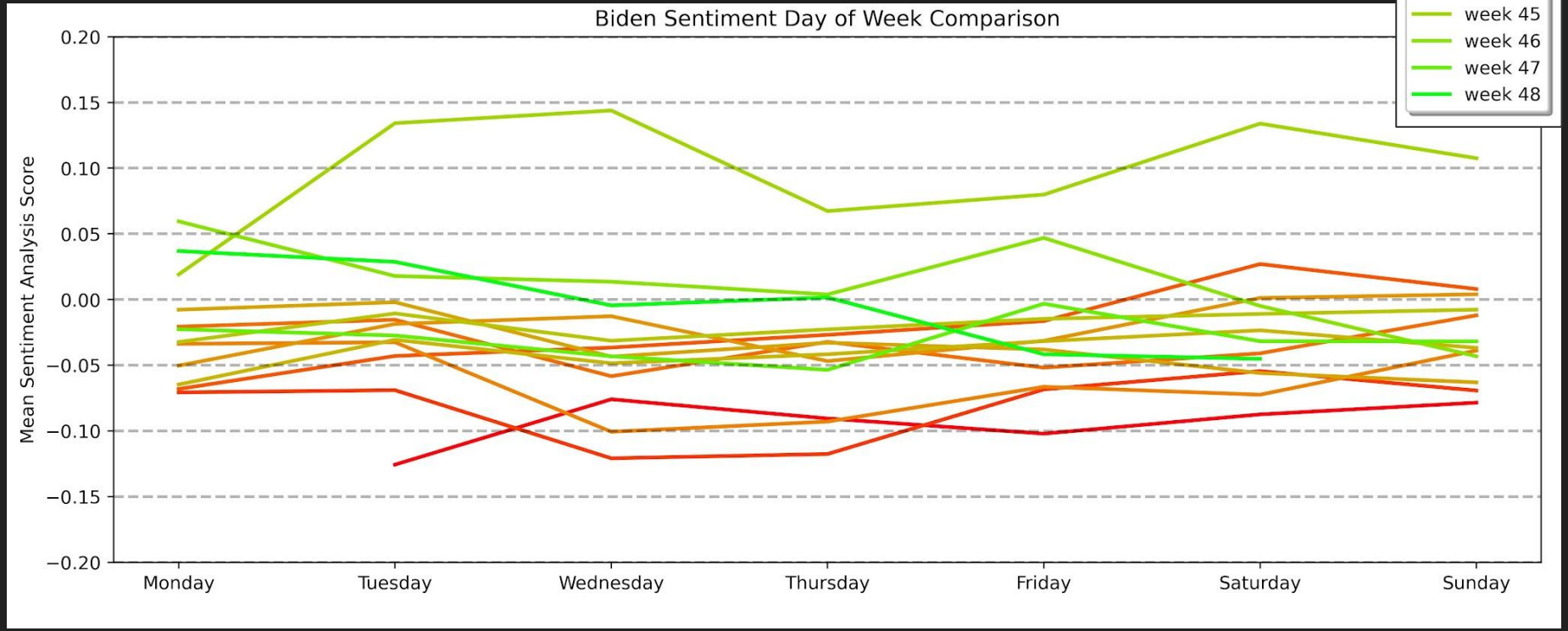


Day of week / week-to-week analysis

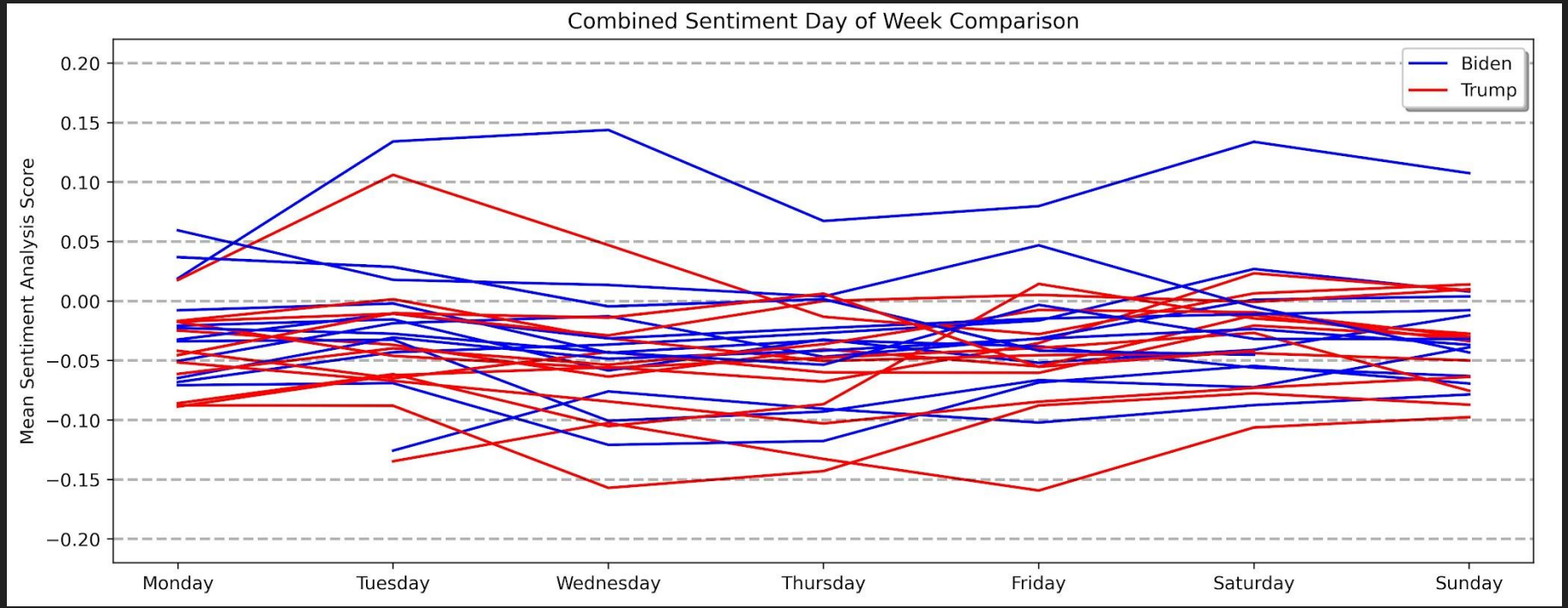
Trump - Day of Week Comparison



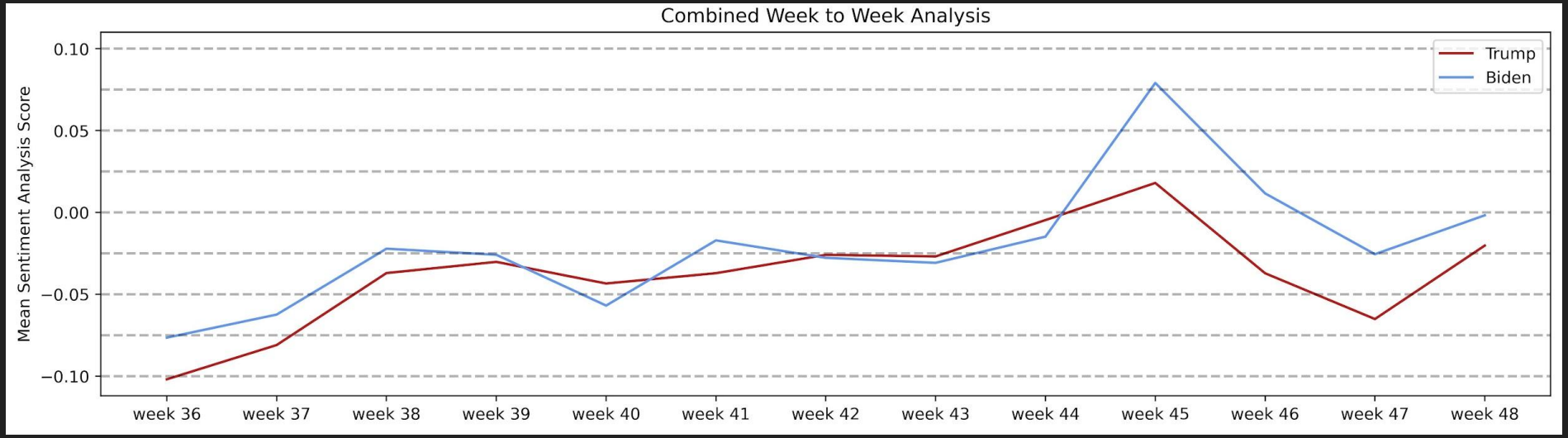
Biden - Day of Week Comparison



Combined Day of Week Comparison



Combined Week-to-Week Comparison



Regression Analysis

Simple Regression

Sentiment Score (Democrats) = -0.675 + 0.0014*(days)

Sentiment Score (Republicans) = -0.990 + 0.0015*(days)

