# Analaying Sinusoidal Patterns in 2020 U.S. Presidential Election Tweets

Andre Ye

June 2021

## Contents

# 1 Introduction

The 2020 U.S. Presidential Election between the Republican incumbent Donald Trump and the Democratic nominee Joe Biden was one embroiled in substantial amounts of controversy and tension. Amidst the coronavirus pandemic, a nationwide reckoning on race relations, and hesitance on election security, discourse on the election has been chaotic and turbulent. Twitter, especially, has been a platform whose design, which restricts tweet lengths to 280 characters, pushes political conversations to evolve and morph rapidly.

Amidst chaos and turbulence, structure is valuable. This paper will attempt to study patterns within Twitter data on the 2020 Presidential Election.

# 2 Data Collection and Preparation Processes

The Twitter API was used to collect tweets with hashtags related to Donald Trump and Joe Biden. Moreover, the country location of the accounts was restricted to the United States. The tweets were cleaned using a preprocessor that removed hashtags and URLs to aid clear sentiment analysis. The resulting dataset consists of 170,021 tweets, with 91,096 Trump-hashtag tweets and 78,925 Biden-hashtag tweets.

Sentiment analysis was performed by the library `TextBlob`, which allows for the calculation of the polarity and subjectivity of text. Polarity pertains to the emotional content of the text from $-1$ (extremely negative) to 1 (extremely positive). Subjectivity pertains to how opinionated the text is, from 0 (extremely factual) to 1 (extremely opinionated). The subjectivity score does not take into account the actual truth or accuracy of the statement, but instead the expression of the text. `TextBlob` uses a simple system of weights for each word; each relevant word – a word is often relevant if it is an adjective – is associated with a weight pertaining to its polarity and subjectivity, and the result is the product of these weights, capped at the minimum and maximum possible output values. While certainly far from a sophisticated system, this method of sentiment analysis serves as a fairly accurate approximation of sentiment, which serves the purposes of this paper well enough. Because weights can be negative, for instance, simple double negatives can be accounted for (see last row). See the table below for examples of `TextBlob`'s sentiment analysis capabilities.

| **Text** | **Polarity** | **Subjectivity** |
|---|---|---|
| "Sentiment analysis is a tool." | 0.0 | 0.0 |
| "I like sentiment analysis, I guess" | 0.0 | 0.0 |
| "I really like sentiment analysis" | 0.2 | 0.2 |
| "I love sentiment analysis" | 0.5 | 0.6 |
| "I love sentiment analysis!!!" | 0.97 | 0.6 |
| "OMG, I love sentiment analysis!!! It is amazing and wonderful." | 0.86 | 0.83 |
| "I hate sentiment analysis!" | -1.0 | 0.9 |
| "I do not hate sentiment analysis!" | 0.5 | 0.9 |

`TextBlob` does have a few limitations, like struggling with contractions, which presumably is not in `TextBlob`'s corpus, complex stacks of negatives, punctuation, and others.

| **Text** | **Polarity** | **Subjectivity** |
|---|---|---|
| "I don't hate sentiment analysis!" | $-1.0$ | 0.9 |
| "I hate people who hate sentiment analysis" | $-0.8$ | 0.9 |
| "I do not not hate sentiment analysis!!!" | 0.5 | 0.9 |
| "Why doesn't everyone love sentiment analysis???!!!????" | 0.98 | 0.6 |
| "no! do not! bad!" | 0.44 | 0.67 |

However, given that tweets on Twitter are restricted to a limited number of characters, it seems that a Twitter user would not rationally use their character count or time to construct rhetorical patterns that would be elaborate or vague enough to fool TextBlob too badly. Given that tweets seem to be simple and direct, especially in relation to the election, we can put a reasonable amount of faith in the indicative capability of the polarity and subjectivity.

Using `TextBlob`'s sentiment analysis capabilities, along with other functionalities, the cleaned dataset that was used for analysis consisted of the following features:

- *Candidate.* Whether the Tweet hashtag implied it concerned Donald Trump or Joe Biden.
- *Datetime.* A Python datetime object indicating the exact second at which the Tweet was posted.

- *Hour.* The *Datetime* column rounded to the nearest hour.
- *Length.* The length of the cleaned tweet, which consists of "purer" content (URLs and hashtags were excluded in text preprocessing).
- *Polarity.* The polarity score of the tweet content.
- *Subjectivity.* The subjectivity score of the tweet content.

## 3 Data Exploration and Inquiry Formulation

The purpose of this section is to conduct a simpler numerical and visual exploration of the data to a) better understand the dynamics of the data and of the phenomena being represented, and b) to identify phenomena of interest and to formulate meaningful paths of inquiry for further investigation.

The means and standard deviations for the key measurement features of the dataset – length, polarity, and subjectivity – by candidate yield interesting results. The spread, or standard deviation, of tweet length, polarity, and subjectivity across candidate are almost identical. Tweets on Trump are, on average, slightly longer than Biden's by about 6 characters, which is insignificant relative to the range of 280 characters. Tweets on Trump and Biden also have similar subjectivity scores, on average. Moreover, the average subjectivity is rather low compared to the highly-opinionated nature of political tweets one may expect. It is unclear whether this low subjectivity is the result of actual lack of opinionated tweets, an excess of factual tweets (perhaps by news sources), an excess of opinionated tweets presented as or using language that appear factual that cannot be detected by the simple sentiment detection algorithm, or a combination of these. There is also a relatively insignificant but present gap between the polarity of tweets; the average polarity score for tweets on Trump is 0.05 more negative than that of Biden's. This is, like with subjectivity, relatively insignificant compared to the range of 2.0.

| | **Length** | | **Polarity** | | **Subjectivity** | |
|---|---|---|---|---|---|---|
| **Candidate** | **Mean** | **STD** | **Mean** | **STD** | **Mean** | **STD** |
| Trump | 184.50 | 80.31 | 0.05 | 0.28 | 0.34 | 0.32 |
| Biden | 178.87 | 81.66 | 0.10 | 0.29 | 0.33 | 0.32 |

The mean of the polarity and subjectivity by hour was computed for each candidate. The Pearson correlation coefficients were computed across these four trends.

- $\rho$(Trump polarity, Biden polarity) = 0.2702
- $\rho$(Trump subjectivity, Biden subjectivity) = 0.2833
- $\rho$(Trump polarity, Trump subjectivity) = $-0.0683$
- $\rho$(Biden polarity, Biden subjectivity) = 0.1522

Each hour, when the polarity of tweets relating to Trump rises, so does the polarity of tweets relating to Biden. This opposes what one may expect – that the polarity of tweets relating to one candidate rises as the other falls. There is a similarly strong correlation between the subjectivity of tweets relating to candidates every hour. This may be the result of the "environment" outside of Twitter more broadly, in which certain events provoke subjective or less subjective tweet content more broadly. It is, furthermore, interesting that there is no clear linear relationship between the polarity and subjectivity of tweets relating to Trump, but a weakly present linear relationship between that of tweets relating to Biden.

We can begin a visual inspection of tweets by plotting the number of tweets released per hour in Figures 3.1 and 3.2. There is a clear sinusoidal relationship between the date-hour and the number of tweets released.

Particularly, there seems to be a consistent sinusoidal variation that occurs on a daily basis – people probably tweet more often during the day than at night – that grows much larger from November 1st, 2020 to November 9th, 2020. This coincides with the approximate range of the time period during which the election process was just about to happen, was in the process of happening, or just happened. We can measure how much of an aberration these last few days were compared to the consistent daily sinusoidal pattern prior by fitting a sinusoidal curve to data prior to November 1st and measure the difference between the predicted sinusoidal pattern and the actual data after November 1st. Moreover, we can compare how this is different across candidate to draw conclusions as to whether the Twitter population was drawn more towards commenting on Trump or Biden near and post-election.

We can also plot the lengths of cleaned tweets across time as well in Figures 3.3 and 3.4; there does not appear to be any sinusoidal relationship, although there appears to be a greater concentration of tweets with lower quantities of tweets. Low-length tweets are especially dense from the November 1st to November 9th window. Because there does not appear to be any clear relationship between the length of the tweet and the date, this paper will not further investigate tweet length.

Visualizations for the polarity and the subjectivity are complex, and for the purposes of this introduction section it is not necessary to display and explore them here – they will be further analyzed in their respective sections.

This paper will pursue the following paths of inquiry:

- What was the sinusoidal pattern for the number of tweets before the election "hit" for each candidate? How much larger was the quantity of tweets during or after the election occurred (window of November 1st to November 9th) compared to what the "normal" quantity would have been?
- What sinusoidal patterns can we identify in the polarity of tweets by candidate over time?
- What sinusoidal patterns can we identify in the subjectivity of tweets by candidate over time?

# 4 Modeling Quantity of Tweets Over Time

The data was isolated to all tweets posted before November 1st, 2020. The four parameters in the sinusoidal model of the form $A\sin(B(t-C))+D$ were optimized using nonlinear least squares. (It should be noted that all sinusoidal-fits in this paper were found using nonlinear least squares optimization, given the number of fits required for exploring the asked questions. Justification for this and specifics of initialization is further elaborated in section 5.3.) The following parameters and RMSE for the corresponding model derived for tweets relating to Trump and Biden are listed in the table below and visualized in figures 4.1 and 4.2.

| **Candidate** | $A$ | $B$ | $C$ | $D$ | **RMSE** |
|---|---|---|---|---|---|
| Trump | $-57.65$ | 0.27 | 3.80 | 113.08 | 39.89 |
| Biden | $-44.55$ | 0.27 | 3.93 | 80.83 | 38.86 |

The sinusoidal models seem to fit well; the root-mean squared error of about 40 for each candidate captures the rough trend of tweets, and is likely exacerbated by the existence of outliers, as demonstrated visually.

The $B$-parameters for each model should be rewritten such that $\frac{2\pi}{p}=B$; $p=\frac{2\pi}{B}=\frac{2\pi}{0.27}\approx 23.27$ hours for Trump and Biden (whose models had very close values of $B$). This confirms our prior hypothesis of the variation being a daily cycle. The amplitude of the quantity of tweets relating to Trump over time is larger in absolute value than that of tweets relating to Biden. This can likely be attributed to the volatility and controversy of coverage

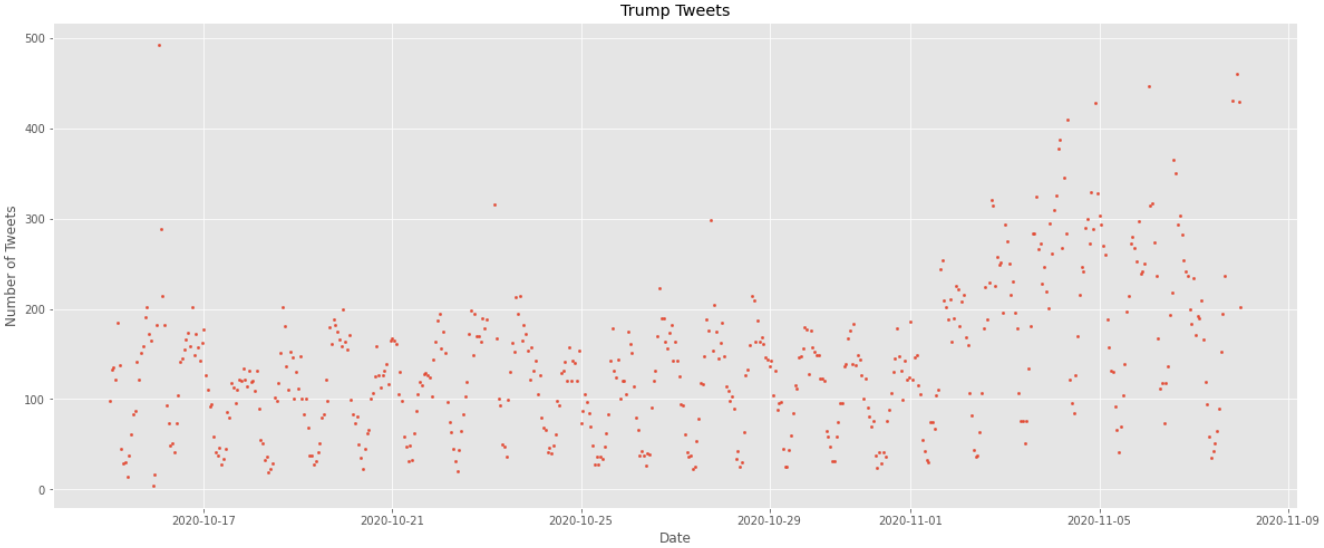

Figure 3.1: Number of Trump-related tweets per hour.

