Twitter Sentiment Analysis of COVID-19 in the State of Virginia

Final Report

Rachel Martonik
Northwestern University
MSDS 498 – Capstone Project
June 7, 2020

Table of Contents

Executive Summary

The COVID-19 pandemic is going to have long-lasting impacts on our public health, economy, and way of life. From the dramatic increase in unemployment to the subsequent loneliness and depression caused by social distancing measures, the mental health impact of COVID-19 will be substantial. The goal of this project is to create an auxiliary surveillance tool for the Virginia Department of Health (DOH) that uses social media data to track sentiment towards COVID-19 and the stay-at-home measures in place in the spring of 2020.

The goal is broadly exploratory, aiming to monitor sentiment towards the virus and the stay-at-home measures by time and location.

Twitter data was collected from April 7, 2020 through May 24, 2020 for two data sets using 1) COVID related keywords, and 2) stay-at-home related keywords. Data was limited to Virginia, using both the Twitter API, and subsequent cleaning and geolocation efforts. Only 11% of the Twitter data was geolocated to a county or city in Virginia and used for analysis. The frequency and sentiment of the Twitter data was analyzed over time and at the county/city level when appropriate. I also performed topic modeling on the stay-at-home data set.

The daily Tweet frequency declined in both data sets over the course of data collection. The COVID data set has a clear weekday vs weekend pattern, with sentiment and frequency both dropping dramatically over the weekend. The same pattern is not seen in the stay-at-home data set.

The COVID data set has an overall sentiment score of .073 (where 1 is totally positive and -1 is totally negative) for the course of data collection and is a politically charged. The mention of Donald Trump in a Tweet has a much larger effect on sentiment than do COVID statistics such as reported cases or deaths, or county-level percent urban population or unemployment rates. The stay-at-home data set had an overall sentiment score of .173 for the course of data collection and is much less political. In fact, Donald Trump is only mentioned in 1.35% of the Tweets compared to 2.45% of the Tweets mentioning haircuts or facial hair. The stay-at-home data is largely focused on life at home during the quarantine. Weak relationships exist between COVID deaths (cumulative weekly), percent urban, and unemployment and average weekly sentiment at the county/city level for this data set.

Despite the geolocation issues and overall noisiness of the data, I was able to track the COVID and stay-at-home related sentiment statewide and by county for those that had enough data. I was able to detect spikes in sentiment and pinpoint the reasons behind them using natural language processing methods. However, the project revealed that simply using keywords to capture Tweets on a certain topic does not necessarily provide a clean measure of sentiment toward that topic. More work would be needed, but using lessons learned from this research, the VA DOH could set up a successful public health Twitter surveillance framework for COVID, perhaps to gather information about family coping methods if schools don't open or attitudes towards a vaccine.

Introduction

The United States is in disarray that started with the onset of COVID-19, an extremely contagious respiratory virus with an estimated mortality rate of 3-4% according to the World Health Organization[1]. The rate for at-risk populations—elderly, and those with underlying conditions such as heart disease, diabetes, asthma, etc.— is much higher. Much of the country was shut down, and on March 30, 2020 Virginia Governor Ralph Northam declared a stay-at-home order until June 10, 2020. States across the country are grappling to design a testing and tracing infrastructure before opening the economy back up. Opinions on how and when to reopen the country are varied.

The COVID-19 pandemic is going to have long-lasting impacts on our public health, economy, and way of life. From the dramatic increase in unemployment to the subsequent loneliness and depression caused by social distancing measures, the mental health impact of COVID-19 will be substantial. For my Capstone project, I am using Twitter data from Virginia to track daily sentiment toward the virus and stay-at-home measures as the virus travels through the state. The goal of this project is to create an auxiliary surveillance tool for the Virginia Department of Health (DOH) that uses social media data to track sentiment towards COVID-19 and the stay-at-home measures in place.

Public Health Surveillance using Twitter Data

"Health organizations require accurate and timely disease surveillance techniques in order respond to emerging epidemics. Such information may inform planning for surges in patient visits, therapeutic supplies, and public health information dissemination campaigns." [2] However, real-time data collection is very expensive and simply not feasible for most organizations. As such, ongoing research continues into using social media for real-time disease and health surveillance.

Using social media data for public health surveillance is not a new idea, and in fact the Center for Disease Control and Prevention (CDC) has been using Twitter for influenza surveillance for almost a decade. While took them a few years to develop an algorithm that filtered out the "Twitter chatter" regarding the flu in general (in the news, etc.), during the 2012 flu season, they found that the number of weekly tweets indicating influenza infection was highly correlated with influenza infection rates. The new system could also detect the direction of change with an 85% accuracy rate during weeks with larger-than-average changes from the previous weeks [2].

According to the Pew Research Center [3], American Twitter users are younger, are more highly educated, have higher incomes and are more likely to identify as Democrats than the population overall. They also differ from the general population on some social issues such as immigration. Additionally they found that 10% of Twitter users produce 80% of the content. While the social media opinions and activity of Twitter users cannot be generalized to the Virginia population overall, the idea is to help Virginia health officials keep the pulse on public opinion regarding COVID and the stay-at-home measures using real-time data.

Project Goals

The goal of this project is not to use Twitter data to estimate COVID-19 prevalence using the frequency of COVID-related Tweets as the CDC has done with influenza surveillance, though I looked to see if COVID sentiment leads or lags reported cases. The goal is broadly exploratory, aiming to monitor sentiment towards the virus and the stay-at-home measures by time and location. In addition to highlighting areas in which sentiment is particularly low (and potentially triggering investigation by the VA DOH), the idea is to set up a framework for the VA DOH to use Twitter for surveillance for the remainder of the COVID lifespan. Right now, we are looking at stay-at-home measures, but when things fully reopen it could be identifying COVID resurgence hot spots (via Tweet frequency or sentiment), or measuring response towards the vaccine once available.

Methods

Data Collection

Data was pulled daily from the Twitter application user interface (API) using the `rtweet` package in R. Data collection was limited to tweets posted in Virginia according to Twitter. Two daily pulls were run:

1. **COVID** pull that uses the search terms: #COVID, COVID, COVID-19, #COVID-19, coronavirus, and #coronavirus. Data collection April 7th through May 22nd.
2. **Stay-at-home** pull that uses the search terms: #stayhome, #stayathome, #Quarantine, quarantine, and #SocialDistancing. Data collection April 9th through May 22nd.

COVID-19 county level data (confirmed cases and deaths) was retrieved from the New York Times GitHub page (https://github.com/nytimes/covid-19-data). Additional county-level demographic data was be pulled from American Community Survey via Census.gov, and unemployment data from the Bureau of Labor Statistics.

Data Cleaning

The goal of the project was to provide the VA DOH with low-level geographical data. A major challenge was filtering the data to this lower geographical level. Using user reported location (pulled from user profiles), I was able to geolocate 11% percent of the data overall to one of the 133 counties and independent cities in VA. See Appendix 1 for additional details on geolocation methods.

Original Tweets only were included in the analysis (no retweets.) In order to assess “individual” users as opposed to news outlets or serial Tweeters, I removed outlier users in regard to Tweet frequency. However, I did not specifically clean out news outlets or businesses from the data. I also removed duplicate Tweet messages. A total of 51,066 Tweets were removed from the COVID data set leaving 112,576 for analysis. A total of 5,156 Tweets were removed from the stay-at-home data set leaving 28,772 for analysis.

Sentiment Analysis using the VADAR Python Package

The VADER (Valence Aware Dictionary and sEntiment Reasoner) Python package is a lexicon and rule-based sentiment analysis tool that is specifically developed for sentiment analysis using social media. VADER has been found to be very good at sentiment analysis of social media data. Unlike other sentiment analysis packages, VADAR provides the portion of the text it finds positive, negative, and neutral and then provides a composite sentiment score [4]. Tweets can contain both positive and negative sentiment, for example:

While everything else is a sad disaster, I am excited to have an expanding set of post-COVID hair style options, now that my locks have grown out to a level I haven't seen since junior high. So there's that.

Negative: 0.1550; Neutral: 0.7940; Positive: 0.0520; Compound: -0.70

The "compound" score is the sentiment of the Tweet, where 1 is totally positive and -1 is totally negative.

Topic Modeling

Topic modeling is a type of unsupervised learning that uncovers "topics" within a text corpus. The topics are represented as a set of the most important words in that topic. The topics help bring structure to a vast amount of unstructured text. Latent Dirichlet Allocation (LDA) is a commonly used algorithm for topic modeling. LDA assumes that each document (Tweet) is a mixture of an arbitrary number of topics selected when training the LDA model. It also assumes that each topic can be represented by a distribution of words [5]. I used the Gensim topic modeling Python package to run the LDA model on the stay-at-home data set.