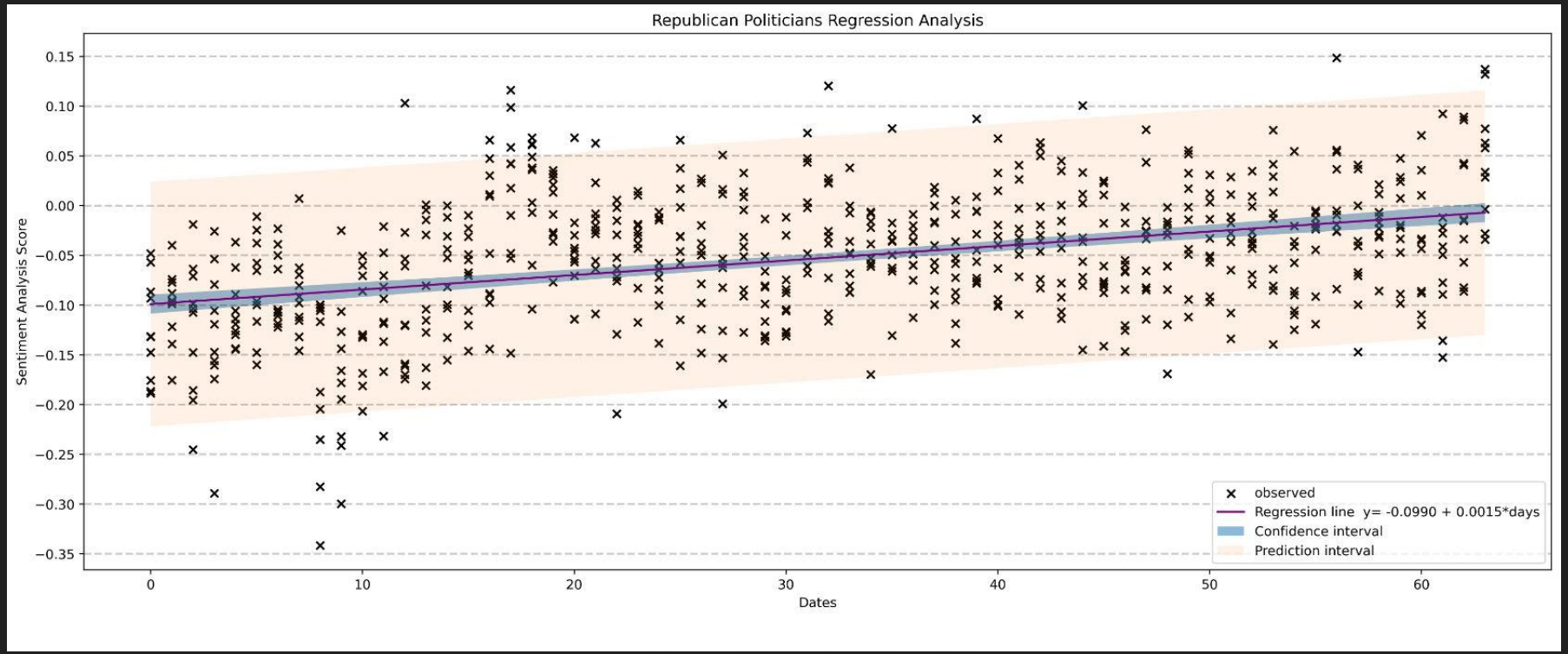
Without Outliers:

Sentiment Score (Democrats) = -0.633+ 0.0013*(days)

Sentiment Score (Republicans) = -0.926 + 0.0013*(days)

The variable days is the number of days from September 1st (0 -> 63)