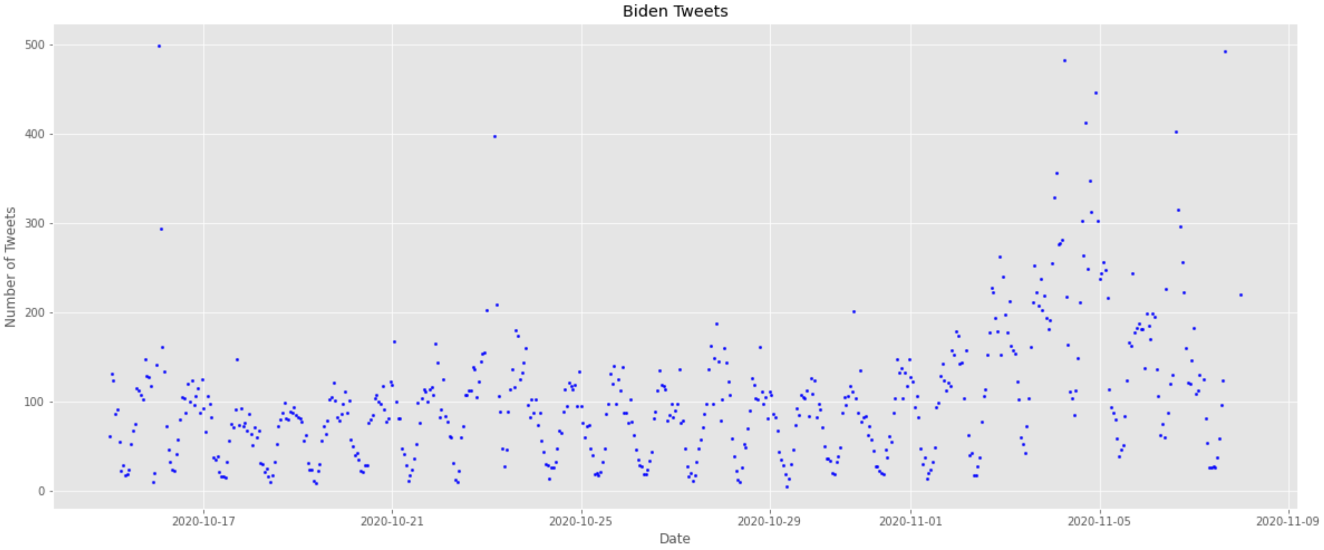

Figure 3.2: Number of Biden-related tweets per hour.

and discussion on Trump and Biden's relatively more "stable" discourse. There seems to be an insignificant shift between the number of tweets per hour between Biden and Trump, suggesting that the Twitter conversation on the two candidates followed relatively the same directions of rising and falling at certain times. Lastly, the mean of the model for tweets relating to Trump is significantly higher than that of the model for tweets relating to Biden; this, again, can likely be related to the Trump's controversial aura relative to that of Biden.

When the model is used to predict on dates past November 1st, it incurs the following errors:

| **Candidate** | **Mean Absolute Error** | **Root Mean Squared Error** |
|---|---|---|
| Trump | 112.53 | 138.42 |
| Biden | 95.54 | 126.12 |

The predictions of the models compared to the real data are displayed in figures 4.3 and 4.4. There is a very clear increase in the number of tweets across both candidates. The root-mean squared error for the model on data after November 1st, 2020 is about 3.5 times that of the RMSE on data before November 1st, 2020 for Trump. That quantity decreases to 3.2 times for tweets relating to Biden. Tweets on Trump very near or during the election thus

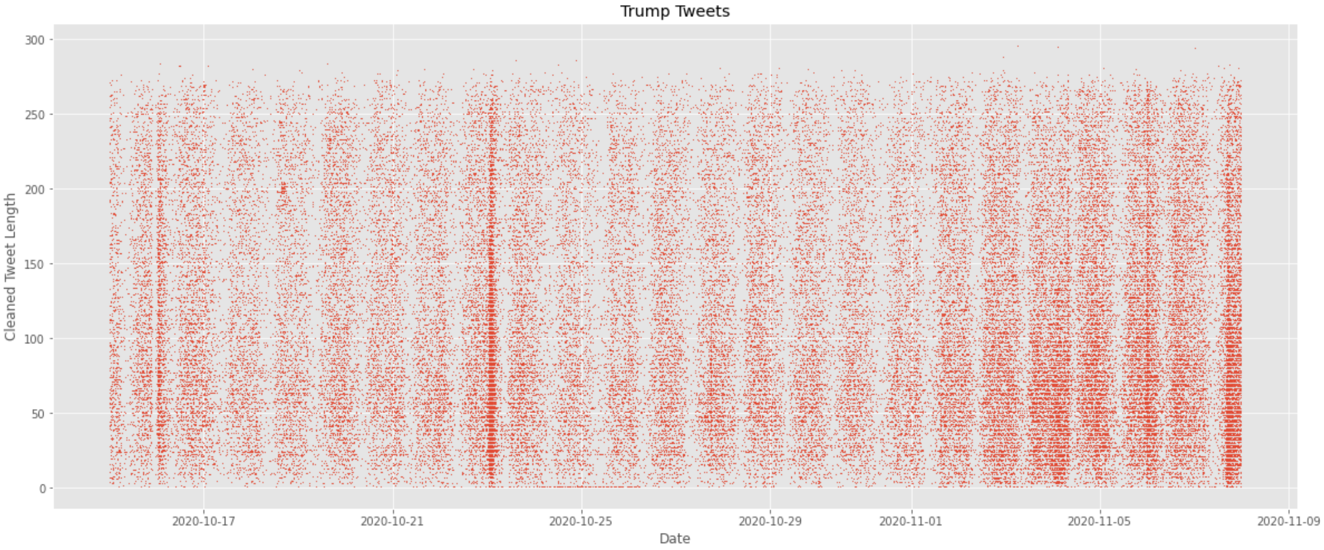

Figure 3.3: Scatterplot of the length of tweets related to Trump and the datetime they were released.

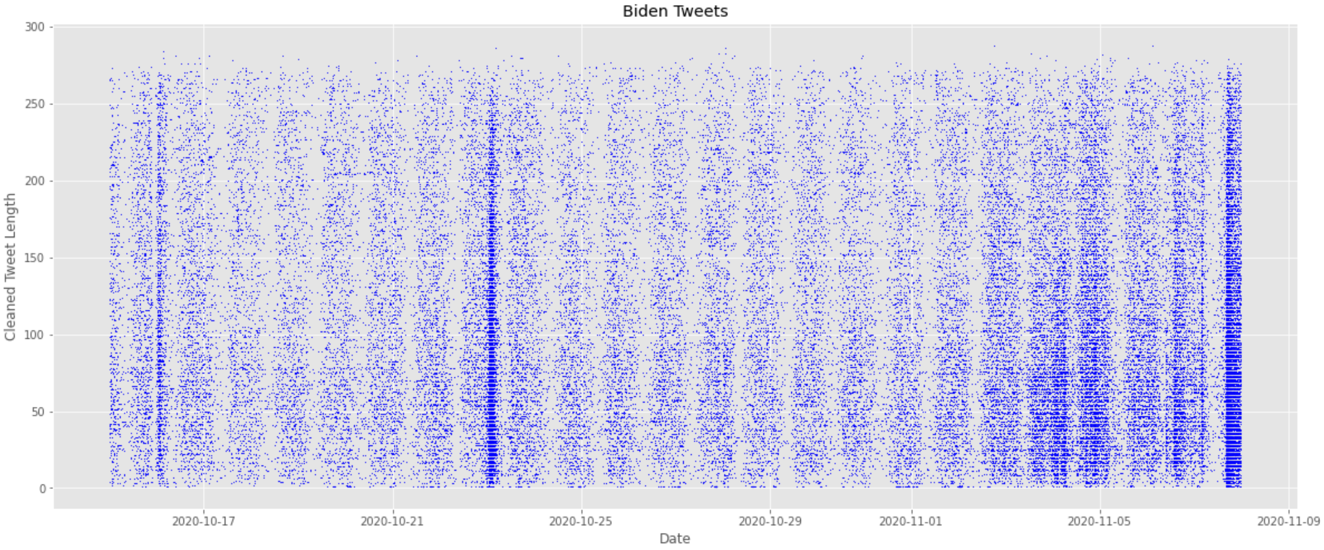

Figure 3.4: Scatterplot of the length of tweets related to Biden and the datetime they were released.

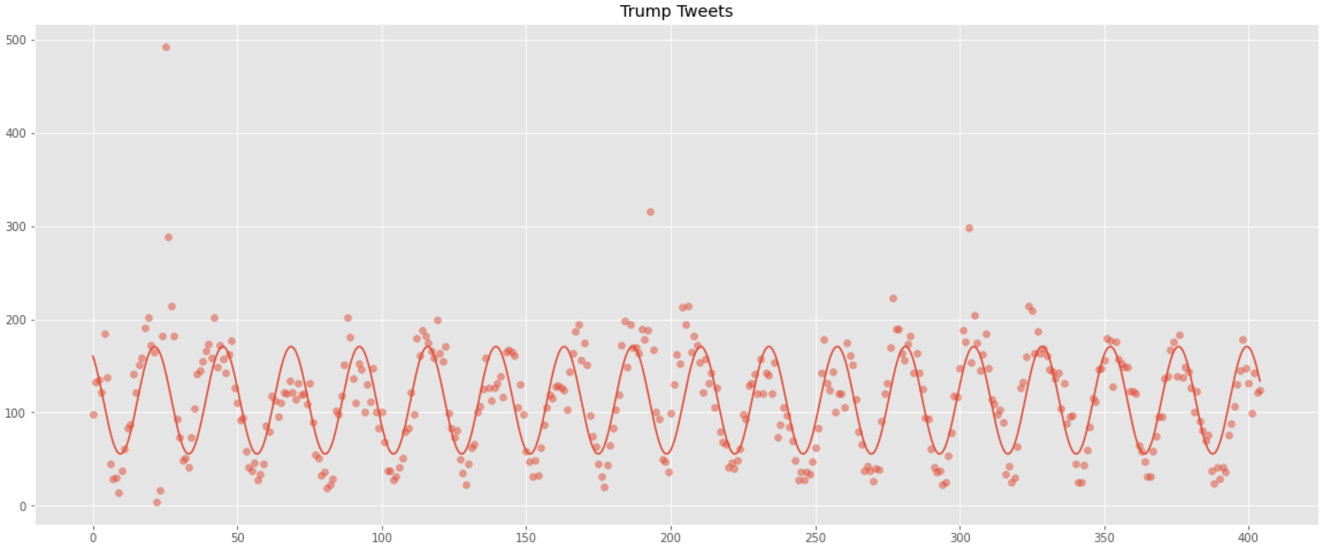

Figure 4.1: Sinusoidal fit of the number of tweets relating to Trump per hour laid over a scatterplot.

deviated more from the predicted cycle than tweets on Biden. Moreover, a graphical analysis of figures 4.3 and 4.4 demonstrate that the sinusoidal fluctuation of tweets from November 1st to the 9th did not "shift up"; rather, the amplitude changed such that the minimum number of tweets from the 1st to the 9th roughly coincides with the predicted model, but the maximum number of tweets surpasses that which is predicted. Furthermore, both in tweets pertaining to Trump and Biden, the amount by which the peak of the real data during the election varies for the predicted cycle seems to rise and fall such that the highest deviation occurs roughly around November 3rd or 4th.

The prior results on tweets on Trump being more volatile or more frequent in general than tweets on Biden were attributed to Trump being a more controversial candidate than Biden. We can reframe this notion of controversy both to explain the results in this section. A Brookings Institute study found that experts ranked Trump as the most polarizing of all presidents, pursuing policies like a travel ban and withdrawal from the Paris Climate Accord. Polarization increases the "need" to share one's opinion – in most cases, where there is agreement, there is not much to say. Especially on a platform like Twitter, in which thoughts are expressed in short form by design, replying is quick, and the time scale is almost exactly in sync with world events, one can reasonably expect for the effects of relative polarization to be well-pronounced on Twitter.

