

Towards Predicting COVID-19 Trends: Feature Engineering on Social Media Responses

Antone Evans Jr.

Department of Computer Science
University of Illinois Springfield
aevan7@uis.edu

Yingyuan Yang

Department of Computer Science
University of Illinois Springfield
yyang260@uis.edu

Sunshin Lee

Department of Computer Science
University of Illinois Springfield
slee675@uis.edu

ABSTRACT

During the course of this pandemic, the use of social media and virtual networks have been at an all-time high. Individuals have used social media to express their thoughts on matters related to the pandemic. It is difficult to predict current trends based on historic case data because trends are more connected to social activities which can lead to the spread of coronavirus. So, it's important for us to derive meaningful information from social media as it is widely used. Therefore, we grouped tweets by common keywords, found correlations between keywords and daily COVID-19 statistics and built predictive modeling. The features correlation analysis was very effective, so trends were predicted very well. A RMSE score of 0.0425504, MAE of 0.03295105 and RSQ of 0.5237014 in relation to daily cases. In addition, we found a RMSE score of 0.07346836, MAE of 0.0491152 and RSQ 0.374529 in relation to daily deaths.

Keywords

Feature engineering, social media, correlation analysis, machine learning, COVID-19, Twitter.

INTRODUCTION

On January 14, 2020, the World Health Organization (WHO) reported the novel coronavirus (COVID-19) to be potentially transmitted by human contact¹. WHO later stated on January 20, 2020, with the increased infections among health workers there proves to be a stronger case with COVID-19 being transmitted among humans². Although, on January 30, 2020, COVID-19 was marked a pandemic which travelled through human transmission³. The United States of America (U.S.A) was hit with its first COVID-19 cases on January 21, 2020, which was later confirmed on February 26, 2020 (Jordan et al., 2020). During this time, the Centers for Disease Control and Prevention (CDC) began to spread information about the COVID-19 virus through social media channels. This increased engagement on public health official's social media accounts as information was disseminated about the developing COVID-19 pandemic occurring across the world. Quickly following this the U.S.A halted as the cases in New York began to rapidly increase starting in late February (Thompson et al., 2020). In return cities around the U.S.A seeing what happened in New York began to enforce social distancing mandates, the wearing

¹ <https://twitter.com/UNGeneva/status/1217146107957932032>

² <https://twitter.com/WHOWPRO/status/1218741294291308545?s=20>

³ [https://www.who.int/director-general/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-\(2019-ncov\)](https://www.who.int/director-general/speeches/detail/who-director-general-s-statement-on-ihr-emergency-committee-on-novel-coronavirus-(2019-ncov))

of mask and the closure of non-essential businesses and activities such as churches, bars, beaches, etc.⁴

Although this pandemic became a common thread around the world, it wasn't the only commonality. The world had lockdowns, social distancing protocols, and mask mandates in common. With social distancing protocols in place, we noticed the use of social networking increase dramatically as users attempted to keep in contact. This made individuals within all generations increase their use of social media. Social media platforms such as Twitter had a 24% increase in daily users in relation to its previous year⁵. Users new and old flocked to Twitter as they began to vocalize their concerns about the government's plan.

Due to this increase in users, Twitter became a high candidate for data within our study. Twitter was used over other social media platforms as this study aimed to determine if based on the nation's understanding towards the pandemic and government's policies implemented. Could common thoughts – keywords, be grouped whether positive or negative in nature and used to model COVID-19 statistics? In addition, Twitter by nature forces individuals to be very direct with their messaging. Individuals are limited by 280 characters; this worked in our favor rather than other platforms as it allowed us to do less data cleaning within individual tweets. We assumed that each tweet is limited to a single thought. Moreover, tweets were used over mass media as we seek to understand how the average user within America feels and not a select group of individuals.

Throughout their vocalization we noticed there to be a common thread within each. Topics such as “working from home” and “shutdown of social activities” came into topic on the platform. We note this as hashtags such as #CoronaLockdown trended at number 1 on March 27, 2020 on Twitter⁶. This hashtag encompassed tweets such as, “Officially working from home ... Good thing there's a bar close by.”⁷. In addition, citizens of countries expressing their continued concern for the health of their nations in relation to the guidelines or lack thereof put into place by their government. We note this as #StayAtHomeAndStaySafe trended to number 1 on March 28, 2020 on Twitter⁸. As tweets such as, “Sure, we don't close the economy over just the flu.”⁹ surfed the internet. Therefore, the goal of this paper is to find a correlation and create models based on these correlations between the daily cases and deaths (DCCD) within the U.S.A using common keywords found within thoughts shared on Twitter.