GOP Politicians Sentiment Regression



GOP Politicians Regression Summary

OLS Regression Results

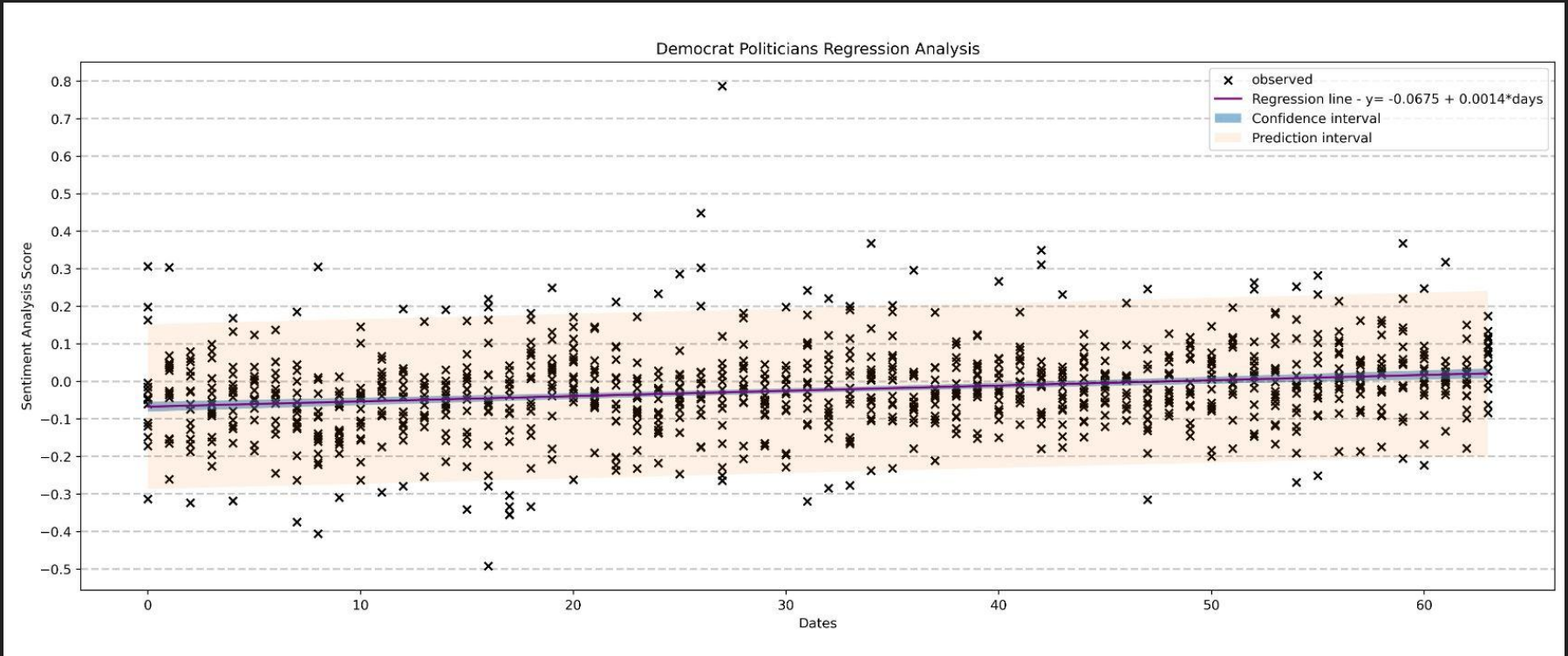
Dep. Variable:	points	R-squared:	0.158
Model:	OLS	Adj. R-squared:	0.156
Method:	Least Squares	F-statistic:	119.3
Date:	Mon, 06 Mar 2023	Prob (F-statistic):	1.41e-25
Time:	09:04:34	Log-Likelihood:	868.15
No. Observations:	640	AIC:	-1732.
Df Residuals:	638	BIC:	-1723.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0990	0.005	-20.301	0.000	-0.109	-0.089
dates	0.0015	0.000	10.922	0.000	0.001	0.002

Omnibus:	6.921	Durbin-Watson:	1.485
Prob(Omnibus):	0.031	Jarque-Bera (JB):	9.250
Skew:	-0.080	Prob(JB):	0.00980
Kurtosis:	3.567	Cond. No.	72.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