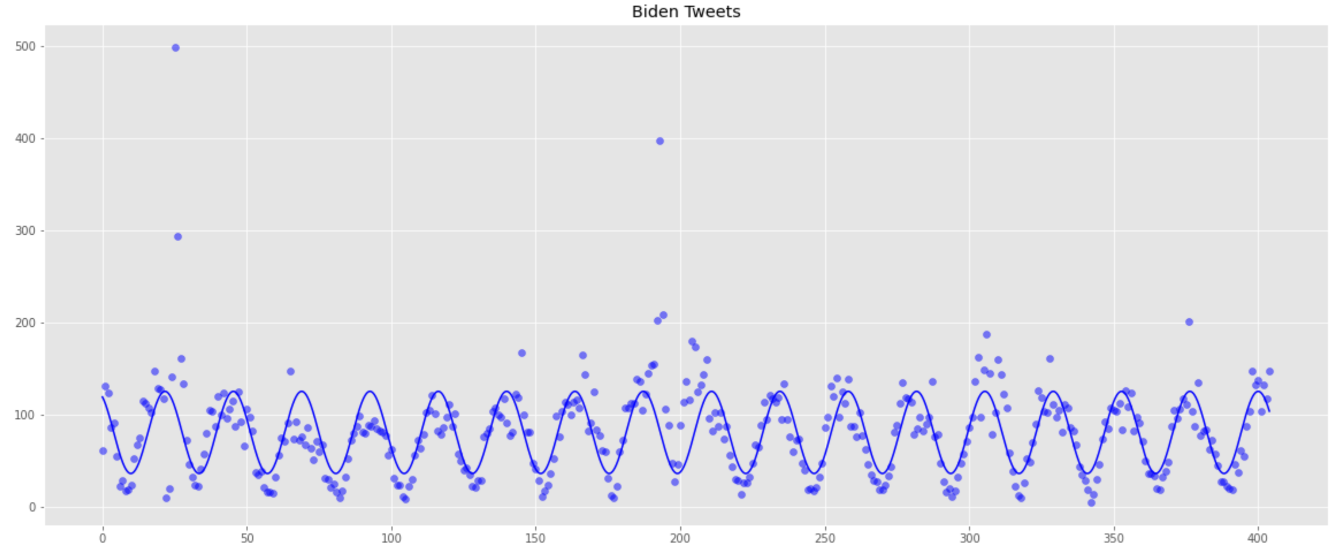

Figure 4.2: Sinusoidal fit of the number of tweets relating to Trump per hour laid over a scatterplot.

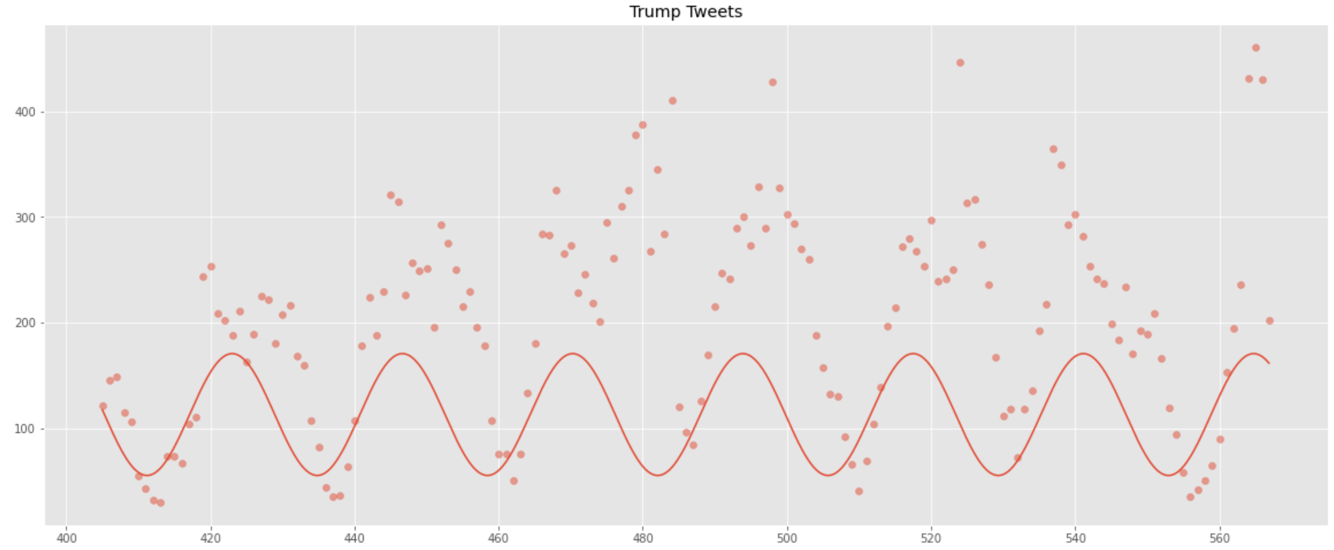

Figure 4.3: Predicted sinusoidal fitted on data before November 1st, 2020 overlaid on a scatterplot of data after November 1st, 2020 for tweets on Trump.

The following sections will transition away from the quantity of tweets, instead focusing on their content via the lens of polarity and subjectivity.

# 5 Modeling Processes and Standards

The following sections, ”Modeling Polarity of Tweets Over Time” and ”Modeling Subjectivity of Tweets Over Time”, attempt to understand how the shape of the distribution of tweet polarity and subjectivity relating to each candidate change over time. Because the polarity and subjectivity of tweets are “messier” than the quantity of tweets over time it is necessary here to establish modeling processes and standards that will be employed in the following sections.

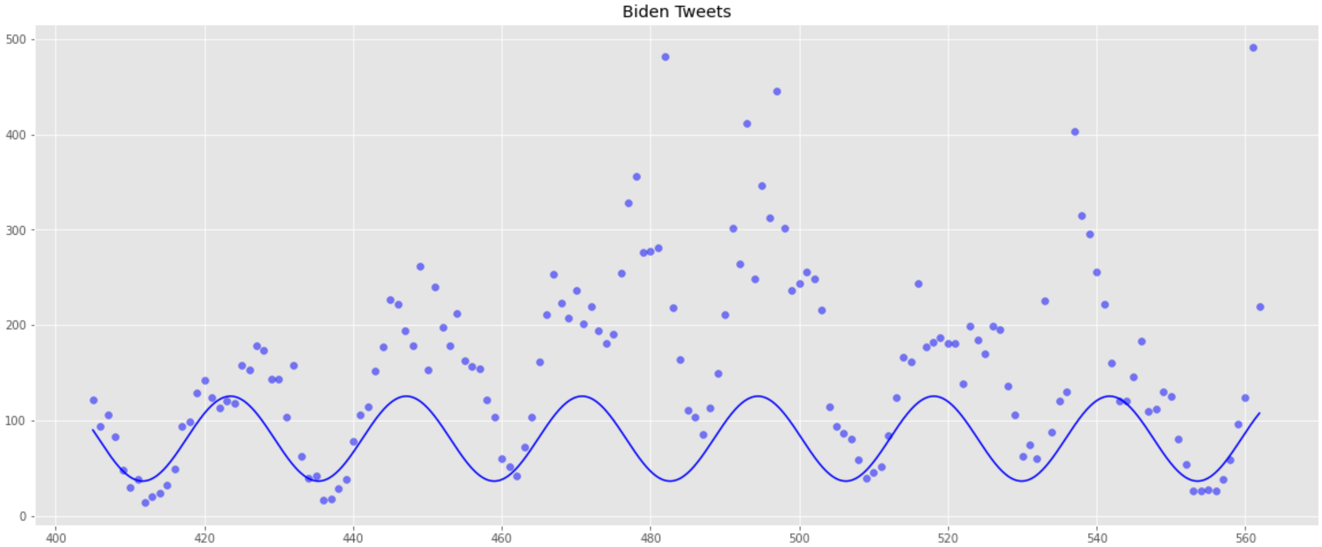

Figure 4.4: Predicted sinusoidal fitted on data before November 1st, 2020 overlaid on a scatterplot of data after November 1st, 2020 for tweets on Biden.

## 5.1 Justifying Focus on Mean and Standard Deviation

Section 4 modeled a fairly straightforward phenomenon - the quantity of tweets released every hour on a candidate. The measured phenomena in Sections 6 and 7 are not so straightforward, and thus require some justification. The following two sections concern the polarity and subjectivity of tweets relating to either candidate; Section 3 demonstrated what "polarity" and "subjectivity" entails. Figures 5.1 and 5.2 show a scatterplot of the polarity of tweets relating to Trump and the polarity of tweets relating to Biden. The subjectivity of tweets are omitted here for brevity, although they were considered in structuring modeling processes and standards.

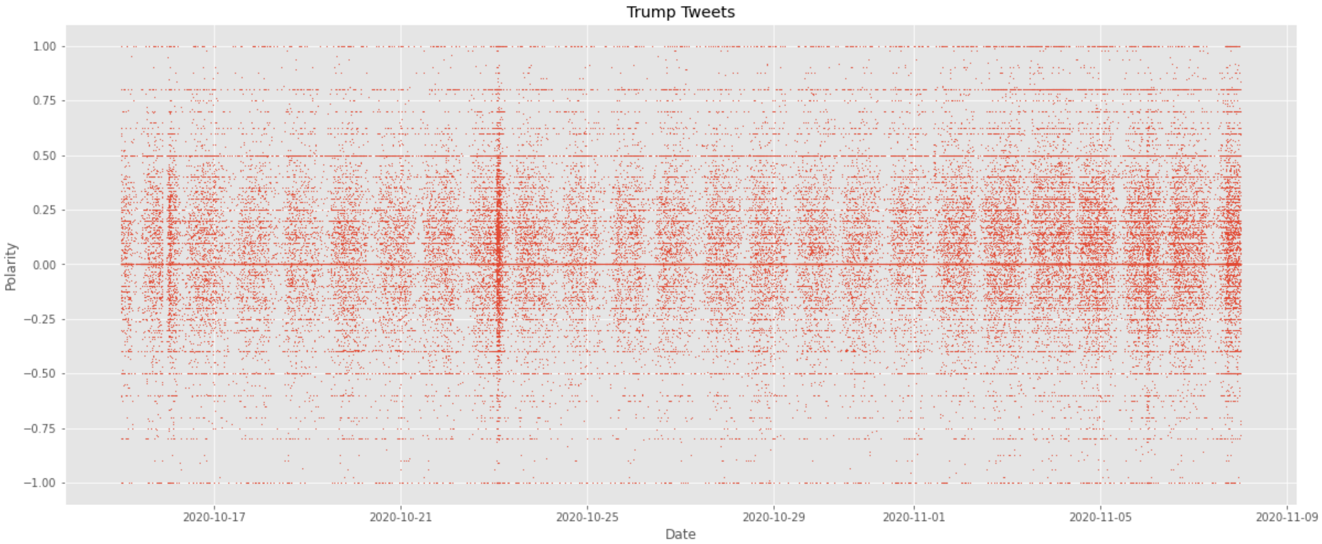

Figure 5.1: Scatterplot of the polarity of related to Trump and the datetime they were released.

There is a much less clear sinusoidal pattern than was observed with the quantity of tweets every hour on a candidate. The most clear pattern one notices is the the tweets form vertical "bands"; this likely corresponds with the daily fluctuation of tweets as explored in the prior section. There appears, however, to be another pattern, in which the distribution of the polarity of tweets changes over time. It seems that the distribution is more uniform in the spaces between the "bands" and more centered towards the middle on the "bands". Moreover, the center, most-dense area of the bands seem to rise and fall slightly in a roughly sinusoidal pattern. We need to transform

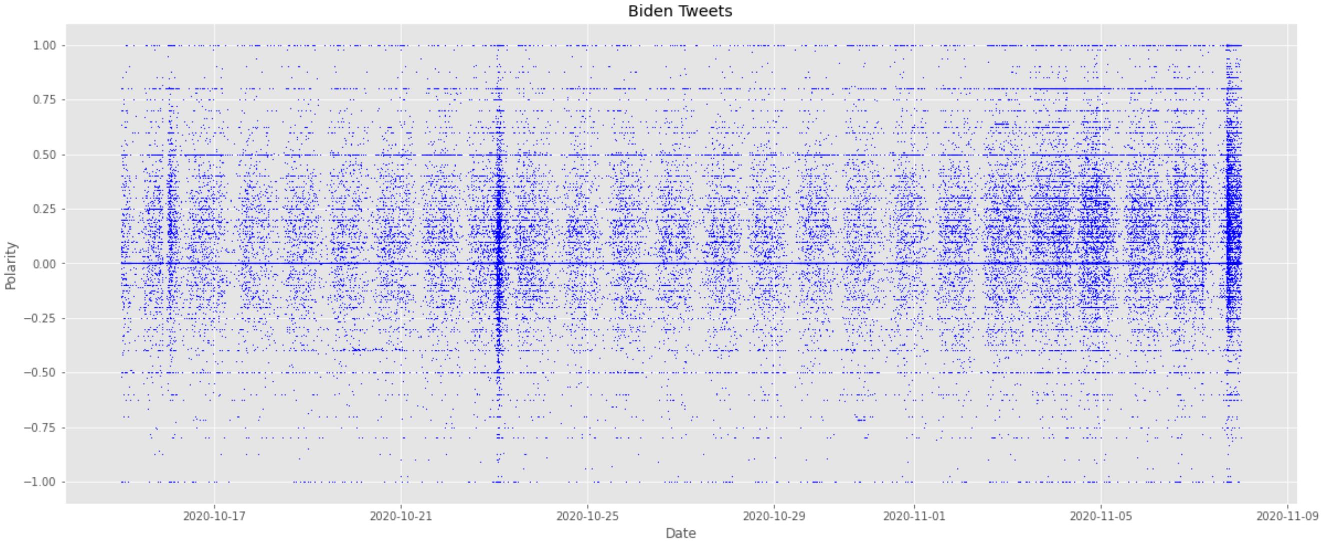

Figure 5.2: Scatterplot of the polarity of tweets related to Biden and the datetime they were released.

the data, however, to make these qualitative assessments concrete.