Visualizations and Code

As part of my final deliverable I am providing visualizations using Tableau Public. In this report I will reference visualizations that can be found here:

COVID data set:
https://public.tableau.com/profile/rachel.martonik#!/vizhome/MeanSentimentColoredbyCases/OverallDailySent?publish=yes

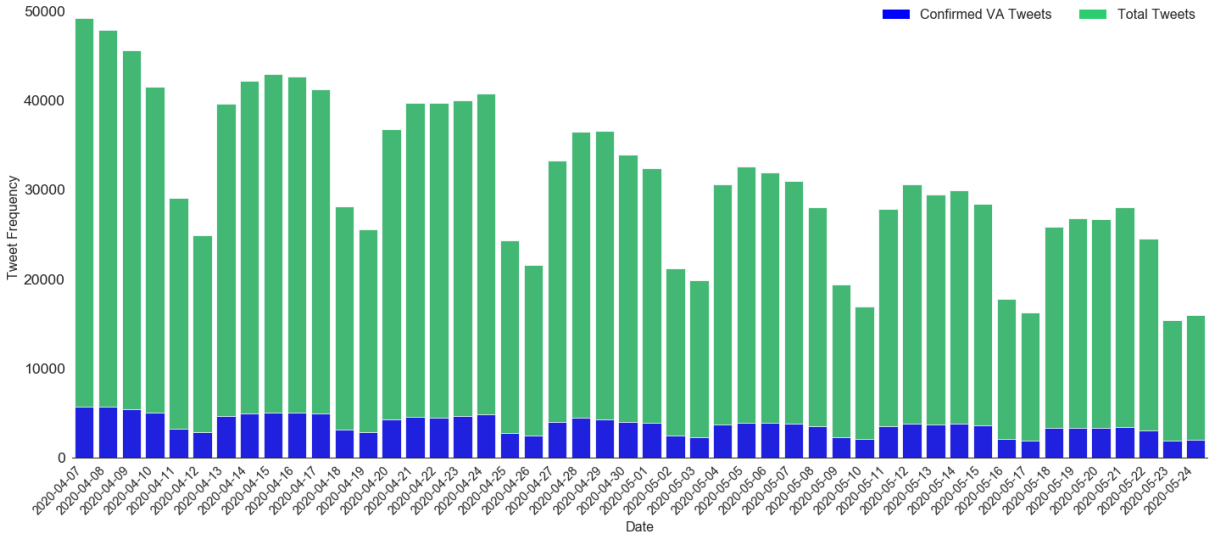
Stay-at-Home data set: https://public.tableau.com/profile/rachel.martonik#!/vizhome/Stay-At-Home/OverallDailySent?publish=yes

All analysis was run using Python Jupyter Notebooks. Data files and notebooks used for analysis can be found here: https://github.com/martonik/COVID-19

Results – COVID Data Set

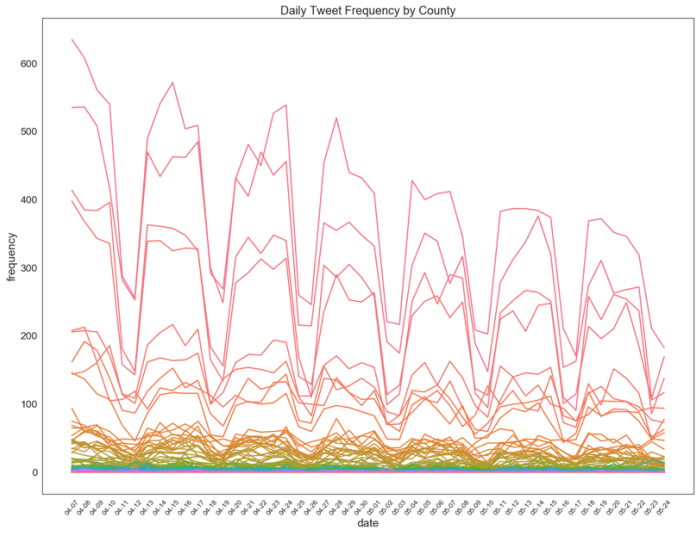
The figure below shows the frequency of COVID related Tweets overall and those confirmed as Virginia tweets.

Figure 1: COVID Related Tweet Frequency - Overall and Virginia



Both overall and in Virginia, the frequency of COVID Tweets drops dramatically over the weekend. There is also a clear downward trend in the number of Tweets, both over the weekends and weekdays. However, the rate of decline seems to be leveling off. The weekday vs weekend trend also appears at the county level and can be seen in the Figure 2 as well as tabs "DTF-lower" and "DTF-higher" visualizations in Tableau.

Figure 2: COVID Related Tweet Frequency - County



Weekend vs Weekday Users

In order to assess "individual" users as opposed to news outlets or serial Tweeters, I removed outlier users in regard to Tweet frequency. The highly uneven distribution of Tweets during the week vs the weekend suggests that there may be some users that are Tweeting as part of their normal work week. The table below shows the breakdown of usernames vs number of Tweets for weekend-only users, weekday-only users, and those that Tweet both on weekdays and weekends as of 5/14/20.

Table 1: Weekend vs Weekday Tweet Stats

Type of User	Unique User Names	% of Total User Names	Tweets	% of Total Tweets
Weekend & Weekday users	7,396	29%	53,578	52%
Weekday only users	14,480	57%	24,890	24%
Weekend only	3,711	15%	24,564	24%
Total	25,587	100%	103,032	100%

Those that Tweet on the weekend (whether it be weekends only or both weekdays and weekends) make up 44% of the users but produce 78% of the Tweets. This undermines the theory that a group of weekday users were disproportionately tweeting. However, note that by definition a user that has only Tweeted once is defined to the group according to which day they Tweeted. The table below shows the breakdown of single-time Tweeters vs those that have more than one Tweet as of 5/14/20.

Table 2: Single vs Multiple User Stats

Type of User	Unique User Names	% of Total User Names	Tweets	% of Total Tweets
Single Tweeters	13,127	51%	13,127	13%
Multiple Tweeters	12,460	49%	89,905	87%
Total	25,587	100%	103,032	100%

Almost half of the users are those with a single COVID tweet in our data set. The other half have at least two tweets and account for 87% of the Tweets. The average number of Tweets per user is 4, with a minimum of 1 and a maximum of 95.

Sentiment Analysis

The average sentiment score for all tweets from 4/7/20 through 5/24/20 is .073 where 1 is totally positive and -1 is totally negative. The table below shows a few of the most positive and negative tweets for reference.

Table 3: Negative and Positive COVID Tweets

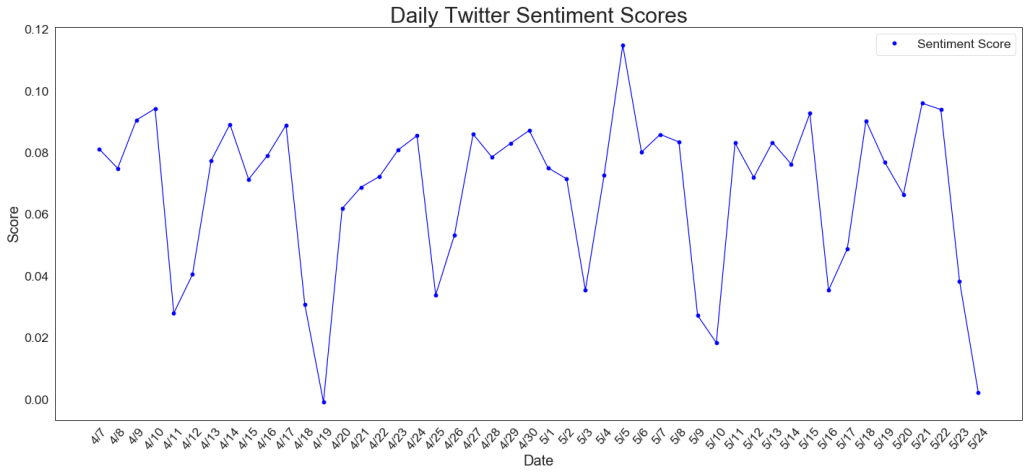
Sentiment Score	Tweet
-0.9947	Trump referring to himself as a "Wartime President" is pathetic. COVID-19 is bad. Worse health crisis in 100 years. WHY do we Americans insist on casting every challenge as "a war"? War on drugs, war on crime, war on poverty (really war on the poor), war on terror?
-0.9863	TO STOP TRYING TO SCREW THE AMERICAN PEOPLE OVER; DO THE REAL JOB SHE WAS ELECTED TO DO!! GOD

-0.9849	A note to all those who asked me "But whuddabout the people who die of the flu every year?": the worst US flu death toll in recent
0.9977	new series starts on Sunday "Love During the Coronavirus"\n\nspoiler alert: keep loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and loving and lov
0.9887	A dear friend of mine lost her father to COVID-19 on Sunday - a loving, gentle, joyful man whose life touched so many. In his honor, please fill this thread with kindness - a moment of hope you felt, a good deed you'll do, words of encouragement. Let's create a thread of love. ❤️
0.9831	Just WOW!!! What a very #kind gesture! The #appreciation from the #public and our @NELFT patients and communities shines daily. THANK YOU! \n\nPeople call us #NHSheroes but we serve our #communities to #care with #honour and #pride doing the job we love! #HCP #NHS #COVID #OneTeam https://t.co/xYnHVuVCI7

Virginia Statewide Sentiment

Figure 3 shows the statewide daily sentiment score (DSS) by date. Similar to the frequency shown in Figure 1, sentiment drops dramatically over the weekend. Sentiment is strongly correlated with the weekends with a Pearson's r of -0.865 and a p-value of <.001. See also "Overall Daily Sent" tab on Tableau. Average weekday sentiment was .0822 and average weekend sentiment was .0033 over the course of data collection.