DEM Politicians Sentiment Regression



DEM Politicians Regression Summary

OLS Regression Results

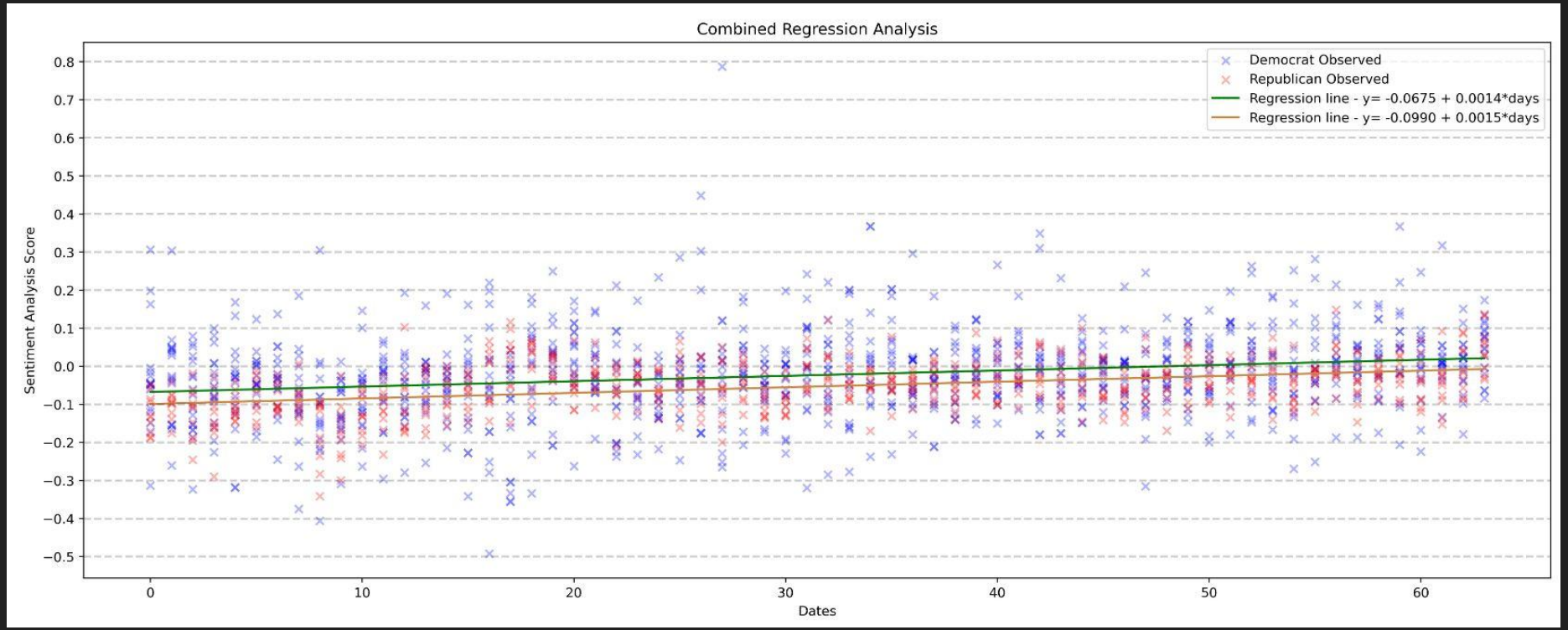
Dep. Variable:	points	R-squared:	0.052
Model:	OLS	Adj. R-squared:	0.051
Method:	Least Squares	F-statistic:	59.09
Date:	Mon, 06 Mar 2023	Prob (F-statistic):	3.36e-14
Time:	09:04:31	Log-Likelihood:	841.88
No. Observations:	1088	AIC:	-1680.
Df Residuals:	1086	BIC:	-1670.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0675	0.007	-10.081	0.000	-0.081	-0.054
dates	0.0014	0.000	7.687	0.000	0.001	0.002

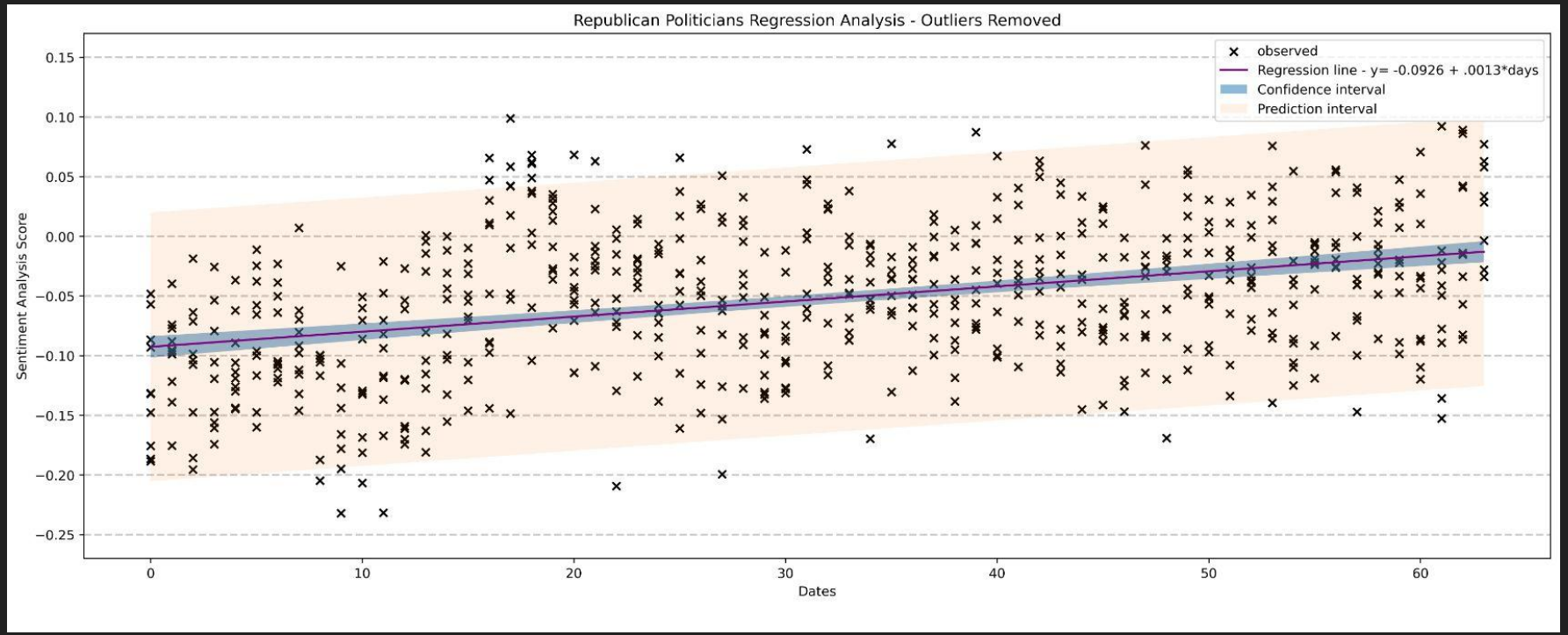
Omnibus:	148.890	Durbin-Watson:	1.958
Prob(Omnibus):	0.000	Jarque-Bera (JB):	687.947
Skew:	0.550	Prob(JB):	4.11e-150
Kurtosis:	6.737	Cond. No.	72.2