The focus for the polarity and subjectivity of the data will be on how the distribution of said metrics changes over time. The mean and the standard deviation are two primary defining characteristics of a distribution. Thus, the following sections will model the mean and standard deviation for the both the polarity and the subjectivity of the data to understand in quantitative terms how the distribution of these metrics differs across time and by candidate.

To take the mean and standard deviation at every datetime would be an unsuccessful strategy, because datetimes are hyper-specific and any tweet is unlikely to share the same datetime as any other tweet. Rather, we utilize a similar "rounding" or "binning" approach as in the prior section such that the mean and standard deviation of the polarity and subjectivity are taken on all tweets relating to a certain candidate within an hour-length interval.

## 5.2 Multi-window Approach to Discovering Layered Patterns

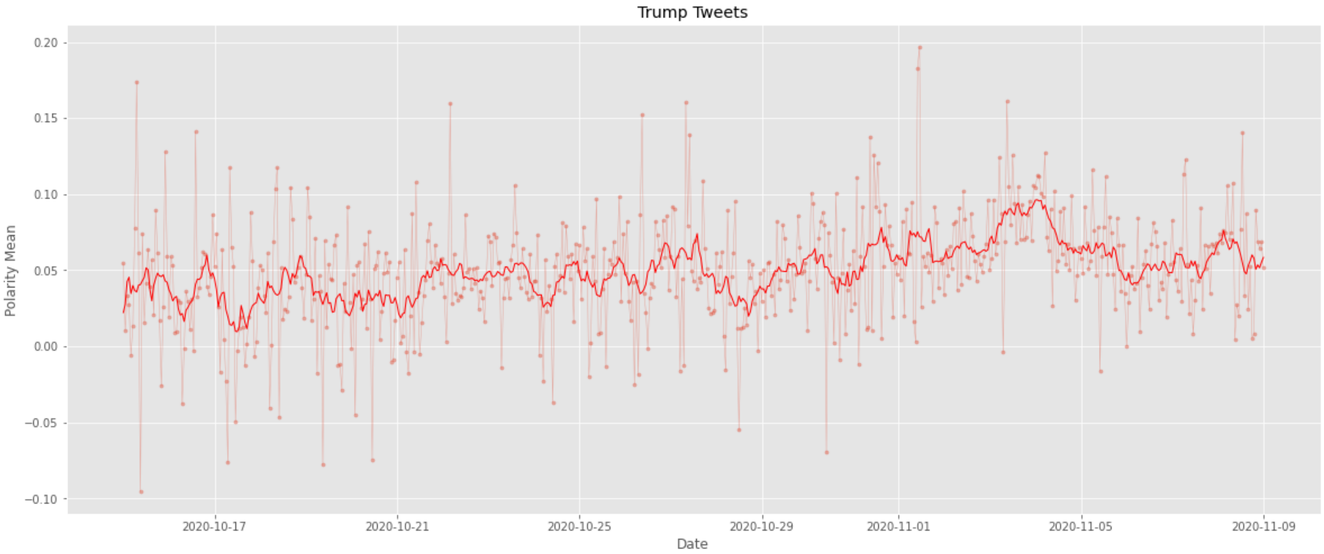

Figure 5.3: Plot of mean of polarity per tweets relating to Trump per hour with 12-hour sliding window.

Figures 5.3 and 5.4 show the mean of the polarity of tweets relating to Trump and Biden (back, faded) behind moving-average-smoothed data (front, bold). A visual inspection of the "raw" means by hour demonstrate

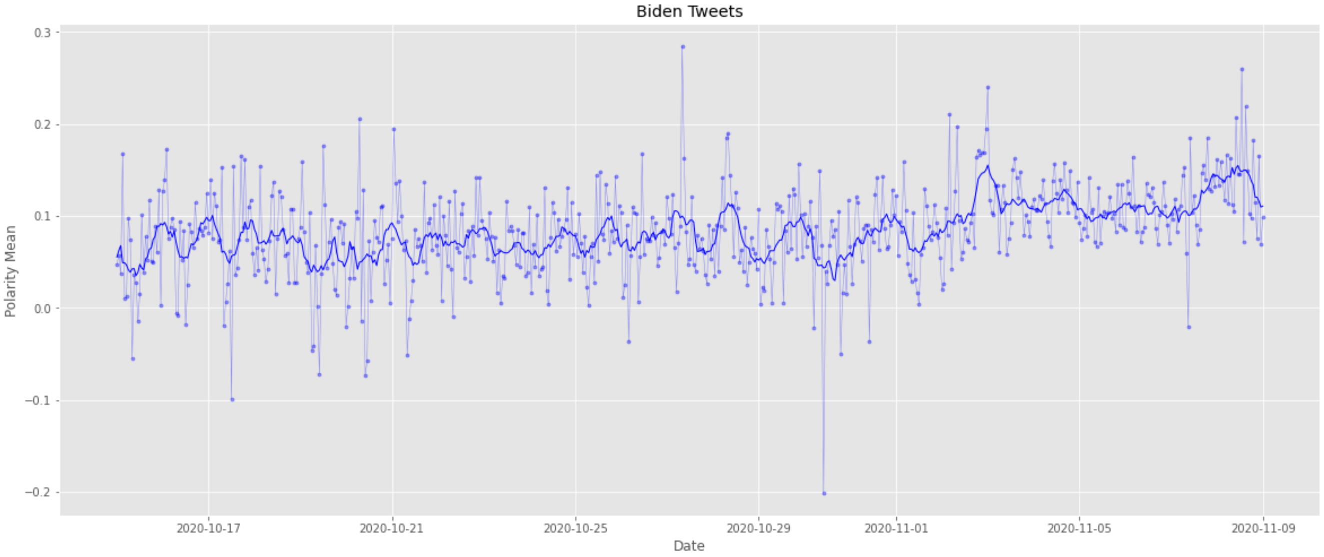

Figure 5.4: Plot of mean of polarity per tweets relating to Biden per hour with 12-hour sliding window.

tremendous volatility. Using a moving average helps to identify the trend of the data, but even then there seems to be several sinusoidal patterns that span a wide range of periods. The purpose of this paper is not to predict the polarity or subjectivity of a tweet, but to look at possible sinusoidal trends that may be occurring. Thus, we will not adopt an epicycle-like approach to "unpacking" layered patterns. Like before, although figures 5.3 and 5.4 only show the mean of the polarity of tweets for the sake of brevity, this decision also incorporated the mean of subjectivity, the standard deviation of polarity, and the standard deviation of subjectivity.

However, we can use moving-average with different window sizes to smooth the data to varying degrees. A moving-average smoothing with a small window size will encourage modeling of smaller-period sinusoidal patterns, whereas a large window size will encourage modeling of larger-period sinusoidal patterns. For purposes of definition, a moving average series $m$ is constructed from a time series $t$ with window size $w$ by

$$m_j = \frac{\sum_{i=j}^{w} t_i}{w}$$

A smaller window size entails an average taken with less elements, and thus each element of $m$ is constructed based on a more "time-specific" interval that allows for higher-frequency sinusoidal patterns to retain their general shape.

For every phenomena that will be modeled in the following two sections, six versions of the data will be generated using windows of lengths 1 hour (no change to data, since the data is bucketed by 1 hour), 2 hours, 5 hours, 12 hours, 24 hours, and 48 hours. A sinusoidal function will be generated for each. Analysis of the results will include

- Derived parameters $a$, $b$, $c$, and $d$ in $a\sin(b(x-c))+d$
- Period, which can be calculated as $\frac{2\pi}{b}$
- Mean Average Error (MAE) of the model on the corresponding smoothed data
- Root Mean Squared Error (RMSE) of the model on the corresponding smoothed data

This paper will investigate the polarity and the subjectivity. For each metric of the content of the tweets, the mean and standard deviation will be modeled. For each metric of the distribution of the polarity/subjectivity of tweets, there are two versions of the data corresponding to tweets relating to either candidate. There are thus

$$2 \text{ metrics of content} \times 2 \text{ metrics of distribution} \times 2 \text{ candidates} = 8 \text{ phenomena}$$

Because each phenomena is modeled with 6 variants of smoothed data, 8 phenomena × 6 window sizes = 48 sinusoidals will need to be generated.

### 5.3 Automated Adjustment for Nonlinear Least Squares Fit

As was briefly discussed in the previous section on modeling the quantity of tweets over time, automated nonlinear least squares fit will be used to calculate the line of best fit. Incomplete attempts to automate a "half-information" approach, in which $a$ and $d$ in $a\sin(b(x-c))+d$ are found given $b$ and $c$ and vice versa via transformation and linear regression fit, are discussed in further inquiry. Ultimately, failure to implement this method and the need for an automated process of fitting sinusoidal functions given the high-volume demand required for sections 5 and 6 as discussed in 5.2 seems to justify utilizing nonlinear least squares.

Nonlinear least squares initial guesses for $a$, $b$, $c$, and $d$ were derived as follows. $x$ and $y$ refer to the set of independent variables and the target variable, respectively.

- $a$: $\sqrt{2}\cdot\sigma(y)$. This was used rather than $\max(y)-\min(y)$ because the data is rather volatile and the actual maximum of the data often does not indicate the "reasonable maximum".
- $b$: $2\pi f$. $f$ is the frequency found via Fourier Transform.
- $c$: 0. The shift can probably be found relatively easily and does not need a strong estimate.
- $d$: $\bar{y}$. The mean of the target variable will suffice.

The solution nonlinear least squares yield for the parameters $a$, $b$, $c$, and $d$ in $a\sin(b(x-c))+d$ are often not in a form optimal for analysis. On the same phenomena, nonlinear least squares may arrive at positive and negative values of $a$ for different window sizes (degrees of smoothing). This makes comparison of the shift $c$ difficult to compare across different window sizes. Moreover, it may be possible for the solution $c$ to be one of the infinitely many that can be derived by adding or subtracting multiples of the period from a "root" shift value. Because it is unreasonable to go about adjusting the parameters for each of the 48 sinusoidal models, we need to develop an automated method to reform the derived solutions in a format conducive to analysis.

To adjust for negative derived solutions for the value of $a$, consider a simplified version of the model $-a\sin\left(b\left(x-c\right)\right)$. We would like to find the shift made to $c$ such that this function is equivalent to itself, but with the coefficient to the sin term being positive ($a$ rather than $-a$). We can do so by shifting the first model by $\Delta$: $-a\sin\left(b\left(x-\left(c+\Delta\right)\right)\right)$. Changing $-a$ to $a$ amounts to a reflection of the sinusoidal across the $x$-axis. The minimum of $-a\sin\left(b\left(x-c\right)\right)$ is the maximum of $a\sin\left(b\left(x-c\right)\right)$, and vice versa. Thus, we can coincide the maxima of the two sinusoidals (they have the same period) by shifting the $-a\sin\left(b\left(x-c\right)\right)$ half a period in either direction. The period is $\frac{2\pi}{b}$; half the period is $\frac{\pi}{b}$. Thus $\Delta=\frac{\pi}{b}$. We can derive the following simple adjustment for negative derived solutions of $a$:

1. Negate $a$.
2. Add $\frac{\pi}{b}$ to $c$.

To account for values of the phase shift being added multiples of the period, we can replace $c$ with $\mod\left(c,\frac{2\pi}{b}\right)$ to account for multiples of $b$.

These two changes were applied to the results of the parameters such that comparison between parameter values across differences in phenomena by tweet content metric (polarity/subjectivity), distribution metric (mean/standard deviation), and candidate (Trump/Biden).

### 5.4 Miscellaneous Changes

As was found in section 4, data "during" the election (after November 1st) is more volatile than the "regular" cycle of tweets. Thus, for the sake of length (this paper attempts to cover too much already), sections 5 and 6 will explore data before November 1st.

It should be noted that all the phenomena may seem to span a rather small range, especially when moving average is implemented to smooth out noise. However, given the largeness of the dataset, even a small shift in the "center" or "spread" of the distribution of content metrics of tweets that demonstrates recurring patterns can be meaningful.

## 6 Modeling Polarity of Tweets Over Time

This section seeks to model the distribution of the polarity of tweets over time by candidate, using the methods and processes as articulated in section 5.