We hypothesized there would be a negative correlation between cases and tweets with a variation of “wear mask” and “mask”. As businesses and governments continued to enforce wearing a mask, we expected cases to decrease and tweets about wearing a mask to increase. It's noted that the use of simple masks can reduce transmission of COVID-19 by 40% (Gakidou, October 2020). In addition, there would be a negative correlation between contact tracing and deaths. As contact tracing increased, we expect deaths to decrease. Based on the research newsletter produced in the JAMA, we note by increasing contact tracing individuals can self-isolate and quarantine faster which can lead to lower mortality rates (Bilinski et al., 2020). With the use of both instances within our model, we expected this to positively increase the likelihood of our model accurately performing when mapping our predicted graph vs. the actual graph.

RELATED WORK

To summarize, we will use keywords to model DCCD within the U.S.A. through searching and tracking common threads between individuals' thoughts throughout the pandemic. To first understand and solve the problem at hand, we first reviewed current literature surrounding this topic.

Tweet Collection and Keyword Search: Researchers such as (Lee et. al., 2012) have used open source tools such as yourTwrapperKeeper to collect tweets. We differ from this as we used a different Social Feed Manager (SFM). Several keyword approaches for pandemic keyword search have been proposed including (Husnayain et al., 2020; Lin et al., 2020; Liu et al., 2020; Monhamadou et al., 2020; Shinde et al., 2020; Li et al., 2020; Oyebode et al., 2020; Tang et al., 2020). These studies focused on the use of Google searches to model COVID-19 cases throughout various countries. These papers aided in our solution as they indicated different keywords and common

⁴ <https://abcnews.go.com/US/News/timeline-100-days-york-gov-andrew-cuomos-covid/story?id=71292880>

⁵ https://www.washingtonpost.com/business/economy/twitter-sees-record-number-of-users-during-pandemic-but-advertising-sales-slow/2020/04/30/747ef0fe-8ad8-11ea-9dfd-990f9dcc71fc_story.html

⁶ <https://us.trend-calendar.com/trend/2020-03-27.html>

⁷ <https://twitter.com/SaraKronenberg/status/1243595836069761028>

⁸ <https://us.trend-calendar.com/trend/2020-03-28.html>

⁹ <https://twitter.com/gdeguzman99/status/1243956649884172289>

threads of keywords. More importantly these papers indicate the significance of those keywords. Therefore, we focused on using Twitter keywords when conducting our research. Twitter was used because of its restrictions on the number of characters within tweets. It allowed tweets to be leveraged and cleaned easily from our algorithms.

Machine Learning: We used Simple Linear Regression (SLR) and Support Vector Machine (SVM) using time series. As we attempted to use common ML algorithms which could be exploited as there used in addition to other algorithms for optimization. In addition, studies such as (Lee et al., 2015) have used ML to predict the location of a user based on their tweet. However, we used the metadata within the location field to group users within the United States (U.S.).

Observation of Government Mandates: As COVID-19 has caused many states to limit social distancing activities to slow down the spread of the virus. We must first understand what these activities are when developing our keyword set. Various studies focused on the effect of these mandates (Alagoz et al., 2020; Chiu et al., 2020; Fang et al., 2020; Lyu et al., 2020). (Alagoz et al., 2020) notes the following COVID-19 mandates put in place by New York, mass gathering restrictions started on 12 March 2020, initial business closures were recommended on 16 March, educational facilities were closed on 18 March, nonessential services closed, and a statewide stay-at-home order was issued on 22 March. Moreover, (Lyu et al., 2020) indicates the following mandates for the state of Virginia which began on May 22, 2020, as the state required residents to wear a face mask in public where people congregate.

Sentiment Analysis: With the use of social media tweets, understanding the emotions behind individuals' messages was an important step. We found knowledge on this domain with the use of (Nemes et al., 2020) paper on sentiment analysis and its relation to social media. Before conducting extensive data cleaning, it was important to divide tweets into sentiments as it gave an opportunity for mapping this data out and finding clusters.

Previous ISCRAM Studies: We can concur with (St. Denis et al., 2020) that first-hand information shared by those affected by a disaster contribute to a richer understanding of how community members are directly impacted by a disaster. We note that this rich information has such an effect on everyday life that it can be modeled almost exactly to that of disasters. Moreover, our study follows the narrative of (Reuter et al., 2018) that Twitter has a dominant impact and is a common tool used within the ISCRAM publications during the period of their study. Furthermore, (Yang et al., 2018) study focuses on the classifications of tweets into different categories based on the "What, When, Where and Who" within the tweet data. Whereas our study dives into creating predictive modeling simply on the combination of these elements. As the "what" represents the different keywords, the "when" is the daily rate of those keywords on a given date, the "where" is within the U.S. and the "who" indicates the average Twitter user who tweets from within the U.S.