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Combined Politicians Sentiment Regression



GOP Politicians Sentiment Regression W/O Outliers



GOP Politicians Regression Summary (W/O Outliers)

OLS Regression Results

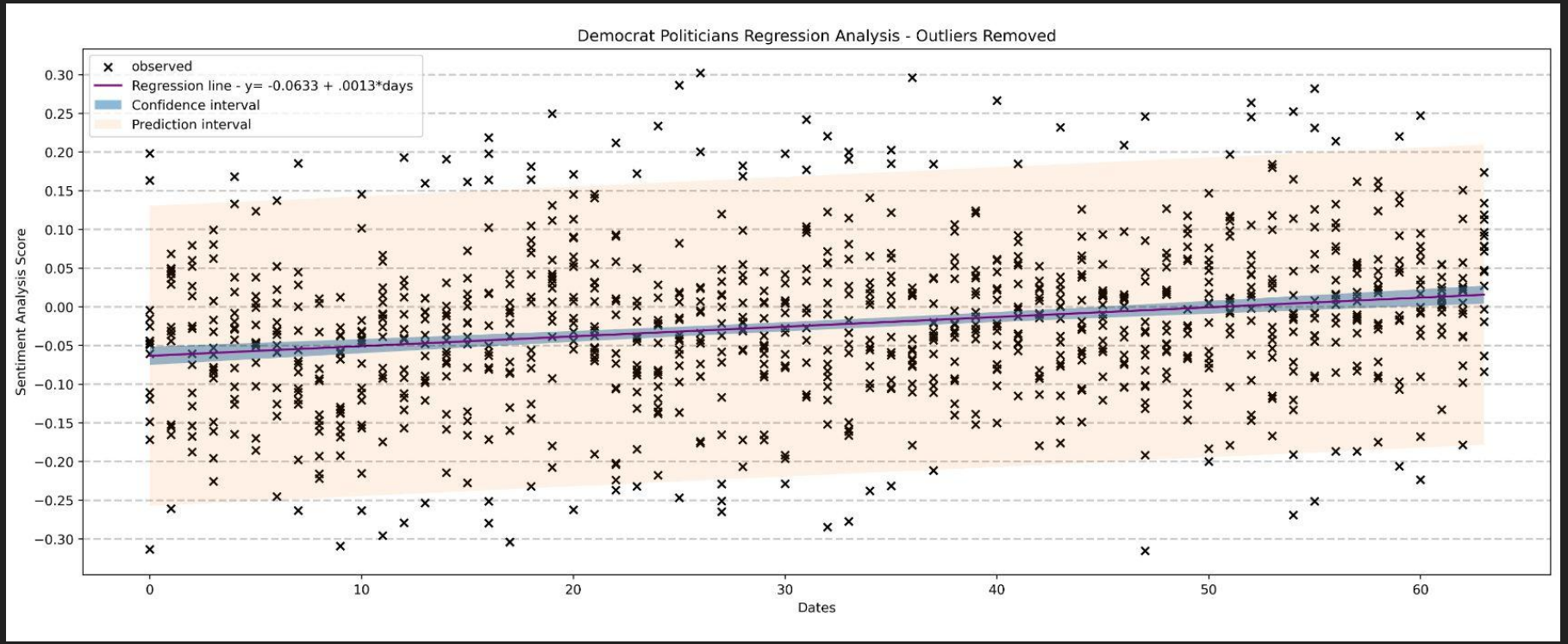
Dep. Variable:	points	R-squared:	0.142
Model:	OLS	Adj. R-squared:	0.141
Method:	Least Squares	F-statistic:	103.4
Date:	Mon, 06 Mar 2023	Prob (F-statistic):	1.43e-22
Time:	09:37:53	Log-Likelihood:	904.54
No. Observations:	626	AIC:	-1805.
Df Residuals:	624	BIC:	-1796.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0926	0.005	-20.328	0.000	-0.102	-0.084
dates	0.0013	0.000	10.167	0.000	0.001	0.002