## 6.1 Mean

Means were taken for the polarity of tweets each hour. Figures 5.5 through 5.10 and figures 5.11 through 5.16 demonstrate sinusoidal fits overlaid with data smoothed by moving average method of windows 1 hour (nothing happens), 2 hours, 5 hours, 12 hours, 24 hours, and 48 hours by candidate. The parameters, period, MAE, and RMSE for each of the models are listed in Tables 6.1 and 6.2.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0093 | 0.7577 | 16.4693 | 0.0426 | 8.2921 | 0.0266 | 0.0365 |
| 2 | 0.0086 | 0.7578 | 10.3322 | 0.0426 | 8.2918 | 0.0201 | 0.0264 |
| 5 | 0.0101 | 0.0067 | 470.0552 | 0.0449 | 940.0930 | 0.0129 | 0.0158 |
| 12 | 0.0100 | 0.0067 | 466.5039 | 0.0446 | 932.9902 | 0.0089 | 0.0107 |
| 24 | 0.0094 | 0.0073 | 428.5540 | 0.0442 | 857.0892 | 0.0061 | 0.0075 |
| 48 | 0.0089 | 0.0079 | 397.6382 | 0.0438 | 795.2565 | 0.0038 | 0.0047 |

Table 6.1: Parameters and results for the mean of polarization of tweets relating to Trump.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0118 | 0.2709 | 0.3560 | 0.0713 | 23.1969 | 0.0354 | 0.0482 |
| 2 | 0.0117 | 0.2708 | 0.3463 | 0.0711 | 23.2003 | 0.0264 | 0.0359 |
| 5 | 0.0109 | 0.2703 | 0.3596 | 0.0712 | 23.2427 | 0.0186 | 0.0244 |
| 12 | 0.0078 | 0.1862 | 0.5300 | 0.0713 | 33.7362 | 0.0126 | 0.0155 |
| 24 | 0.0087 | 0.0522 | 0.2250 | 0.0714 | 120.3597 | 0.0071 | 0.0086 |
| 48 | 0.0057 | 0.0163 | 193.3154 | 0.0712 | 385.1710 | 0.0051 | 0.0063 |

Table 6.2: Parameters and results for the mean of polarization of tweets relating to Biden.

Interestingly, the sinusoidal fit for tweets relating to Trump with a window size of 1 and 2 in Figures 6.1 and 6.2 have very short, 8.3-hour periods. This is approximately $\frac{1}{3}$ of the day, so there may be some sort of pattern at play. However, there does not seem to be much visual corroboration for the sinusoidal fit. The encouragement of a low-period sinusoidal fit yields a small – perhaps negligible – amplitude, likely because a low-period high-frequency sinusoidal fit does not align with the patterns sufficiently. A small amplitude on already smoothed data on the mean of the actual phenomenon being modeled amounts to too many levels of aggregation and information loss. Given that the fits for a 1 hour and 2 hour window moving average smoothed data have such a small amplitude, the sinusoidal fits for 6.1 and 6.2 seem not to be modeling any actually sinusoidal behavior.

Figures 6.3 through 6.6 show the sinusoidal fits for longer-running trends based on data smoothed with larger window sizes. It should be noted that the smoothing of the data itself yields very interesting results, most pronounced in the 24-hour and 48-hour window smoothed data shown in figures 6.5 and 6.6. The data here shows clear patterns of rising and falling, although it is not sinusoidal in the sense of the curve falling as much as it rises. Moreover, these patterns of rising and falling seem to be located at relatively regularly spaced intervals. However, it seems that because of the upward trend – a feature the sinusoidal model cannot account for – the derived fit is not so much sinusoidal as it is an augmented line pointing upwards. Table 6.1 demonstrates clearly the drastic increase in the period of the derived sinusoidal fit from $\approx 8.2918$ hours with a window of 2 hours to $\approx 940.0930$ hours with a window of 5 hours. Correspondingly, the shift $c$ jumps from $\approx 10.3322$ hours to 470.0552 hours as the entire "character" or "spirit" of the sinusoidal drastically transforms.

Interestingly, isolating the four sinusoidal fits with window sizes 5, 12, 24, and 48 hours (these have similar "characters") yields three interesting trends across all six results.

- As window size increases, $a$ decreases.
- As window size increases, $b$ increases (period decreases).
- As window size increases, $c$ decreases.

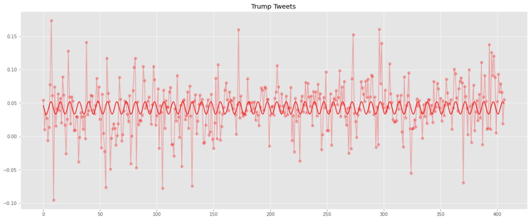

Figure 6.1: Trump Tweets, 1 hr. window.

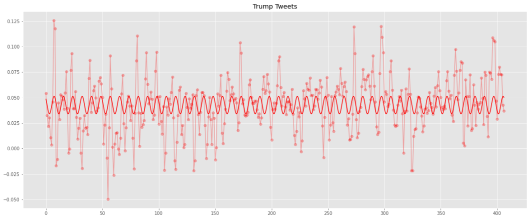

Figure 6.2: Trump Tweets, 2 hr. window.

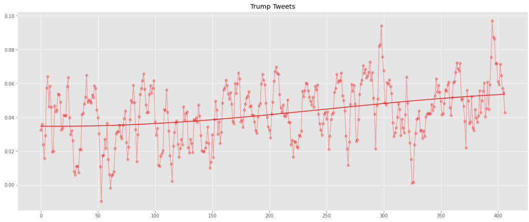

Figure 6.3: Trump Tweets, 5 hr. window.

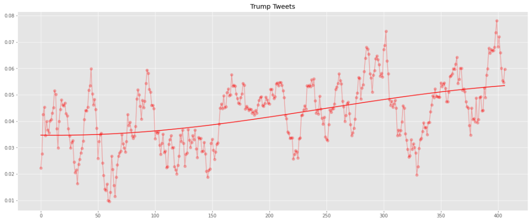

Figure 6.4: Trump Tweets, 12 hr. window.

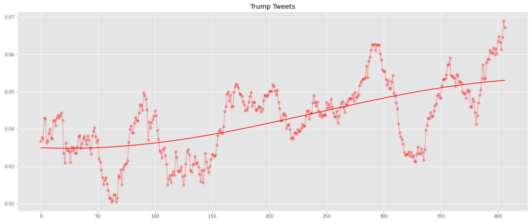

Figure 6.5: Trump Tweets, 24 hr. window.

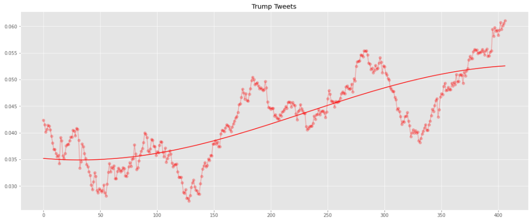

Figure 6.6: Trump Tweets, 48 hr. window.

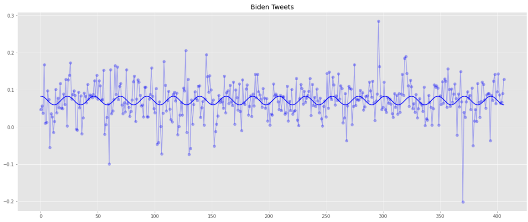

Figure 6.7: Biden Tweets, 1 hr. window.

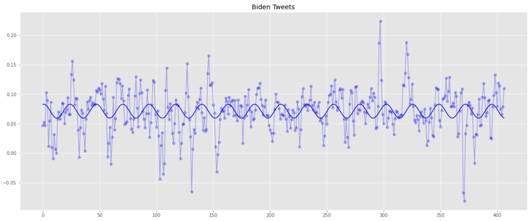

Figure 6.8: Biden Tweets, 2 hr. window.

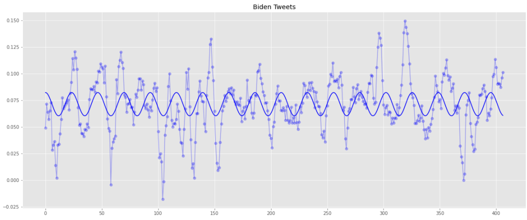

Figure 6.9: Biden Tweets, 5 hr. window.

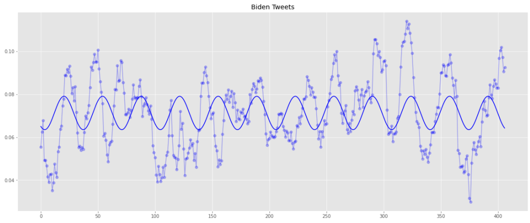

Figure 6.10: Biden Tweets, 12 hr. window.

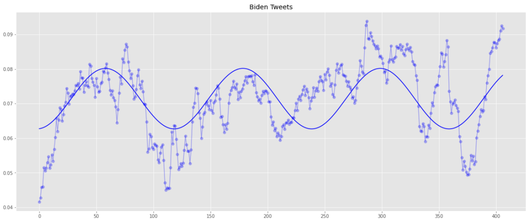

Figure 6.11: Biden Tweets, 24 hr. window.

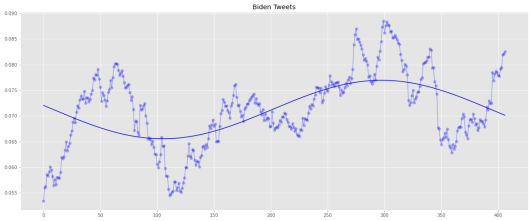

Figure 6.12: Biden Tweets, 48 hr. window.

- As window size increases, $d$ decreases.
- As window size increases, MAE decreases.
- As window size increases, RMSE decreases.

Of particular interest is the decrease in error as window size increases. As the window size increases, the sinusoidal fit can achieve a lower error. This is likely because greater smoothing of the data limits the range of the data and hence reduces the error a model that amounts to an augmented line makes. This will be expected with all results, so we will need to interpret RMSE with a grain of salt, understanding that a lower RMSE does not necessarily indicate a better sinusoidal fit – it is likely the result of a larger moving average window size reducing the range and hence the error of a model, even if the model itself hasn't "discovered" any new profound insights about sinusoidal curvature or some other pattern. This highlights the importance of visual analysis in guiding our conclusions.

Patterns for the smoothed mean of polarity of tweets relating to Biden, as displayed in Figures 6.7 through 6.12, are noticeably more conducive to "nontrivial" sinusoidal models (a "trivial" sinusoidal model being one whose "sinusoidality" is not exploited, like the derived model in Figure 6.5). These models, which are able to capture a sinusoidal pattern that models on tweets relating to Trump, have a worse RMSE (see Table 6.2), but this is to be expected.

The sinusoidal fits for tweets relating to Biden with a window size of 1, 2, and 5 hours are fascinating in that their periods are 23.1969, 23.2003, and 23.2427 hours, respectively. The sinusoidal models certainly seem to match tweets relating to Biden much more closely and meaningfully than tweets relating to Trump in that clusters of lower values coincide approximately with the minimums of the predicted fit and clusters of higher values coincide approximately with the maximums. Moreover, their solutions for the $a$, $b$, $c$, and $d$ parameters are very close, further corroborating the existence of such a phenomenon. That the periods are so close to 24 hours suggests that the polarity of tweets relating to Biden oscillates on a daily cycle.

Figure 6.11 is also especially interesting in terms of the smoothed data, which forms three very smooth and clear cycles. This sinusoidal fit has a period of 120 hours, or about 5 days. It also seems, therefore, that there is an oscillation in the polarity of tweets relating to Biden that rises and falls every five days.

It should also be noted that all the means ($d$ parameter) for sinusoidal fits for tweets on Biden are significantly higher than means for sinusoidal fits for tweets on Trump, suggesting that generally discourse on Biden was more positive than discourse on Trump.

There are many more observations one could extract, but perhaps the most important for this section is the existence of well-modelable patterns in tweets relating to Biden and the lack of such clear patterns in tweets relating to Trump. One possible attribution of this is to the volatility and relative controversy of Trump as a candidate compared to Biden.

## 6.2 Standard Deviation

The standard deviation represents the spread of the distribution of polarity for tweets relating to some candidate over time. A distribution that is more centered and dense in the middle will have a smaller standard deviation than one that is more dissipated and uniform.

Figures 6.13 to 6.18 show sinusoidal fits for the standard deviation of the polarity of tweets relating to Trump for different window sizes. Figures 6.19 to 6.24 show the same with tweets relating to Biden. Tables 6.3 and 6.4 list sinusoidal fit information for models fit to data of different candidates and window sizes.

A broad visual inspection of figures 6.13 to 6.24 shows that the standard deviation seems to be generally difficult to model. It seems, however, that the standard deviation of tweets relating to Trump is more modelable than the standard deviation of tweets relating to Biden. Visually, the sinusoidal fits for tweets relating to Trump seem more "in sync" with the rise and fall of the data than for tweets relating to Biden.