Figure 3: Statewide Daily Sentiment Score (DSS)



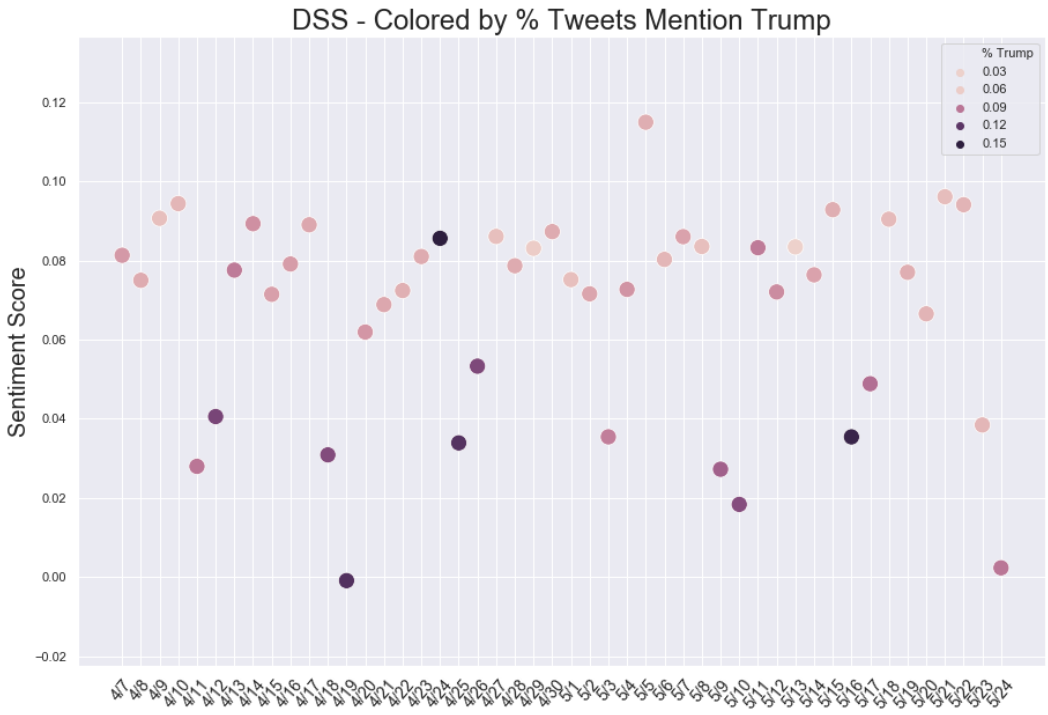
Despite the weekend dips, the plot shows that generally statewide DSS has remained stable over time with the exception of April 19th with a low score of .0016 and May 5th with a high spike of .113. Additionally, Wednesday, May 20th is low for a weekday with a score of .066 and Sunday, May 24th is low with a .0023.

The Trump Effect

In an effort to identify the cause for the sentiment drop on April 19th, I reviewed the most frequent words in the Tweets for that day. I noticed more references to President Trump (Trump, realdonaldtrump, president) than for other days. The percentage of daily Tweets that mention Donald Trump in some capacity is negatively correlated with sentiment. The more Tweets that reference Trump, the lower the DSS with a Pearson's r of -.654 statistically significant with a p-value of <.001. Figure 4 shows DSS colored by percentage of daily Tweets

that mention Trump. It is worth mentioning that the same correlation does not exist between Virginia Governor Ralph Northam and sentiment score.

Figure 4: Statewide Daily Sentiment Score (DSS) – Trump Effect



Twelve percent of the Tweets on April 19th mention Trump—which is the third highest day—and comes on the heels of Trump's controversial April 17th Tweets about "liberating states" on lockdown, including Virginia:

LIBERATE VIRGINIA, and save your great 2nd Amendment. It's under siege!

Anecdotal evidence suggests the response to Donald Trump over a weekend with fewer Tweets overall and lower weekend sentiment scores caused this large drop.

Figure 4 shows that while April 24th had a high proportion of Tweets mentioning Trump, sentiment score was not correspondingly negative. According to NBC News, "at an April 23 press briefing, President Donald Trump mused about the possibility of using "very powerful light" and injecting disinfectant into the body to kill COVID-19 – a suggestion that, in the case of disinfectant, was roundly criticized by experts as dangerous. A day later, he said he was being "sarcastic."[6]

Both "bleach" and "disinfectant" were in the top thirty most frequently used words that day, suggesting that perhaps this was more of a viral moment than an everyday normal reaction to Trump.

May 5th – Giving Tuesday Now

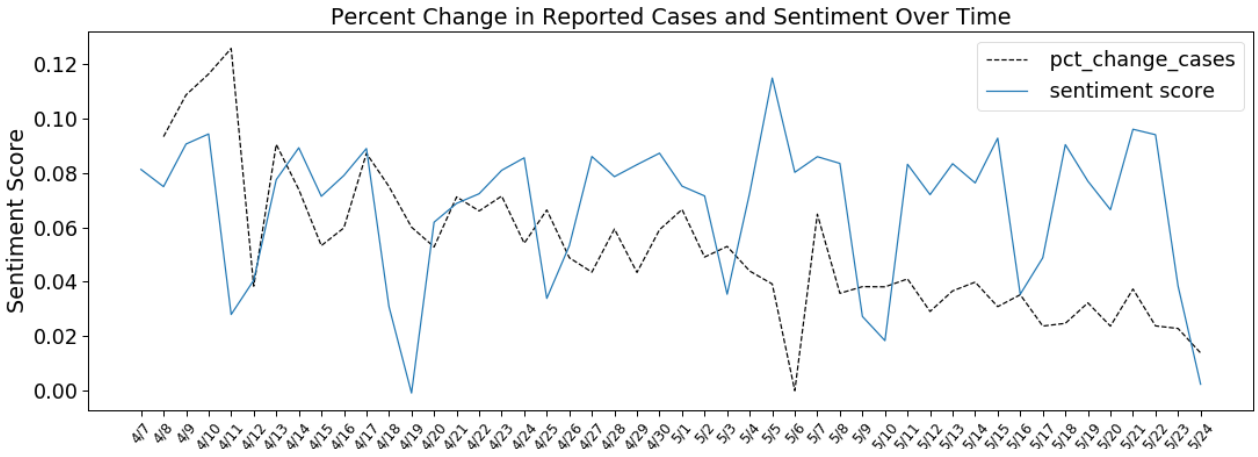
The May 5th positive spike in sentiment is related to the #GivingTuesdayNow campaign [7] – "a day of global action for giving and unity in response to COVID-19." Removing Tweets with #GivingTuesdayNow type hashtags dropped the sentiment score to .088 in line with the .0822 weekday average. It should be noted that the #GivingTuesdayNow Tweets (and positive sentiment) were disproportionately in Arlington county, Alexandria city, and Richmond city and the bump was not seen across the rest of the state.

Sentiment and Reported COVID Cases and Deaths

There was a total of 3,332 confirmed COVID cases and 69 deaths as of April 7, 2020 when data collection started. As of May 24, 2020—and only forty-six days later—there was 36,244 confirmed cases and 1,171 reported deaths in Virginia [8]. Despite the reported COVID cases and deaths steadily increasing, there seems to be no correlation between statewide DSS and the number of reported cases or deaths in Virginia.

While the number of reported COVID cases in the state continues to rise, the rate of increase has slowed—the flattening of the curve—since the beginning of April which can be seen in the Figure 5. The figure shows statewide DSS (blue) and the daily percent change in reported COVID cases (dashed black) over the data collection period. While both lines exhibit daily variations, the percent change of reported cases steadily declines over time while DSS remains fairly steady. An assumption of this analysis was that a spike in reported cases or deaths would be proceeded by or comingled with a negative spike in sentiment. From this visualization alone, there is no evidence of such a pattern. Additional time series analysis is needed to statistically test this hypothesis while adjusting for the clear weekend/weekday sentiment bias.

Figure 5: Statewide DSS and Percent Change in Reported COVID Cases



Sentiment by County

A goal of this research is to see how sentiment towards COVID-19 trends over time. Statewide, the number of COVID Tweets is declining and sentiment and frequency drop on the weekends.

Generally, sentiment has remained relatively steady throughout data collection. The same overall patterns are found the county-level data.

The "DSS_county" tab in the Tableau file is essentially the county-level tracking component of this project. This visualization shows COVID DSS by county over time. Color denotes sentiment where red is negative, and blue is positive. The width of the line denotes the overall sample size of the data. The wider the line, the more Tweets in the sample for that county. Not surprisingly, counties with smaller sample sizes present more variance in DSS. Despite the variability, the data does present unique stories for the different counties. The four largest localities in terms of sample sizes (Alexandria city, Arlington County, Fairfax County, and Richmond city) clearly show the weekend drop in sentiment. Fairfax County shows a severe drop on Saturday May 16th and Richmond trends more negatively overall. Both Fairfax and Richmond show a severe drop on May 24th. Bedford county, Newport News city, Rockingham County, York County, and Williamsburg city are among the most negative areas. In many locations, for example Manassas city, a sustained drop in sentiment over a one- to two-week period is observable.

Fairfax County on May 16th

With an average DSS of 0.082, Fairfax County saw its lowest DSS by far on May 16th with a score of -0.045. While no single item stood out as I reviewed the data, there seemed to be a mixture of bad news for Northern Virginia's largest suburban county that day, including:

- Washington Post reports of largest one-day increase in nationwide death toll since outbreak began [9]
- Washington Post reports that DC, MD, and VA saw cases double in one week [10]
- Virginia started Phase 1 reopening plan on 5/15, though Northern Virginia was excluded due to high levels of the virus

May 16th saw the most frequent usage of the words "death", "deaths", and "died". It was also the second highest Trump referenced day of data collection.