Omnibus:	2.065	Durbin-Watson:	1.520
Prob(Omnibus):	0.356	Jarque-Bera (JB):	1.871
Skew:	0.038	Prob(JB):	0.392
Kurtosis:	2.743	Cond. No.	73.1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

DEM Politicians Sentiment Regression W/O Outliers



DEM Politicians Regression Summary (W/O Outliers)

OLS Regression Results

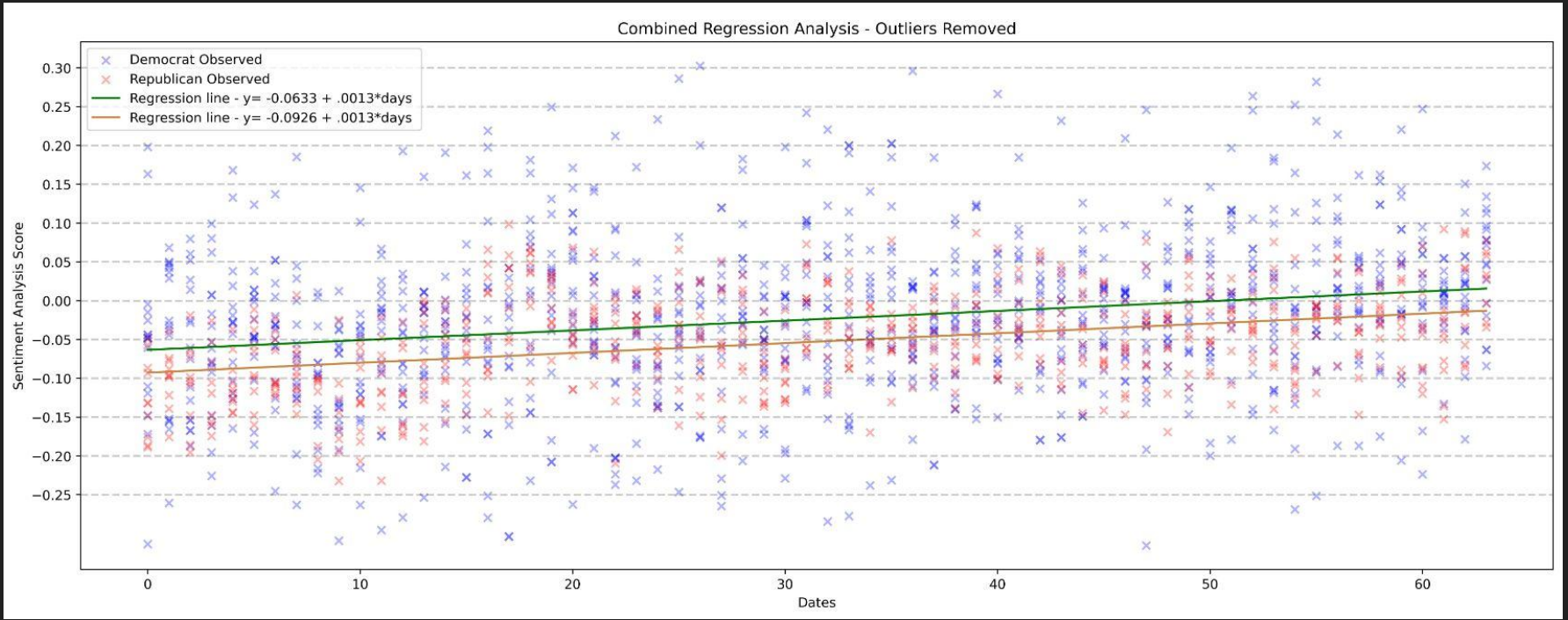
Dep. Variable:	points	R-squared:	0.052
Model:	OLS	Adj. R-squared:	0.051
Method:	Least Squares	F-statistic:	58.50
Date:	Mon, 06 Mar 2023	Prob (F-statistic):	4.53e-14
Time:	09:37:51	Log-Likelihood:	956.93
No. Observations:	1065	AIC:	-1910.
Df Residuals:	1063	BIC:	-1900.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.0633	0.006	-10.531	0.000	-0.075	-0.052
dates	0.0013	0.000	7.649	0.000	0.001	0.002