Another interesting phenomenon is that both the mean ($d$, see Tables 6.3 and 6.4) and the amplitude ($a$) of the sinusoidal fit for tweets relating to Trump are higher than the mean of the sinusoidal fit for tweets relating to Biden, with a few minor exceptions. This supports the previous finding that tweets relating to Trump are more volatile than tweets relating to Biden.

The derived periods for the sinusoidal fit are also interesting in that there emerges sets of well-corroborated period values. We also observed this in the prior exploration of the mean polarity. Four of the six models fitted

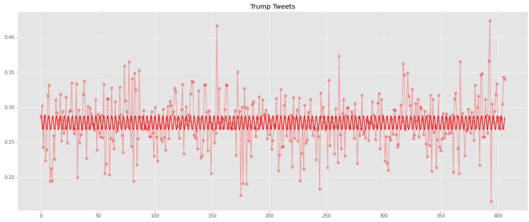

Figure 6.13: Trump Tweets, 1 hr. window.

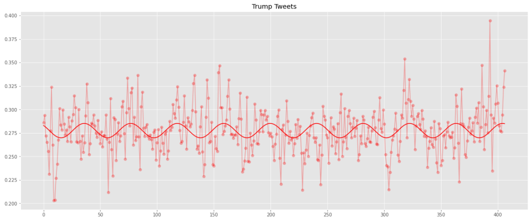

Figure 6.14: Trump Tweets, 2 hr. window.

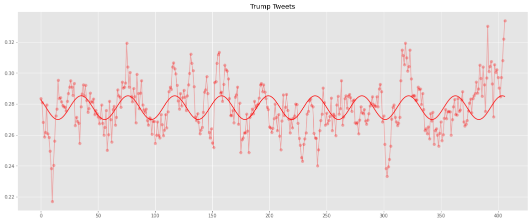

Figure 6.15: Trump Tweets, 5 hr. window.

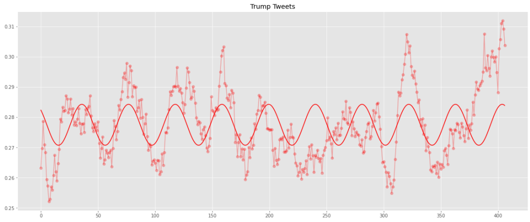

Figure 6.16: Trump Tweets, 12 hr. window.

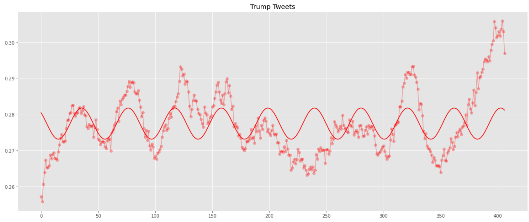

Figure 6.17: Trump Tweets, 24 hr. window.

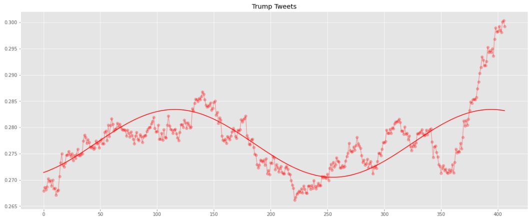

Figure 6.18: Trump Tweets, 48 hr. window.

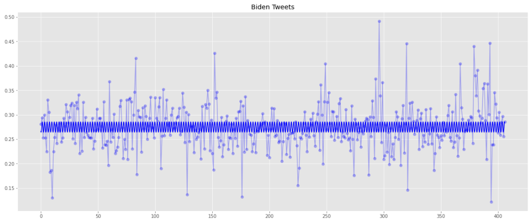

Figure 6.19: Biden Tweets, 1 hr. window.

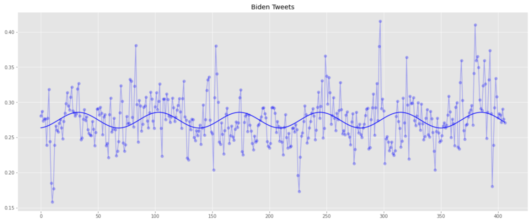

Figure 6.20: Biden Tweets, 2 hr. window.

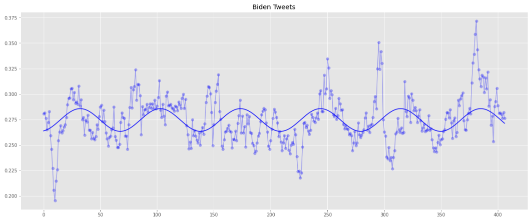

Figure 6.21: Biden Tweets, 5 hr. window.

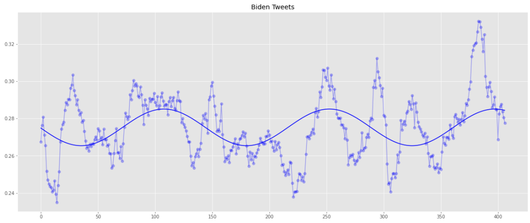

Figure 6.22: Biden Tweets, 12 hr. window.

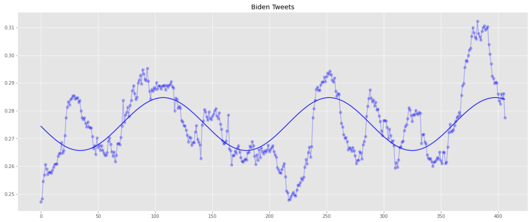

Figure 6.23: Biden Tweets, 24 hr. window.

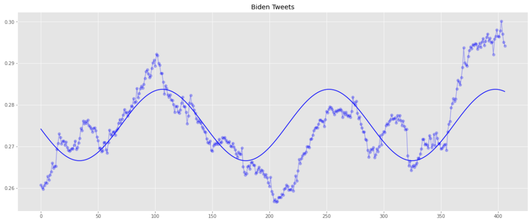

Figure 6.24: Biden Tweets, 48 hr. window.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0099 | 1.8247 | 9.5991 | 0.2775 | 3.4435 | 0.0274 | 0.0360 |
| 2 | 0.0077 | 0.1536 | 21.6766 | 0.2775 | 40.9193 | 0.0185 | 0.0243 |
| 5 | 0.0077 | 0.1535 | 21.8276 | 0.2776 | 40.9332 | 0.0114 | 0.0148 |
| 12 | 0.0068 | 0.1537 | 21.6456 | 0.2775 | 40.8713 | 0.0084 | 0.0104 |
| 24 | 0.0043 | 0.1541 | 21.6557 | 0.2775 | 40.7808 | 0.0061 | 0.0081 |
| 48 | 0.0065 | 0.0226 | 0.1368 | 0.2770 | 278.5876 | 0.0032 | 0.0045 |

Table 6.3: Parameters and results for the standard deviation of the polarization of tweets relating to Trump.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0105 | 2.9981 | 4.6075 | 0.2749 | 2.0958 | 0.0349 | 0.0471 |
| 2 | 0.0112 | 0.0891 | 35.3790 | 0.2745 | 70.5087 | 0.0243 | 0.0330 |
| 5 | 0.0111 | 0.0891 | 35.3925 | 0.2745 | 70.5331 | 0.0162 | 0.0217 |
| 12 | 0.0098 | 0.0434 | 72.4241 | 0.2752 | 144.7610 | 0.0123 | 0.0160 |
| 24 | 0.0095 | 0.0433 | 72.6742 | 0.2752 | 145.2608 | 0.0087 | 0.0113 |
| 48 | 0.0086 | 0.0432 | 72.8307 | 0.2752 | 145.5736 | 0.0055 | 0.0071 |

Table 6.4: Parameters and results for the standard deviation of the polarization of tweets relating to Biden.

on tweets relating to Trump, for instance, had a period of around 41 hours, or about 1.7 days. The periods of the sinusoidals (displayed in Figures 6.14 to 6.17) seem to be convincing, especially in Figure 6.15 with a 5-hour window. Similarly, for sinusoidals fitted on tweets relating to Biden, two of six (window sizes 2 and 5 hours) had a period of around 70.5 hours (about 2.9 days) and three of six (window sizes 12, 24, 48) had a period of around 145 hours (about 6 days). Interestingly, this period is almost exactly half of the period derived for models of window sizes 2 and 5 hours.

Moreover, both for models fitted on tweets relating to Trump (table 6.3) and tweets relating to Biden (table 6.4), there is a drastic change in the predicted sinusoidal as the window size increases slightly from 1 hour to 2 hours. A similar phenomenon was observed in section 6.1 in table 6.1 for tweets relating to trump, in which an increase in the smoothing window from 2 to 5 hours increased the period of the predicted sinusoidal from 8.29 to 940.09 hours. This functions as a measure of how volatile the series is; for a two-hour window to yield a drastically different result than a one-hour window (no change), the distance between adjacent data points must be generally so large such that their average yields a result very different from the original dataset. That the standard deviation - itself a measure of spread and perhaps volatility - is subject to extreme volatility on the scale of the hour highlights the inherent chaos in the phenomenon of investigation here.

## 6.3 Summary and Hypotheses

In this section, we investigated the distribution of the polarity of tweets relating to Trump and Biden via the lens of the mean and the standard deviation. We made several findings:

- The mean polarity of tweets relating to Trump could not be modeled non-trivially with the current methods, likely because there was an underlying trend that could not be captured by the sinusoidal model in its current form.
- The mean polarity of tweets relating to Biden was found to have two likely sinusoidal patterns: roughly 1-day period (corroborated by 3/6 models) and a roughly 5-day period.
- As window size decreases, the error metrics (MAE and RMSE) naturally decreases because the range of the data decreases. Thus, these error metrics must be interpreted with caution.
- The mean ($d$ parameter) of sinusoidal models fitted to the mean polarity was general higher for tweets relating to Biden than for tweets relating to Trump, suggesting that discourse on Biden was more positive in polarity than discourse on Trump.
- The mean ($d$ parameter) of sinusoidal models fitted to the standard deviation of polarity was generally higher for tweets relating to Trump than for tweets relating to Biden.

- The standard deviation of polarity for tweets relating to Trump was found to have one likely non-trivial sinusoidal pattern with a period of about 41 hours (corroborated by 4/6 models).
- The standard deviation of polarity for tweets relating to Biden was found to have two likely sinusoidal patterns with a period of about 70.5 hours (corroborated by 2/6 models) and 145 hours (corroborated by 3/6 models).
- For the standard deviation of polarity of tweets across both candidates, there were drastic differences between the periods of the sinusoidals found for unsmoothed data (window size 1 hour) and data smoothed with a window size of 2 hours, suggesting that the standard deviation of polarity of tweets is very volatile on an hourly basis.

Despite the appearance of similarities in the polarity across candidates, we found many interesting expected and unexpected results. That discourse on Biden was more positive than Trump is to be expected; Biden won the popular vote in the 2020 election, and Twitter users are more likely to support the Democratic Party than the Republican Party. A Pew Research Study found that 60% of Twitter users lean towards the Democrat party, whereas 35% of Twitter users lean towards the Republican party.

The periods of the sinusoidal patterns found are difficult to explain, but it's also difficult to deny their existence.

# 7 Modeling Subjectivity of Tweets Over Time

Subjectivity, as outlined in section 3 on data exploration and inquiry formulation, concerns how factual or opinionated the content of a tweet is. Note that this does not indicate the actual truth of the content of the tweet, and thus so indicates more the nature of the expression of tweet. Recall that 0 indicates "extremely factual" and 1 indicates "extremely opinionated".

## 7.1 Mean

The mean subjectivity of tweets relating to Trump are displayed in figures 7.1 to 7.6; the mean subjectivity of tweets relating to Biden are displayed in figures 7.7 to 7.12. The corresponding information for the derived models is listed in tables 7.1 and 7.2.

From a visual inspection, it seems that this layout is the "opposite" of the layout in section 6.1 (modeling the mean of the polarity of tweets over time). In Section 6.1, the fits for most of the tweets relating to Trump were trivial, whereas those for tweets relating to Biden were not. Here, in section 7.1, we see the opposite: many of the fits for tweets relating to Biden, particularly for data smoothed with a window of five hours or larger (figures 7.9 to 7.12) seem to be somewhat trivial. Trivial models are not necessarily unhelpful, though. Particularly, the fits for 48-hour window tweets relating to Trump and Biden appear very similar; their parameters are also very close – notably, periods of about 349 hours and 329 hours, respectively; on such a large scale, that the periods would be so close is interesting. This suggests that the general trend of the subjectivity of tweets was similar across candidates. when smoothed out with a large window size, the data for Trump and Biden have remarkably similar shapes (see figures 7.6 and 7.12). The existence of this similarity affirmed by the high positive correlations between the polarity across candidate, as well as the subjectivity across candidate, as found in section 3 during data exploration.