In Relation to Reported COVID Cases and Deaths

An assumption going into this analysis was that if the number of cases or deaths in an area saw a spike, Twitter sentiment would respond negatively around the time of the increase. The tabs "DSS_county_cases" and "DSS_county_deaths" track DSS alongside cumulative reported cases and deaths (per 1000 people), respectively. Using these visualizations alone, it is difficult to see any correlation between DSS and reported cases and deaths.

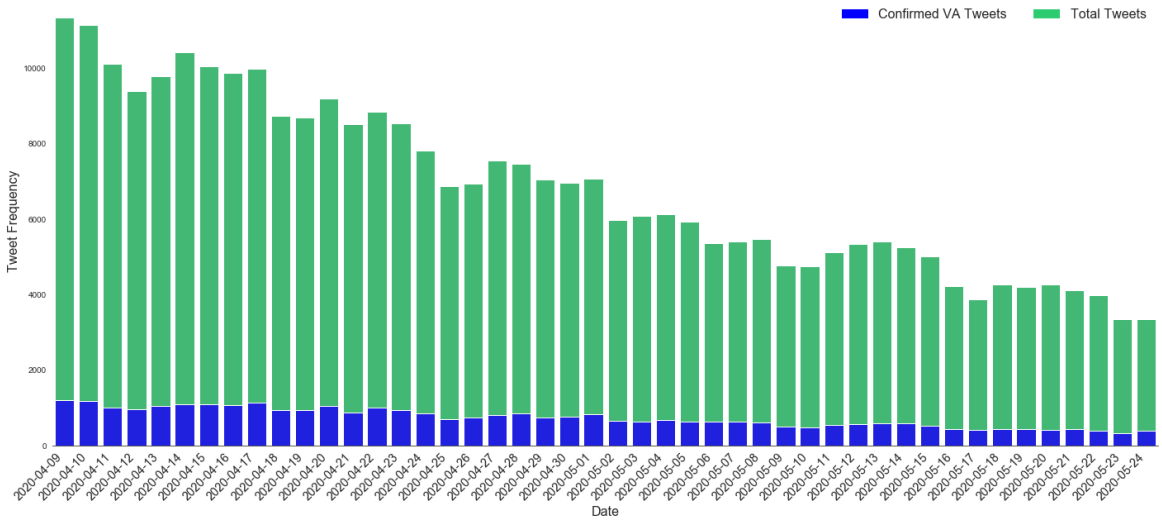
Visualizations linking DSS ranking and COVID case ranking (per 1000 people) do not show an overarching relationship between DSS and the COVID numbers. See "Slope - Confirmed Cases - 5/24 SnapShot" for DSS on May 24th and see "Slope - Cases Ranked by Avg Sent Ranked" for average DSS on Tableau (also shown in Figure 6). This figure shows the cumulative confirmed cases as of 5/24 ranked (high to low) on the left and average DSS (low to high) on the right, colored by sentiment. If sentiment was strongly related to COVID cases, we would expect the county rankings to be similar on both sides (high cases = low sentiment).

Figure 6: Cumulative Confirmed COVID Cases and May 24th DSS – Ranked by County

Results – Stay-At-Home Data Set

Figure 7 shows the frequency of stay-at-home related Tweets overall and those confirmed as Virginia Tweets.

Figure 7: Stay-at-Home Related Tweet Frequency - Overall and Virginia



Frequency of stay-at-home related Tweets has dropped over time. While weekend frequency is lower than weekdays within the same week, we do not see the same discrepancy that is present in the COVID data set. See also the "Overall Daily Sent" tab in Tableau.

Figure 8 shows the Tweet frequency by four regions: Northern Virginia, the greater Richmond area, the Hamptons Road area, and the remainder of the state. The vertical black line denotes when Phase 1 reopening began (May 15th) for all but Northern Virginia.

Figure 8: Tweet Frequency by Region– Stay-at-Home

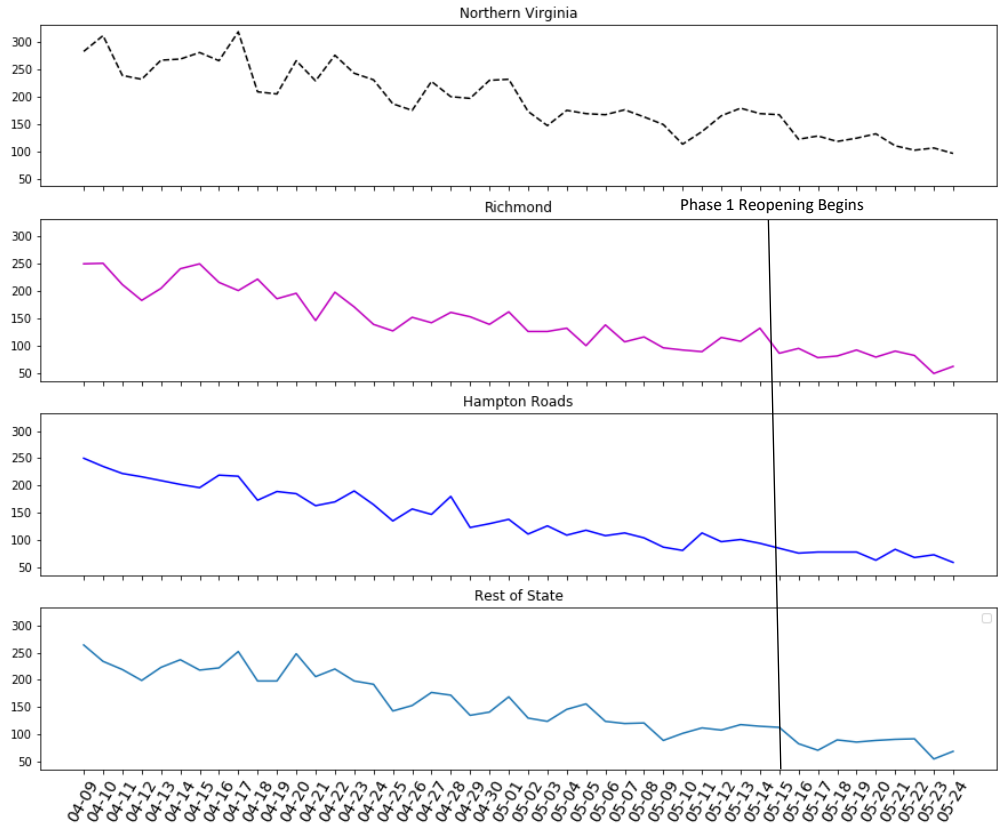


Table 4 shows the percentage of Tweets by county/city for the COVID and stay-at-home data sets for the top 10 areas. It also includes the percent of VA population and the Tweet-to-population ratio for the stay-at-home data. While Richmond city only represents 3% of the population, it is responsible for 19% of the entire data set, drastically overrepresenting the state's capital.

Table 4: Top Counites/Cities in Twitter Data

COUNTY	PERCENT VA POPULATION	PERCENT COVID TWEETS	PERCENT STAY-AT-HOME TWEETS	STAY TWEET TO POPULATION RATIO
RICHMOND CITY	0.03	0.14	0.19	7.32
VIRGINIA BEACH CITY	0.05	0.05	0.08	1.53
FAIRFAX	0.14	0.16	0.07	0.49
ARLINGTON	0.03	0.11	0.06	2.20
NORFOLK CITY	0.03	0.04	0.06	2.05
ALEXANDRIA CITY	0.02	0.10	0.06	3.17
LOUDOUN	0.05	0.04	0.05	1.12
PRINCE WILLIAM	0.05	0.02	0.02	0.42
ROANOKE CITY	0.01	0.02	0.02	1.91
NEWPORT NEWS CITY	0.02	0.01	0.02	1.04

Sentiment Analysis

The average sentiment analysis for all tweets from 4/9/20 through 5/24/20 is .179 (compared to .073 of the COVID data set) where 1 is totally positive and -1 is totally negative. Table 5 shows a few of the most positive and negative tweets for reference.

Table 5: Negative and Positive Stay-at-Home Tweets

Score	Tweet
-.9796	This quarantine is getting to meeeee omggggg 😭😭😭😭😭😭😭😭😭
-0.9755	I've moved on from the constantly sad part of quarantine to the straight up angry part. I am just SO angry that people are this fucking stupid. I'm angry that because of these idiots, the rest of my year will be cancelled. I'm just so angry.
-0.9676	I hate to say that but really this corona virus destroyed everything and even my quarantine is not doing good like others like I'm so confusing about how my certificate year will pass cause my future and career and this is so annoying and my insomnia is gettin worse buhhhh
0.9906	Quarantine's got me doing crazy (but great) things, like becoming a patron of Welcome to Night Vale on @patreon!! First time supporting anyone on this platform!! I could only give that honor to the people giving me comfort and amazing stories since 2013 :) https://t.co/H7K5BKwKnJ
0.9855	Thx to all who wished me a Happy Birthday yesterday. Weird to have a bday during quarantine, but I had so many people text, call, DM
0.9843	Help spread positivity, hope, and joy! Share your inspirational stories.\n\n#HopeFromHome #Hope #Kindness #joy #Inspiration #StayHome #CERT https://t.co/pqIXAU4uhZ

Figure 9 shows the statewide DSS by date and Figure 10 shows sentiment score collapsed by week of data collection.