Omnibus:	15.256	Durbin-Watson:	1.906
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.049
Skew:	0.153	Prob(JB):	2.69e-05
Kurtosis:	3.617	Cond. No.	73.1

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Combined Politicians Sentiment Regression W/O Outliers



Multiple Regression Model

Sentiment Analysis Score = -0.0734 + 0.0014day - 0.0264GOP + 0.0174Woman + 0.0146On_Ballot

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.45793 -0.04688 -0.00071  0.04417  0.82382

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0733776  0.0057093 -12.852  < 2e-16 ***
day          0.0013577  0.0001255  10.817  < 2e-16 ***
republican  -0.0263802  0.0055160  -4.783 1.88e-06 ***
woman        0.0173731  0.0059734   2.908  0.00368 **
on_ballot    0.0146232  0.0053892   2.713  0.00673 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09458 on 1659 degrees of freedom
Multiple R-squared:  0.0948,    Adjusted R-squared:  0.09262
F-statistic: 43.44 on 4 and 1659 DF,  p-value: < 2.2e-16
```

Research Insights

Sentiment Analysis Scores

- Ignoring near-zero scores (which most-likely represent a failure in analyzing sentiment) LSA Scores for tweets mentioning Republicans and Democrats have a bimodal distribution around -0.5 and +0.5.
- Sentiment Analysis scores for Republicans are much more strongly correlated with one another than the Sentiment Analysis scores for Democrats.
- There was a very strong correlation between the Presidential candidate, Vice-presidential candidate, and political party. This held true for the combinations of Pres, Vice, and Party within a single party, as well as for the opposing party.
    - The exception is Pence who had weaker correlations on average

Significant Granger Causality Results

Trump Polling -> Trump Sentiment ***

Trump Polling -> Biden Sentiment ***

Biden Polling -> Trump Polling ***

^Item 1 Granger-Causes Item 2^

Sentiment Score Volatility

- Percent Change as a metric for volatility was not useful as there were many near-zero values. These would lead to +-700% spikes.
- Used Z-Scores as a metric for volatility instead

Day of Week/Week Analysis

- Individual days of the week (monday, etc.) did not seem to have an effect on sentiment score
- As scores trended higher over time, which week it was had a statistically significant effect on sentiment scores.

Econometric tests

- People Spoke more positively about a candidate if:
  - They were a Democrat
    - This makes sense because Democrats ultimately won the presidential election and a plurality of congressional elections
  - They were a woman
  - They were up for election/on a 2020 ballot
- People spoke more positively about politicians as Nov. 3rd approached

Next Steps

Next Steps (summary)

- Further construct Econometric models
- Incorporate other politicians + come up with a more solid methodology of choice for politicians
- Time-of-day analysis
- Analyze:
  - popular, non national politicians (mayors of NY, Chicago etc.)
  - Gov. institutions such as SCOTUS, CDC, etc.
- Compare Sentiment scores of name mentions vs. *@username* tagging.
- Create a deep-learning model to analyze relationships between certain topics and sentiment score

Econometrics

- Check if being a Woman and being a Democrat are correlated with each other (autocollinearity). If so, the regression needs to be adjusted to account for collinearity
- Test Tweets/day as a variable

Deep Learning Model

1. Use an LDA topic model library to identify and determine which topics are driving sentiment over time. (Currently planning on using LDA DAVIS' Gensim-LDA library)
2. Tune the model by testing various hyperparameter configurations
3. Once I find the optimal model configuration. Create a model to find out how positive or negative the tweets associated with the various topics are. This would allow us to find out what caused changes in trend

Hyper-Parameter Tuning

There are 3 important Hyperparameters in the Gensim-LDA library.

- Alpha = The Document Density within the corpus
- Beta = The density text within the Documents
- *K* = The number of words within a topic

I would choose the tweets from 7 days spread across the 64 days. While the day of the week doesn't seemingly impact the sentiment analysis score, the day of week does affect reporting of news, so I will remain cognizant of the day of the week when choosing testing data.

Hyper-Parameter Tuning II

The testing would give two outputs

- Coherence = the score of a single topic by researching degree of semantic similarity between high-scoring words within the topic
- Perplexity = A measure of how good the model is. [Lower the better]

After testing the various combinations of Hyperparameters, I would apply the variables that would maximizes the Coherence Score while minimizing the Perplexity Score. I would then run this model on the entire dataset.