Incredibly, five sinusoidal models (window sizes 1, 2, 5, 12, 24) for tweets on Trump arrived at almost the same model, with a period of about 70 hours, or about 2.9 days. That five models of drastically different window sizes all arrive at the same conclusion suggests that the mean subjectivity of tweets relating to Trump is a very robust phenomena.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0140 | 0.0895 | 35.1951 | 0.3469 | 70.2103 | 0.0343 | 0.0453 |
| 2 | 0.0140 | 0.0895 | 62.7221 | 0.3469 | 70.1788 | 0.0257 | 0.0333 |
| 5 | 0.0139 | 0.0895 | 35.1941 | 0.3469 | 70.2083 | 0.0175 | 0.0222 |
| 12 | 0.0134 | 0.0894 | 35.2288 | 0.3469 | 70.2784 | 0.0125 | 0.0155 |
| 24 | 0.0117 | 0.0893 | 35.2831 | 0.3468 | 70.3865 | 0.0097 | 0.0119 |
| 48 | 0.0105 | 0.0180 | 0.6587 | 0.3461 | 348.6673 | 0.0054 | 0.0070 |

Table 7.1: Parameters and results for the mean of the subjectivity of tweets relating to Trump.

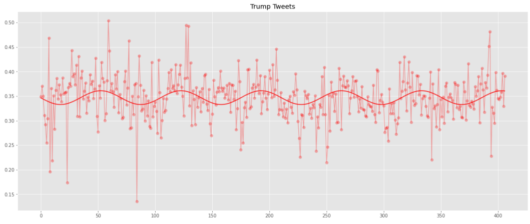

Figure 7.1: Trump Tweets, 1 hr. window.

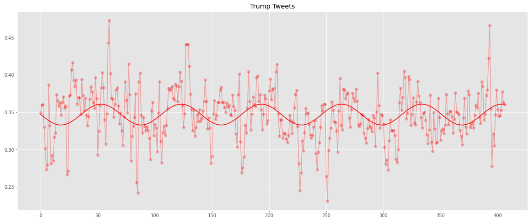

Figure 7.2: Trump Tweets, 2 hr. window.

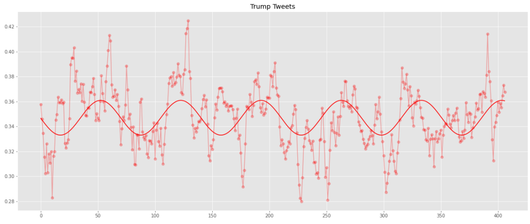

Figure 7.3: Trump Tweets, 5 hr. window.

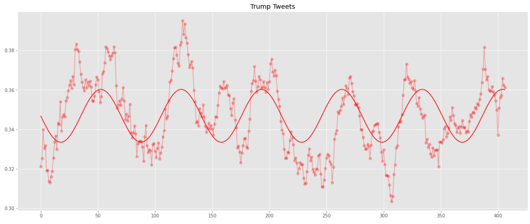

Figure 7.4: Trump Tweets, 12 hr. window.

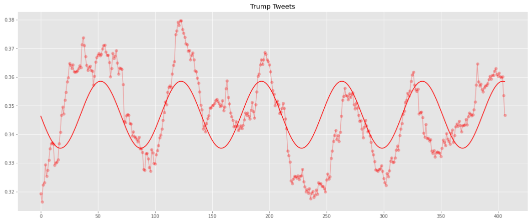

Figure 7.5: Trump Tweets, 24 hr. window.

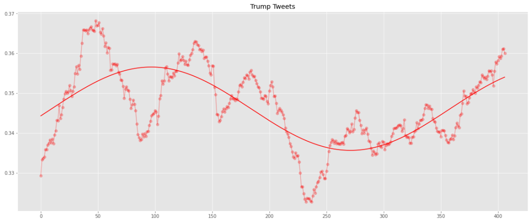

Figure 7.6: Trump Tweets, 48 hr. window.

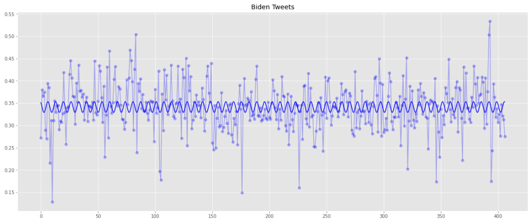

Figure 7.7: Biden Tweets, 1 hr. window.

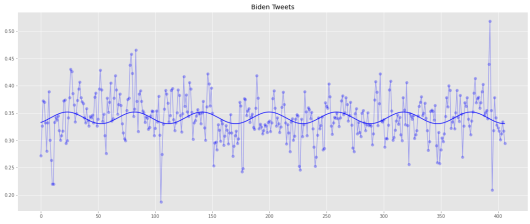

Figure 7.8: Biden Tweets, 2 hr. window.

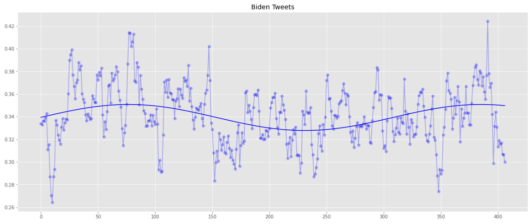

Figure 7.9: Biden Tweets, 5 hr. window.

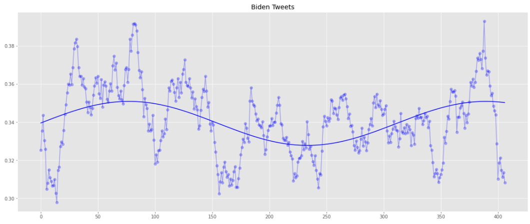

Figure 7.10: Biden Tweets, 12 hr. window.

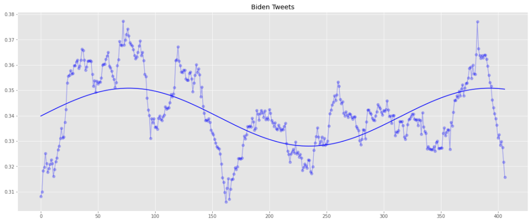

Figure 7.11: Biden Tweets, 24 hr. window.

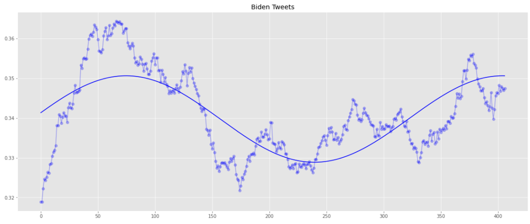

Figure 7.12: Biden Tweets, 48 hr. window.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0120 | 0.9277 | 9.3964 | 0.3412 | 6.7727 | 0.0414 | 0.0541 |
| 2 | 0.0107 | 0.1067 | 0.7446 | 0.3411 | 58.8805 | 0.0291 | 0.0389 |
| 5 | 0.0115 | 0.0204 | 0.0204 | 0.3393 | 308.6041 | 0.0196 | 0.0246 |
| 12 | 0.0116 | 0.0201 | 0.0202 | 0.3394 | 312.3177 | 0.0134 | 0.0169 |
| 24 | 0.0114 | 0.0199 | 0.0200 | 0.3394 | 315.6470 | 0.0101 | 0.0122 |
| 48 | 0.0109 | 0.0191 | 0.0195 | 0.3397 | 329.4072 | 0.0059 | 0.0073 |

Table 7.2: Parameters and results for the mean of the subjectivity of tweets relating to Biden.

Although it will be difficult to meaningfully compare the amplitude of sinusoidals fit on the same "root" data smoothed with different window sizes because the amplitude will be expected to get smaller, we can compare the amplitude of sinusoidal fits for tweets across candidate as long as the sinusoidals are trained on data with the same window size. It seems that generally the amplitude of the mean subjectivity is larger than that of Biden's. Because many of the derived sinusoidals for the mean subjectivity of Biden's data are semi-trivial, however, the amplitude may not be meaningful to compare.

## 7.2 Standard Deviation

The standard deviation for tweets relating to Trump are displayed in Figures 7.13 to 7.18; the standard deviations for tweets relating to Biden are displayed in Figures 7.19 to 7.24. The corresponding results for each model are listed in tables 7.3 and 7.4.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0052 | 2.5775 | 3.8240 | 0.3121 | 2.4377 | 0.0159 | 0.0209 |
| 2 | 0.0046 | 0.9103 | 24.8426 | 0.3120 | 6.9025 | 0.0112 | 0.0148 |
| 5 | 0.0040 | 0.3243 | 11.5171 | 0.3121 | 19.3726 | 0.0068 | 0.0088 |
| 12 | 0.0027 | 0.1362 | 30.1609 | 0.3121 | 46.1231 | 0.0040 | 0.0049 |
| 24 | 0.0026 | 0.0555 | 56.6240 | 0.3121 | 113.1150 | 0.0024 | 0.0029 |
| 48 | 0.0021 | 0.0562 | 55.9772 | 0.3121 | 111.8227 | 0.0012 | 0.0015 |

Table 7.3: Parameters and results for the standard deviation of the subjectivity of tweets relating to Trump.

| Window Size | A | B | C | D | Period | MAE | RMSE |
|---:|---:|---:|---:|---:|---:|---:|---:|
| 1 | 0.0073 | 2.0551 | 3.6910 | 0.3098 | 3.0573 | 0.0197 | 0.0272 |
| 2 | 0.0059 | 0.7897 | 4.8370 | 0.3097 | 7.9560 | 0.0138 | 0.0190 |
| 5 | 0.0055 | 0.1067 | 3.3378 | 0.3097 | 58.8912 | 0.0083 | 0.0109 |
| 12 | 0.0052 | 0.1067 | 2.7120 | 0.3097 | 58.9080 | 0.0049 | 0.0064 |
| 24 | 0.0043 | 0.1063 | 3.8589 | 0.3097 | 59.1275 | 0.0033 | 0.0045 |
| 48 | 0.0025 | 0.0334 | 94.2179 | 0.3097 | 188.0248 | 0.0023 | 0.0029 |

Table 7.4: Parameters and results for the standard deviation of the subjectivity of tweets relating to Biden.

The fits for the standard deviation of the subjectivity of models seem for Trump tweets seem to be varied across window sizes. The fits for 7.13 (1 hour window) and 7.14 (2 hour window) have a frequency that seems too high to be meaningful. The fits for 7.15 (5 hour window) and 7.16 (12 hour window) also do not seem consistent enough. However, there does seem to be a meaningful trend with a period of approximately 112-113 hours (about 4.6 days) corroborated by models of window size 24 and 48 hours. Three sinusoidal models with window sizes 5, 12, and 24 (corresponding to figures 7.21 to 7.23)for tweets relating to Biden corroborate a period of about 59 hours, or approximately 2.5 days.

The amplitude for sinusoidals modeling tweets relating to Biden is always higher than the amplitude for the corresponding sinusoidal (by window size) for Trump. This might suggest that tweets on Biden are more "volatile" than tweets on Trump. Here, we need to reconcile several notions of volatility that we have explored throughout this paper:

- *Standard deviation.* A higher standard deviation indicates a larger spread within an interval and therefore is "volatile" in the sense of spreading a large range, rather than being concentrated near one location.
- *Amplitude.* A larger amplitude indicates that the sinusoidal pattern fluctuates through a larger range of points, suggesting that the phenomena is less stable in the sense of oscillating with a smaller range.
- *Difficulty to be modeled by sinusoidal function.* The sinusoidal function itself is a statement on how predictable the phenomenon itself is – it should rise and fall similar heights at similar spaced intervals. Data cannot be modeled well by a sinusoidal function could be said to be volatile in its unpredictability.
- *Difference between derived models of adjacent window sizes.* As was established prior, another measure of volatility can be derived from analyzing the differences in the model parameters for different data when the window size for data smoothing is slightly changed, like from 1 hour to 2 hour. In this interpretation of volatility, volatility is the opposite of "robustness" to changes in window size.

In this context, we are borrowing from the first and second interpretations of volatility, in that the amplitude of the sinusoidal model modeling the standard deviation of the subjectivity. That is, the quantity of volatility is itself volatile, but in a relatively predictable way.

### 7.3 Summary and Hypotheses

In this section, we investigated the distribution of the polarity of tweets relating to Trump and Biden via the lens of the mean and the standard deviation. We made several findings:

- The means of subjectivity yields the "opposite" results by candidate than the means of the objectivity in that tweets on Trump can be modeled with nontrivial sinusoidals but tweets on Biden return relatively trivial sinusoidals.
- The general trend of the mean subjectivity of tweets was similar across candidates (when smoothed with a large window size).
- The mean subjectivity of tweets relating to Trump was found to have one likely sinusoidal pattern with a period of approximately 70 hours (about 2.9 days), corroborated by five models.
- The standard deviation of subjectivity of tweets relating to Biden was found to have one likely sinusoidal pattern with period 59 hours (about 2.5 days).
- The amplitude for sinusoidals fitted to tweets relating to Biden is higher than the corresponding sinusoidal for tweets relating to Trump.