Figure 9: Statewide Daily Sentiment Score (DSS) – Stay-at-Home

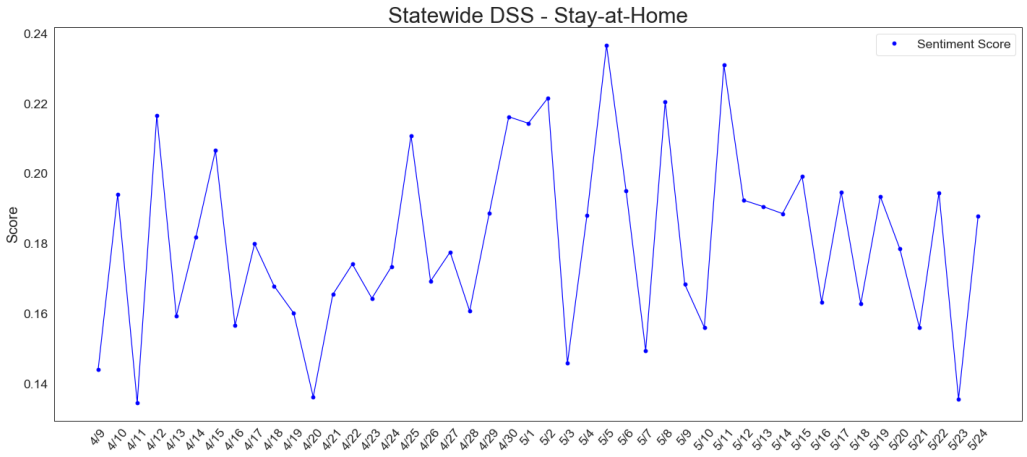
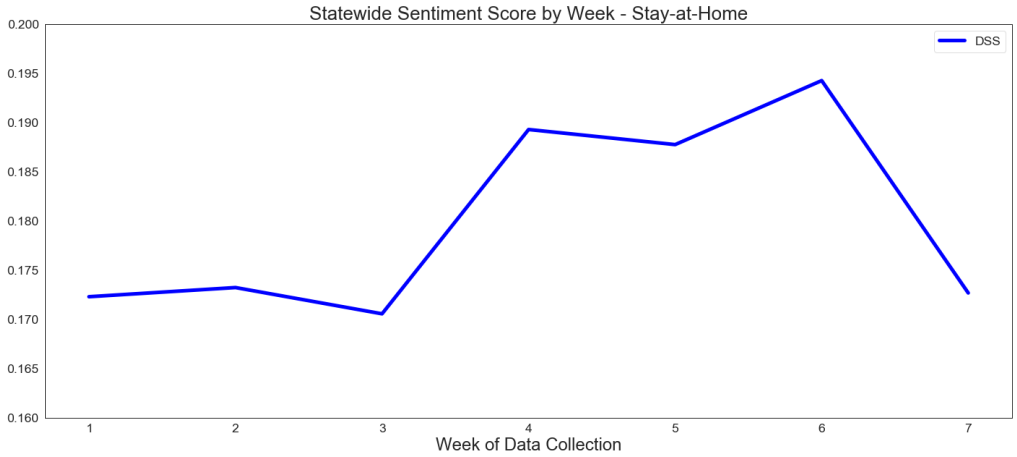


Figure 10: Statewide Sentiment Score by Week– Stay-at-Home



Unlike the COVID data, stay-at-home Tweet frequency does not have a strong relationship with sentiment nor do we see the sentiment drop on the weekends. See also the "Overall Daily Sent" tab in Tableau. The lowest sentiment score was 0.134 on April 20th and the highest was 0.236 on May 5th. The COVID low was April 19th (although April 20th was the lowest weekday score) which was seen to have a large Trump effect. The COVID high was also May 5th as a result of a COVID charity campaign with the hashtag #GivingTuesdayNow.

While the COVID sentiment remained generally stable throughout data collection, the stay-at-home sentiment was lower in the first two weeks (4/9-4/19), higher in weeks 3-6 (4/20-5/17) and dropped again in week 7 (5/18-5/24).

Sentiment by County

Tracking by County

The stay-at-home data does not work well for DSS tracking for most of the counties/cities simply because there is not enough data to make the trends meaningful. The Tableau tab "DSS_Stay_County" shows only 22 areas that had daily Twitter data, compared to 42 in the COVID set. The small sample sizes in most counties create volatility and the spikes in sentiment may be caused by only one or two Tweets with polarizing scores. To reduce the volatility, I collapsed DSS into four regions: Northern Virginia, the greater Richmond area, the Hamptons Road area, and the remainder of the state. See the"DSS_Stay_Region" and Figure 11 below. The vertical black line denotes when Phase 1 reopening began (May 15th) for all but Northern Virginia.

Figure 11: DSS by Region– Stay-at-Home

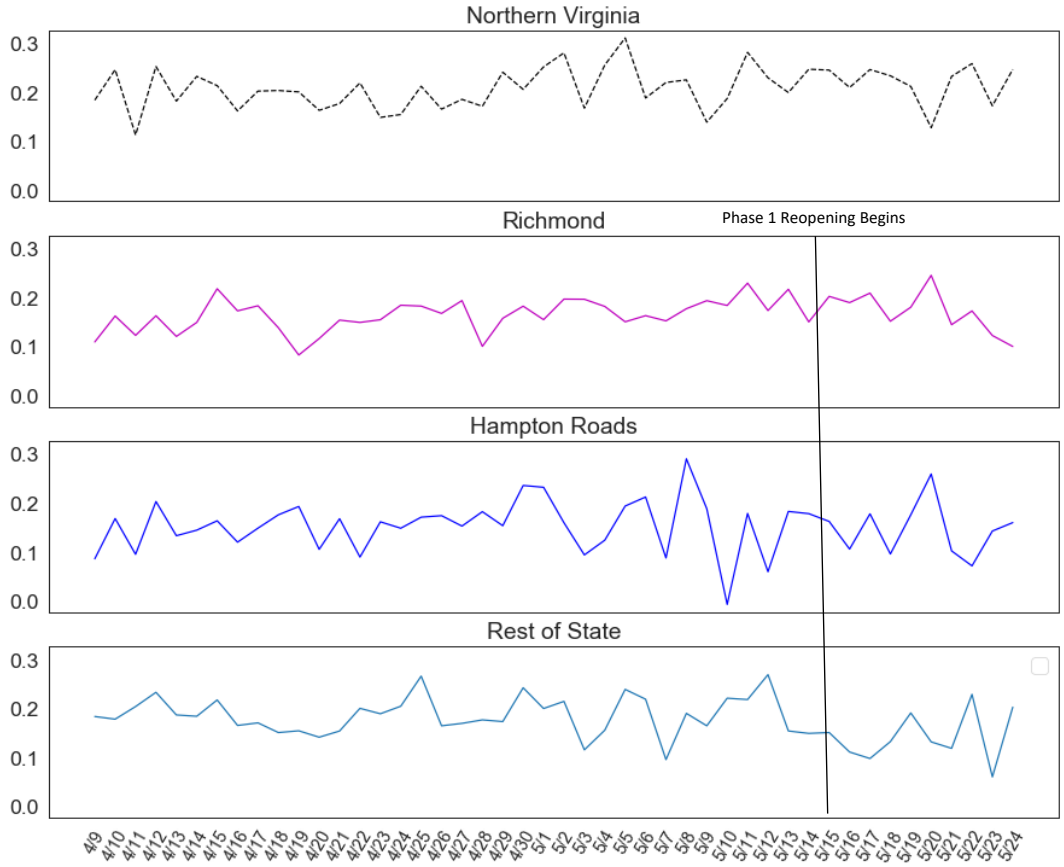


Table 6 shows the average DSS stay-at-home score by region. Northern Virginia is generally the most positive while the Hampton Roads area is the most negative.

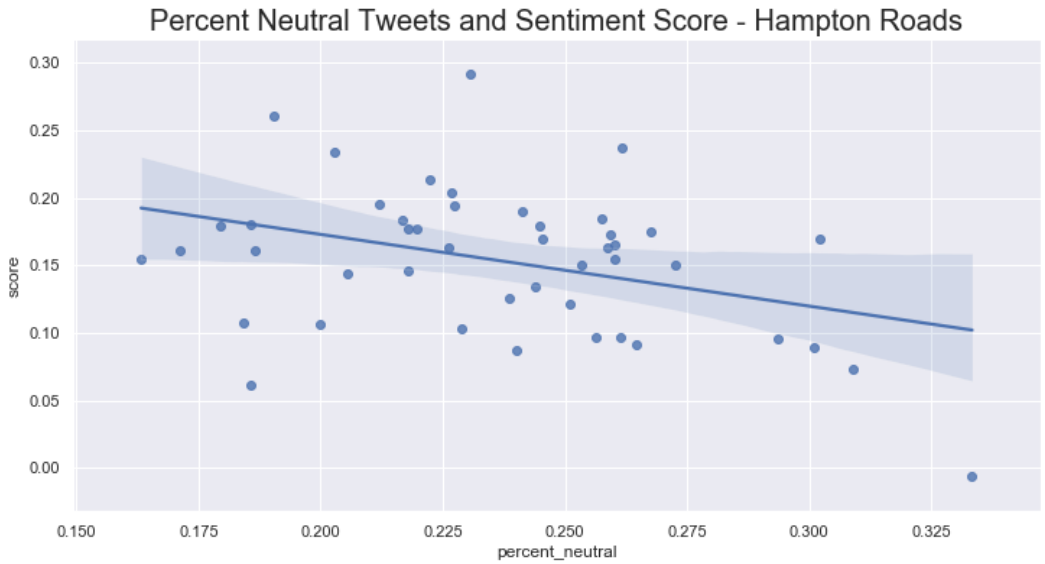
Table 6: DSS by Region

Region	Avg Score
Northern Virginia	.210
Richmond Area	.167
Hampton Roads	.154
Rest of State	.178

Hampton Roads on May 10th

The was a large dip in sentiment in the Hampton Roads area on May 10th with a DSS of -.0057. The dip can be seen in most of the areas that make up this region (Virginia Beach, Williamsburg, etc.) I could not find a pattern to the Tweets to indicate why sentiment dropped that day other than the percentage of neutral Tweets (where sentiment=0), was relatively high. While not statistically significant, there is a negative correlation between the number of neutral Tweets and sentiment score. The Tweets were not particularly negative that day, but they were also not particularly positive and the neutrality of the data skews the sentiment score downward.

Figure 12: DSS by Percentage of Neutral Tweets– Hampton Roads



County Level Differences

The COVID-19 pandemic remains a fluid situation as do statewide limitations on personal travel and social gatherings. To assess the data over time, I collapsed it into one-week periods (Monday-Sunday), although further analysis could use smaller time steps. Assessing the data by week enabled patterns to emerge across counties. However, before reporting these results, it is worth acknowledging how noisy the Twitter data actually is. While efforts were made to filter by county, the geolocation process also removed a lot of data simply because we could not pinpoint the user's location, greatly reducing the data that could be used. Additionally, while outliers were removed in terms of frequency, a lot junk Tweets remain (i.e. business announcements and advertisements.) Some of the results reported below were only observed using smoothed data – that is, weekly averages using daily averages. Some of the results changed or disappeared when only using weekly averages. Additionally, as already mentioned, some counties have relatively small sample sizes which can make the data unreliable. I believe these trends are worth reporting, but urge caution when drawing conclusions.

Reported COVID Cases

There was no relationship between overall average sentiment and reported COVID-19 cases (per 1000 residents) as of May 24th at the county level. However, there is a potential positive relationship between reported COVID cases and sentiment (higher cases=higher sentiment) during the last two weeks of data collection. However, as this was only seen in the smoother data set, I would recommend further study as data collection continues to see if this trend persists.

Reported COVID Deaths

There is a small positive correlation between overall average sentiment and reported COVID-19 deaths(per 1000 residents) as of May 24th ($r=.297$, $p<.05$). There is a small positive correlation between reported deaths (higher deaths=higher sentiment) in the last two weeks of data

collection (r=.401, p<.10, r=.36, p<.10, respectively). Appendix 2 shows sentiment by reported deaths over the 7 weeks of data collection.

Unemployment

According to Bloomberg News [11], while the May jobs report added 2.5 million jobs, a record number of Americans remain unemployed as a result of the nationwide Coronavirus lockdown that started in March. As such, I expected to see differences in sentiment in regard to stay-at-home measures in counties depending on the COVID-related unemployment impacts. The county level unemployment data was only available for March at the time of this analysis. The impact of the Virginia lockdown can be seen in the data with February to March increases in unemployment across all counites. The average rate of change was .187, with a low in Alleghany county of .048 (4.2% to 4.4%) and a high of .853 in Galax City (from 3.4% to 6.3%).

There was no relationship between average sentiment and March unemployment or Feb-to-March change in unemployment. However, for all but week 4 of data collection, there is a negative correlation between March unemployment rate and sentiment scores (higher unemployment = lower sentiment). The correlation ranges from a low of r=-.36 in week 1 to a high of r=-.58 in week 7, significant at the P<.10 level. Appendix 3 shows sentiment by March unemployment over the 7 weeks of data collection.

Urban vs Rural

There was no relationship between overall average sentiment and percent of an area that was urban, or by population density of urban-only areas. However, for weeks 3, 4, and 5 of data collection, there is a negative correlation between percent urban and sentiment scores (higher urban=lower sentiment). The correlation with r=-.51 in week 3, r=-.473 in week 2, and r=-.491 in week 5 all significant at p<.05. Appendix 4 shows sentiment by percent urban over the 7 weeks of data collection.

What People are Tweeting About

To use Twitter as a tracking tool, it is important to understand what people are actually saying in their social media posts. This data set was pulled using “stay-at-home” keywords. As such, we can assume each Tweet had some reference to the coronavirus induced stay-at-home orders given to all Virginians. However, the Tweets can be divided into many subgroups. Using LDA, I was able to pull out 25 “topics” based on the words used in the Tweets. I assigned each Tweet to a topic based on its highest probability of belonging to that group based on the words in the Tweet. I was then able to review the topics with the associated Tweets. While this exercise allowed me to group the topics mathematically, it should be noted that the topics showed extensive overlap and were subjectively named by me. Two basic categories emerged along with many subgroups which are listed below. Color coded word clouds were pulled from Tweets assigned to that topic.

1 - Coronavirus

- Illness itself, spread, testing, pandemic

- Government response, action
- Quarantine
- Support

2 - Life during quarantine

- Documenting life – "Day X of quarantine…"
- Working from home, being suck at home
- Staying safe
- Keeping busy
  - Baking, cooking, creativity, reading, family fun, tv shows
- Personal hygiene
  - Facial hair, haircuts, etc.
- Mental health
  - Depression, anxiety, "going crazy"
- Birthdays during quarantine
- Boredom
- Life *after* quarantine
- Relationship - Missing people, things, dating

The stay-at-home data set is far less political than the COVID data set. While the average sentiment score of stay-at-home Tweets that mention Trump is negative (-.083), we simply don't see the same effect on DSS because the number of Trump tweets is very small. Only 1.35% of the stay-at-home Tweets mention Trump compared to 7.91% in the COVID data set.

The stay-at-home data set is less about COVID-19 specifically than I had anticipated. Only 7.45% of the Tweets mention "COVID", "COVID-19" or "Coronavirus", and jumps to 8.30% if you include the word "pandemic." The data is more focused on life (at home) during quarantine. Almost 2.5% of the Tweets mention "hair", "facial hair" or "haircut," and another 1.3% are about birthdays. 1.6% mention being "bored" or "boring," and another 2.7% talk about "life after quarantine."

Conclusions & Limitations

Tracking Sentiment

A goal of this research was to see if tracking COVID and stay-at-home Twitter sentiment at the county level was possible. Ultimately, I was able to track the COVID and stay-at-home related sentiment statewide and by county for those that had enough data. I was able to detect spikes in sentiment and pinpoint the reasons behind them using natural language processing methods. However, this worked better with the COVID data than with the stay-at-home data given the volume of Tweets.

The frequency of Tweets in both data sets has declined steadily over time, suggesting subject matter fatigue. As phase 1 reopening gets underway in Virginia, I would expect stay-at-home Tweets to decline, unless there is a spike in cases/deaths as a result. There is a relationship between reported deaths and positive sentiment in the stay-at-home data, suggesting perhaps that more affected areas are more supportive of stay-at-home measures.

There is a relationship between stay-at-home sentiment and March unemployment (higher unemployment=lower sentiment) when looking at weekly sentiment rates. I recommend reassessing this relationship when the April and May unemployment numbers are released. I would expect this discrepancy to hold or grow for unevenly distributed unemployment across the state.

Geolocation Issues

For this analysis we were only interested in Virginia residents. Twitter does not require users to enter a location, but provides an optional free text field called “location” in a user’s profile. While the vast majority of users in our data did provide some text (i.e. the field was not blank), only 11% were coded to a county or city in Virginia using the Google Maps API. An enormous amount of data was discarded, and while most of the users were clearly not in Virginia, some of them were. This impacts the integrity of the data and is an impediment to analyzing Twitter data at lower geographical areas.

Sentiment Towards What

The assumption was that by collecting Tweets using specific COVID or stay-at-home keywords I was capturing sentiment towards those things. While in theory I may have done this at a macro level, I am not sure how useful either DSS metric would be to the Department of Health. The CDC learned while studying flu related Tweets that there was large amount of “Twitter chatter” (i.e. media attention) surrounding the flu that was confounding any relationship between Tweet volume and actual infection rates. By definition, this pandemic has touched the life of every single American and therefore the volume of “Twitter chatter” in the COVID data set is no doubt enormous.

The COVID data set was overwhelmingly political, with Donald Trump having much more of an impact on DSS than reported COVID cases, deaths, unemployment or the urban/rural divide.

Additional analysis is needed to tease out the weekend vs weekday phenomenon present in the data.

The stay-at-home data set was far less noisy, but contained much more about everyday life during the quarantine than anticipated. I expected this data set to be more politically charged given some of the political discourse and dialog we have seen about states infringing on citizen's rights with lockdowns and the requirements to wear masks. What can we learn from Tweets about birthdays and haircuts (or the lack thereof)? One thing that is clear from the data is that people are staying home.

From a public health perspective, I would not recommend using these sentiment scores as an attitudinal measure towards COVID or stay-at-home measures. The COVID data set is too politically charged, and the stay-at-home data focuses too much on everyday life. That is not to say there is nothing useful in this data. In particular, I think the stay-at-home data set could be mined further to identify how people are coping with the "new normal" during the lockdown. Overall sentiment might not be valuable, but further analysis on the impacts of unemployment and insights into mental health/wellbeing could be useful to inform policy decisions.