Explaining this phenomenon is a tall order. An explanation will need to both take into account that the general trend of the mean subjectivity is very similar across candidates and that there are many differences in the derived periods and amplitudes for sinusoidals across candidates. This highlights the layers and complexity of this data. One possible explanation for the similarities in the shape of any phenomena across two candidates when sufficiently smoothed out (e.g. 48 hours) being similar while lower-level details of the standard deviation and mean differing at smaller period scales is that the large, general trend of the data is informed more by the environment than the candidate and the smaller, higher-frequency patterns are informed more by the candidate than the environment.

## 8 Key Findings

This paper found three key general findings:

- The quantity of tweets released every hour is sinusoidal in character, and has a period of one day. During the election (dates after November 1st), tweets relating to Trump deviated more from the predicted cycle than tweets relating to Biden.

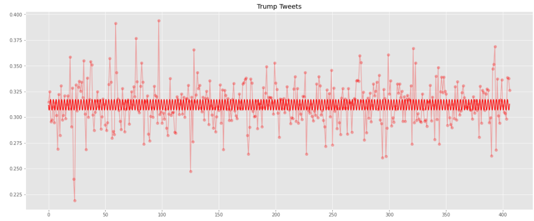

Figure 7.13: Trump Tweets, 1 hr. window.

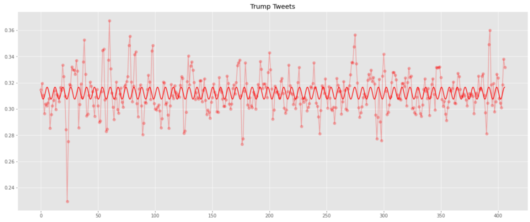

Figure 7.14: Trump Tweets, 2 hr. window.

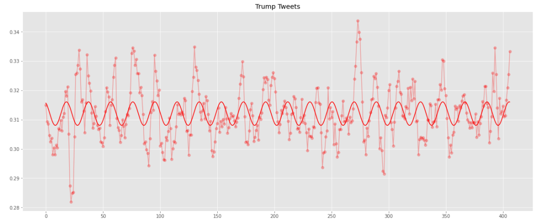

Figure 7.15: Trump Tweets, 5 hr. window.

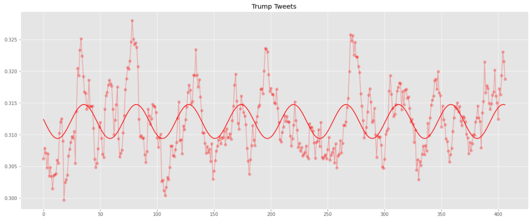

Figure 7.16: Trump Tweets, 12 hr. window.

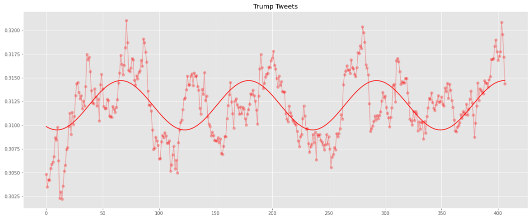

Figure 7.17: Trump Tweets, 24 hr. window.

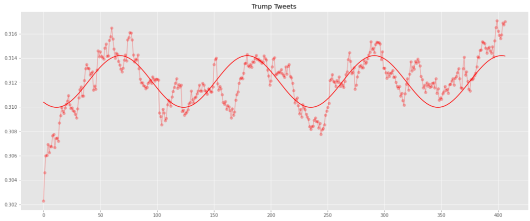

Figure 7.18: Trump Tweets, 48 hr. window.

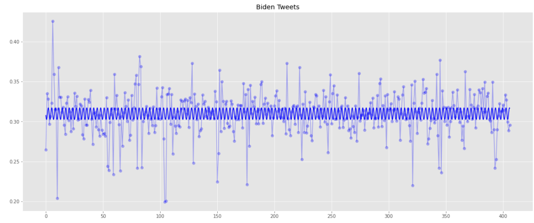

Figure 7.19: Biden Tweets, 1 hr. window.

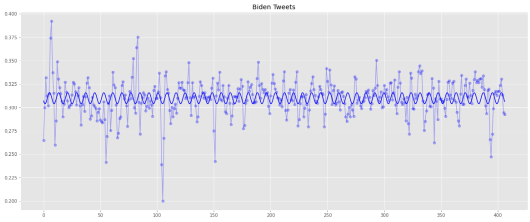

Figure 7.20: Biden Tweets, 2 hr. window.

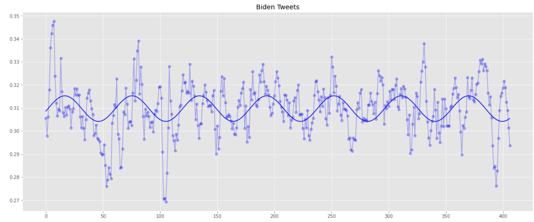

Figure 7.21: Biden Tweets, 5 hr. window.

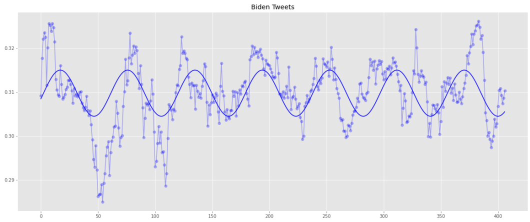

Figure 7.22: Biden Tweets, 12 hr. window.

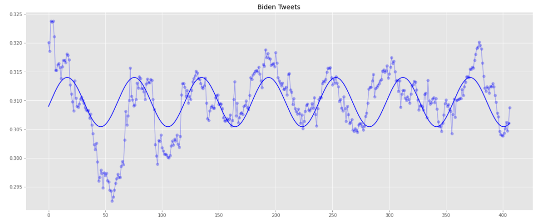

Figure 7.23: Biden Tweets, 24 hr. window.

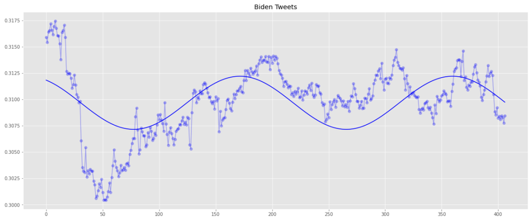

Figure 7.24: Biden Tweets, 48 hr. window.

- There exist many fitting sinusoidal models for the polarity and subjectivity of tweets relating to Trump and Biden. Given that there are multiple interpretations of what “volatility” entails, the results demonstrate that tweets on both Trump and Biden are volatile in different ways.
- If several sinusoidal models fitted on the same “root” data smoothed with different window sizes arrive at the same curve, then it is likely that the phenomena is “real” or “honest”.

# 9 Further Inquiry

## 9.1 Expanding and Improving the Processes of This Paper

This paper, for the sake of length, left out many paths of inquiry in its processes. The following are directions for further inquiry that directly expand upon shortcomings of this paper.

- In several instances a phenomena was composed of both an oscillating pattern and an underlying trend, like a $u$-shape or a line. Further inquiry would model such phenomena using models composed of both a trend component and a sinusoidal component.
- This paper partially sought to analyze the distribution of polarity and subjectivity across time, but investigated only the mean and the standard deviation. There are, however, many more attributes of distributions are worth investigating, like skewness and kurtosis.
- This paper evaluated the goodness (absolute and relative) of models via the mean average error and the root mean squared error. However, these metrics were difficult to utilize given the methods used of smoothing via windows of different sizes.
- It would be interesting to see if the RMSE could be utilized, even given the range-shrinking phenomenon of moving average smoothing. Further inquiry would find a way to verify the “goodness” of a sinusoidal model on data given both the RMSE and the window size. For instance, comparing how much the RMSE would decrease if the data were smoothed and the model did not increase in “goodness” to the actual decrease to see if the model actually decreased the error meaningfully.

## 9.2 Beyond the Processes of This Paper

The following are paths of inquiry that further explore the broad phenomena discussed as the topic of this paper, but do not stem from the methods that have been proposed and discussed.

- In this paper, we looked for whether there were sinusoidal trends in the data. This could be better put into context by comparing sinusoidal models (or models formed as sums of sinusoidals and other models) to models like polynomial fits, linear fits, sigmoid fits, and so on.
- Further inquiry would improve the sophistication of the metrics used for sentiment analysis. Polarity and subjectivity, as found by `TextBlob`, are rather rudimentary and hence do not reveal the true richness of the content of the data.
- Further inquiry would explore whether sinusoidal patterns can simply appear by chance. For instance, a graph generated with each step adding a random number selected from $(-n, n)$ to its previous location will generally look vaguely sinusoidal. Understanding this does not diminish the existence of sinusoidal patterns in the data, but reframes the phenomena in a different and important context.

## 9.3 Successes and Shortcomings of the “Half-Information” Iterated Method of Sinusoidal Fitting

As was noted in section 5.3, ”Automated Adjustment for Nonlinear Least Squares Fit”, the half-information iterated method was attempted, ultimately unsuccessfully. The “half-information” approach relies on having either the information about parameters outside the function or the information about parameters inside the function. In the case of $y = a\sin(b(x-c)) + d$, $a$ and $d$ must be known to find $b$ and $c$, and $b$ and $c$ must be known to find $a$ and $d$. To find $a$ and $d$ given $b$ and $c$, a transformed array $x_t$ is formed such that $x_t = \sin(b(x-c))$. $a$ and $d$ can be found by performing linear regression on $y = ax_t + d$. To find $b$ and $c$ given $a$ and $d$, $y$ must be transformed

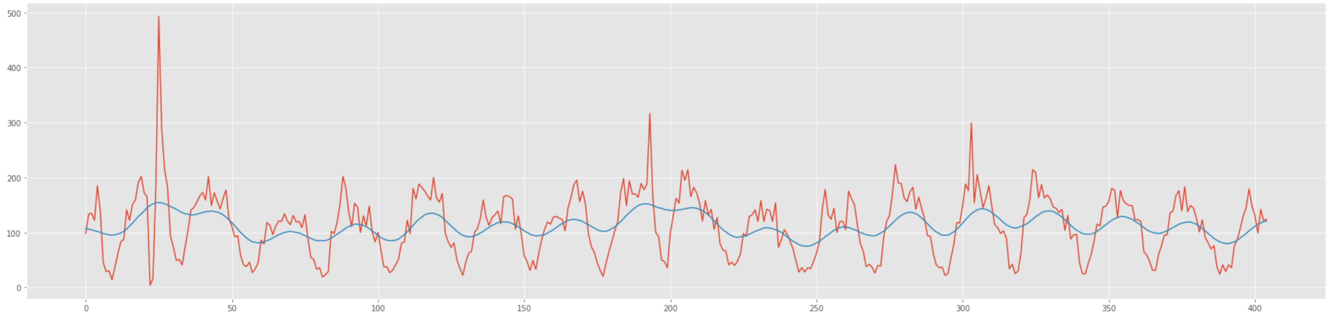

Figure 9.1: Number of tweets relating to Trump per hour in red, smoothed data in blue.

such that $y_t = \arcsin\left(\frac{y-d}{a}\right)$. $y_t$ is further altered either as $y_t + 2\pi n$ or $\pi - y_t + 2\pi$ based on the estimated phase shift and the period. $b$ and $c$ are thus found with linear regression on $y_t = b(x-c)$. One consideration is that $\frac{y-d}{a}$ must be between $-1$ and $1$ for all values of $y$. Although we can initialize $a$ and $d$ such that this requirement is true, there is no guarantee that it remains true after $a$ and $d$ are changed via several rounds of finding the other "half" of information. In the case that $\frac{y-d}{a}$ is not between $-1$ and $1$ for all values of $y$, which one would imagine a sinusoidal of best fit would gravitate towards on most datasets (some points would be above the highest point or below the lowest point of the sinusoidal), $a$ and $d$ must be adjusted such that $-1 \leq \frac{y-d}{a} \leq 1$ for all $y$. This would seem to "reset" progress somewhat. Because updating $b$ and $c$ is dependent on finding $a$ and $d$, it seems that the domain inflexibility of arcsine can "shut down" the iterative process.

Rough estimates of $b$ and $c$ are required for finding $b$ and $c$. There was success on this front. Autocorrelation,the correlation of a series with itself shifted by some lag quantity, allows for a fairly accurate detection of the period. A series with period $p$ will have a high correlation with itself shifted by $p$ units. To find the shift, it was found that three applications of moving averages with window sizes of half the period yielded a smooth curve whose derivative at all points was "honest" in indicating where the maxima and minima were. Without these applications of moving averages, there are many "fake" maxima and minima that an algorithm may mistake for a real peak (see figure 9.1). Once the data has been smoothed, we can calculate the shift by finding the distance between the starting location and the location halfway between the first minimum and the first maximum.

Further inquiry would investigate whether the problems with domain restrictions are fatal to using an iterated "half-information" method, or perhaps find a way around it.