Twitter as Surveillance Framework

This project was largely exploratory, and what it revealed is that simply using keywords to capture Tweets on a certain topic does now necessarily provide a clean assessment of sentiment toward that topic. However, lessons learned from this project could be applied to a public health Twitter surveillance framework going forward. Perhaps Tweet frequency, as the CDC already uses flu monitoring, is a better measurement given the inherent noise in Twitter data. Additional research is needed to either refine the keywords used (filter out the "chatter"), or to filter down specific topics for analysis.

The stay-at-home data set contained a rich description of how Virginia residents are living life during the quarantine and could be further mined for insights. If public schools do not reopen in the fall, it would behoove the state to track similar data to assess how families with children at home are coping with this unprecedented situation.

A promising application of sentiment tracking will be when a vaccine does become available. If DOH can filter out the "Twitter chatter" surrounding the vaccine, it would be a worthwhile subject to track to assess attitudes and opinions toward the vaccine, and potentially address reluctance or trust issues that emerge.

References

[1] World Health Organization. (2020) Q&A: Influenza and COVID-19 - similarities and differences. Retrieved April 7, 2020, from www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-similarities-and-differences-covid-19-and-influenza#:~:text=Mortality%20for%20COVID%2D19,quality%20of%20health%20care.

[2] Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PloS one, 8(12), e83672. https://doi.org/10.1371/journal.pone.0083672

[3] Wojcik, S., Hughes, A. (2019) Sizing Up Twitter Users. Retrieved May 24, 2020 from https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

[4] Pandey, P. (2018) Simplifying Sentiment Analysis using VADER in Python (on Social Media Text). Retrieved April 7, 2020, from https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f

[5] Lane, Howard, Hapke (2019). Natural Language Processing in Action. Manning Publications. Shelter Island, NY.

[6] Farley, R., Kiely, E. (2020) Fact Check: The White House Spins Trump's Disinfectant Remarks. Retrieved June 1, 2020, from https://www.nbcwashington.com/news/coronavirus/fact-check-white-house-spins-trump-disinfectant-remarks/2284542/

[7] GivingTuesday. (2020) GivingTuesday Announces Day of Global Action for Giving and Unity in Response to COVID-19. Retrieved May 16, 2020, from https://www.givingtuesday.org/blog/2020/03/givingtuesday-announces-day-global-action-giving-and-unity-response-covid-19

[8] NYT Github: https://github.com/nytimes/covid-19-data

[9] Hawkins, D., Bellware, K., Mettler, K., Beachum, L., Hassan, J., Thebault, R., Armus, T. (2020) U.S. sees largest one-day increase in coronavirus death toll since the outbreak began; San Francisco area asked to shelter at home. Retrieved June 1, 2020, from https://www.washingtonpost.com/world/2020/03/16/coronavirus-latest-news/

[10] Cox, E., Wiggins, O., Olivo, A. (2020) D.C., Virginia and Maryland coronavirus cases double in a week to exceed 20,000. Retrieved June 1, 2020, from https://www.google.com/search?q=%E2%80%A2+Washington+Post+reports+that+DC%2C+MD%2C+and+VA+saw+cases+double+in+one+week&oq=%E2%80%A2%09Washington+Post+reports+that+DC%2C+MD%2C+and+VA+saw+cases+double+in+one+week&aqs=chrome..69i57.370j0j7&sourceid=chrome&ie=UTF-8

[11] Dimitrieva, K. (2020) U.S. Hiring Rebounds, Defying Forecasts for Surge in Joblessness. Retrieved June 6, 2020, from https://www.bloomberg.com/news/articles/2020-06-05/u-s-jobless-rate-unexpectedly-fell-in-may-as-hiring-rebounded

[12] Eichstaedt, J. “What We Can Learn From Twitter Analysis About COVID-19.” COVID-19 and AI: A Virtual Conference. April 1, 2020.

[13] Giachanou, Anastasia & Crestani, Fabio. (2016). Like It or Not: A Survey of Twitter Sentiment Analysis Methods. ACM Computing Surveys. 49. 1-41. 10.1145/2938640.

Appendix 1 – Geolocation by Data Set

The project focused on county-level (and independent cities) data, and a major challenge was getting reliable location data. According to the Twitter website, there are three ways a tweet can be tagged with geolocation metadata:

1. **Geotagged by user** – User provides exact location when issuing Tweets. While very precise, only 1-2% of Tweets are geo-tagged. (Project data has less than 1%)
2. **Mentioned locations in tweets** – Twitter parses tweets for location mentions. Accuracy is likely to be lower as this simply refers to a location and does not necessarily mean the user is there.
3. **User**-reported location via profile – Users can report their location in their Twitter user profile. This is the largest source of location data. (99.5% of project data)

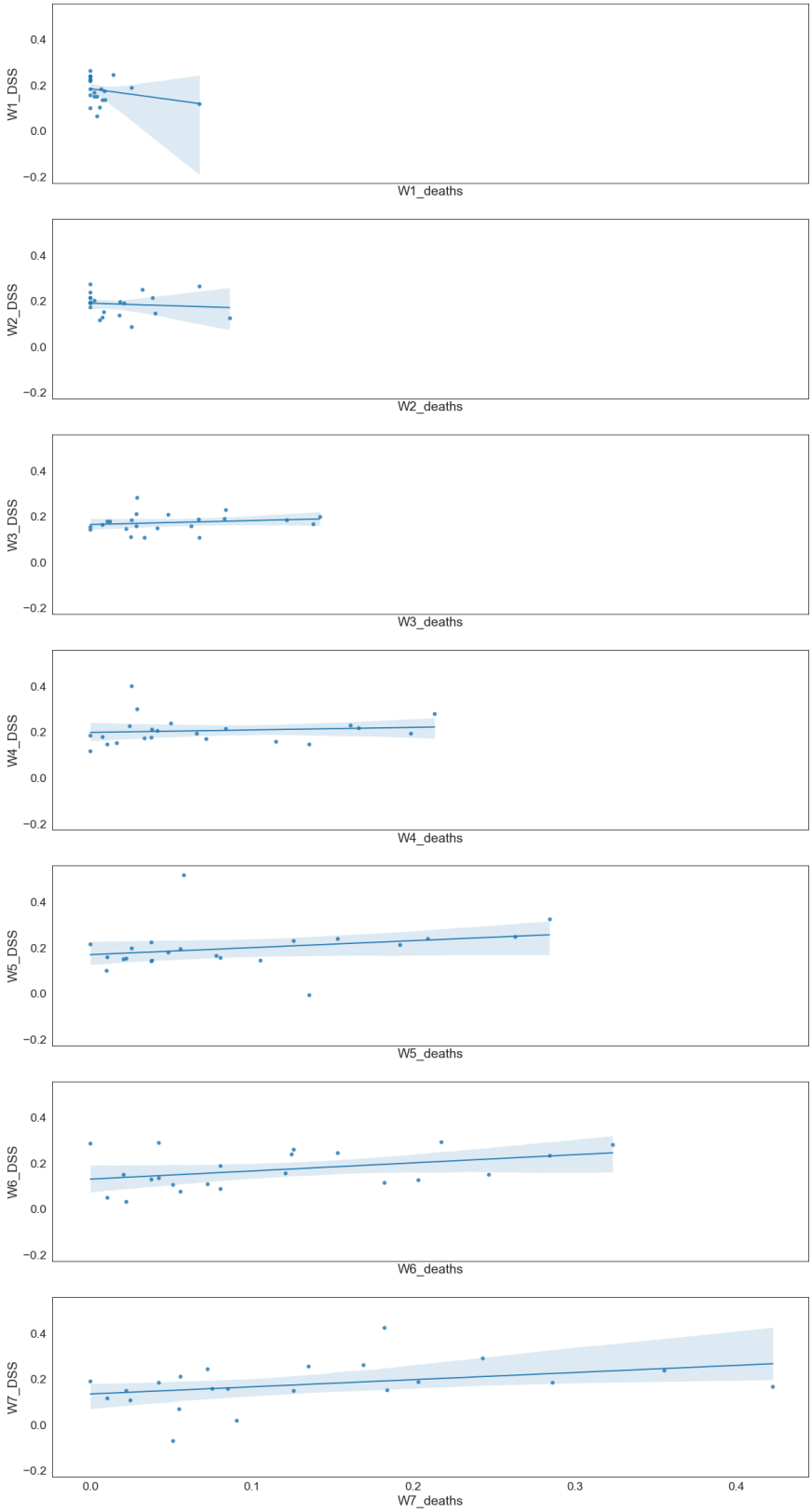
Data collection was filtered to Virginia using the Twitter API search using rtweet's `lookup_coords` function. Almost all users provide a location in their user profile, however many of the locations are not in VA. Many are in DC, MD, NC, WV, among others. Some user-reported locations are not real or identifiable (i.e. space, candy land, my house.) All this plus the open-ended nature of user location text field makes for very messy data.

User provided locations were sent through the google maps API which returned a city, county and state. Non-Virginia locations were excluded. Locations that were unable to be coded to one of the 133 counites or independent cities in Virginia were excluded (for example, "Virginia" or "NoVa" or "Southern Virginia"). The table below shows the final breakdown of location data.

Data Cleaning – Geolocation Data, Outlier Users, & Duplicates

	COVID	%	STAY-AT-HOME	%
TOTAL RECORDS	1,491,411		314,311	
CODED TO VIRGINIA	163,642	11.0%	33,928	10.8%
USED IN ANALYSIS (OUTLIERS AND DUPLICATES REMOVED)	112,576	68.8%	28,772	84.8%

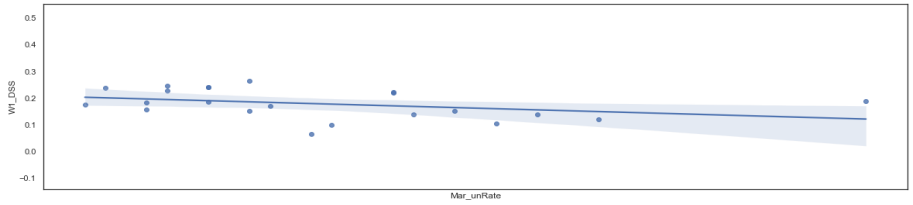
Appendix 2 - Sentiment by Reported Deaths by Week



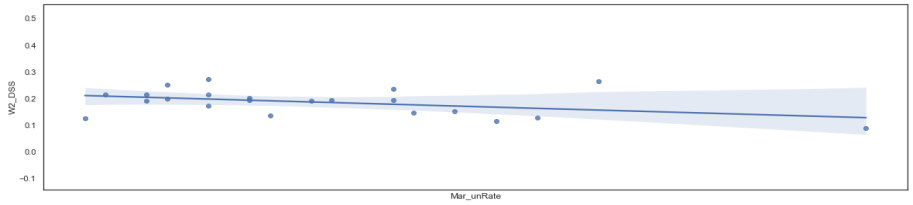
r=.401, p<.10,

r=.36, p<.10,

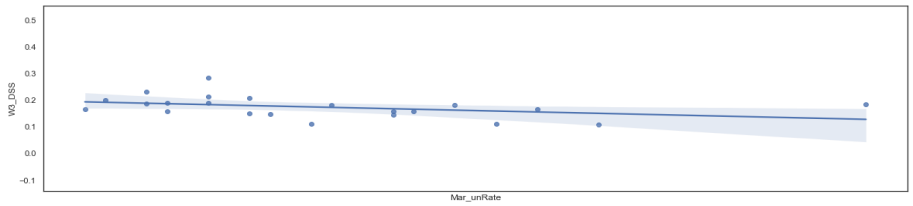
Appendix 3 - Sentiment by March Unemployment By Week



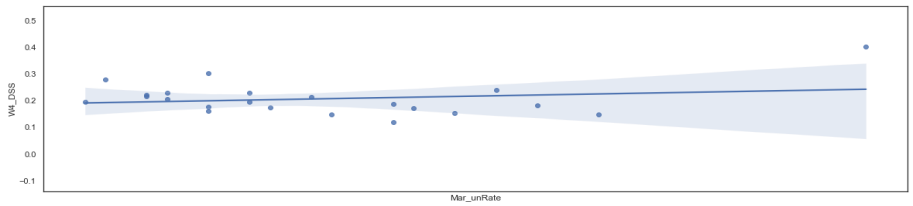
r=-.36, p<.05,

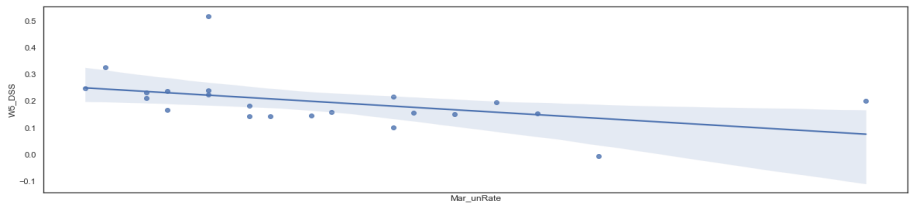


r=-.409, p<.05,

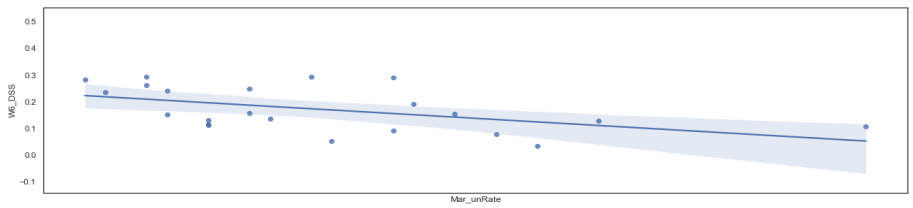


r=-.388, p<.05,

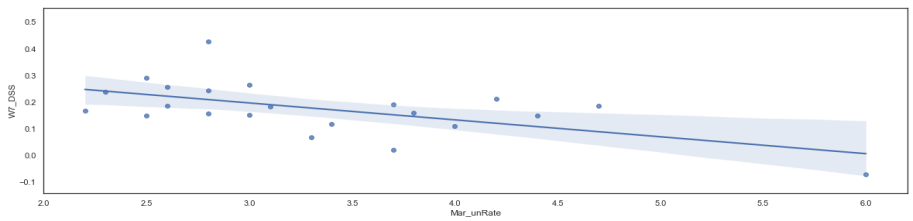




r=-.436, p<.05,

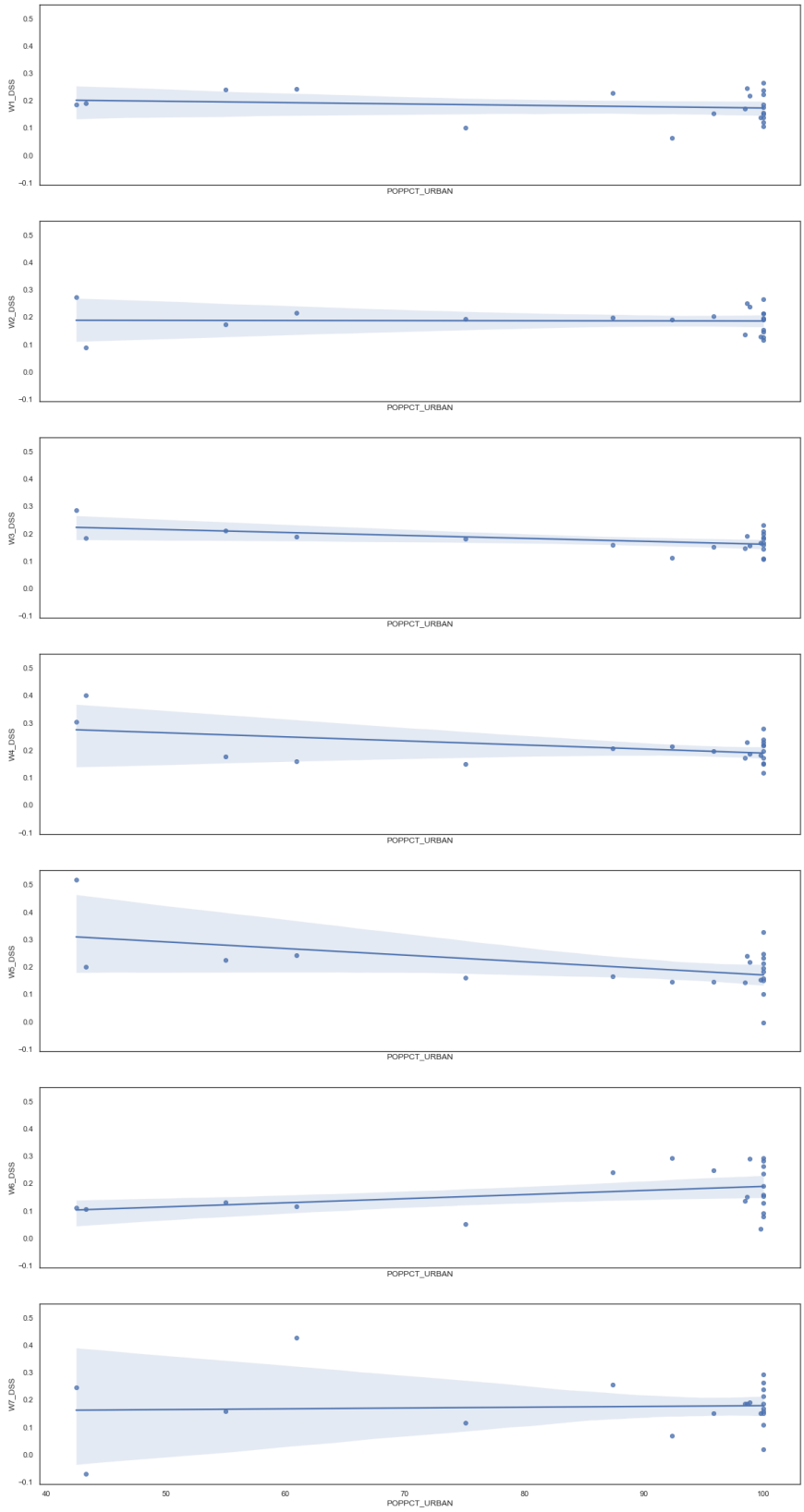


r=-.497, p<.05,



r=-.586, p<.05,

Appendix 4 - Sentiment by Percent Urban By Week



r=-51., p<.05,

r=-.473, p<.05,

r=-.491, p<.05,