We used a combination of similar studies with a focus on finding the best keyword feature when modeling COVID-19 statistics. So, we put an emphasis on understanding if social media platforms truly had the power to influence individuals. By our definition, we define social media influence by the accuracy of our model using only the correlation produced by keywords. If we are truly able to see a decrease or otherwise an increase in DCCD due to the knowledge and concern voiced by individuals on social media. Then, social media would be considered to have a huge influence on our social activities and willingness to follow COVID-19 protocol throughout this pandemic.

METHODS

Data

The DCCD from February 2020 – July 2020, in the U.S. were collected from the WorldOMeter¹⁰ and Google COVID-19 Tracker websites¹¹. As the reporting of DCCD during the first months of tracking were not consistent across platforms. The Google COVID-19 Tracker was used to verify the accuracy of the numbers reported on the WorldOMeter website.

The daily tweets related to coronavirus were gathered through the use of a SFM produced by The George Washington University¹². The dataset can be found at <https://tweetsets.library.gwu.edu/>. The study was able to gather over 75 million tweets which spanned across February 2020 to July 2020. As the SFM used the Twitter API to gather tweets, our raw data included both the tweet data in addition to the metadata related to the posters

¹⁰ <https://www.worldometers.info/coronavirus/country/us/>

¹¹ <https://news.google.com/covid19/map?hl=en-US&gl=US&ceid=US%3Aen>

¹² <https://library.gwu.edu/scholarly-technology-group/social-feed-manager>

profile.

Data Cleaning and Exploration

After data collection was completed, the first step was to understand our data. As indicated by Figure 1, 63% of the data within our dataset were in English, all other tweets were removed from the dataset.

Percentage of Different Tweets by Language which Mentions Coronavirus

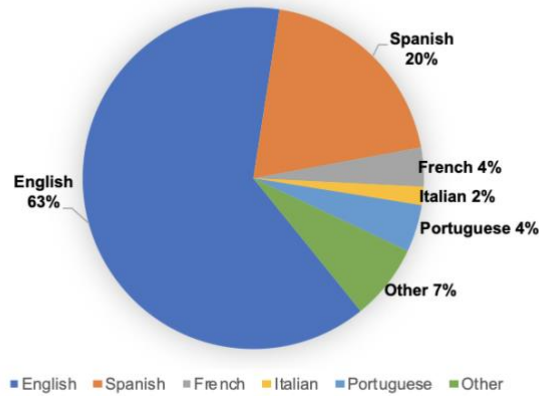


Figure 1. Percentage of Different Tweets by Language which Mentions Coronavirus

Following this step, we investigated those English tweets to identify how many were original tweets – tweets which were not retweeted or otherwise quoted. As seen by Figure 2, 96% of the tweets within our dataset were noted to be original tweets. After exploring the 4%, it was noted that quoted tweets are seen to be expressing one’s thoughts on the original tweet which it is quoting, or it is simply agreeing with the original user. By this assumption the study did not remove quote tweets as additional individuals used this Twitter feature to develop their own thoughts on posted issues.

Percentage of Original vs. Quoted Tweets Mentioning Coronavirus

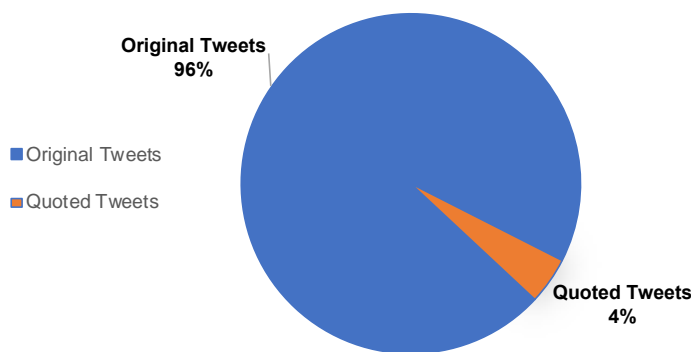


Figure 2. Percentage of Original vs. Quoted Tweets Mentioning Coronavirus

With the initial data cleaning completed, we utilized our knowledge of the dataset and our study’s goal to develop an architecture as indicated in Figure 3.

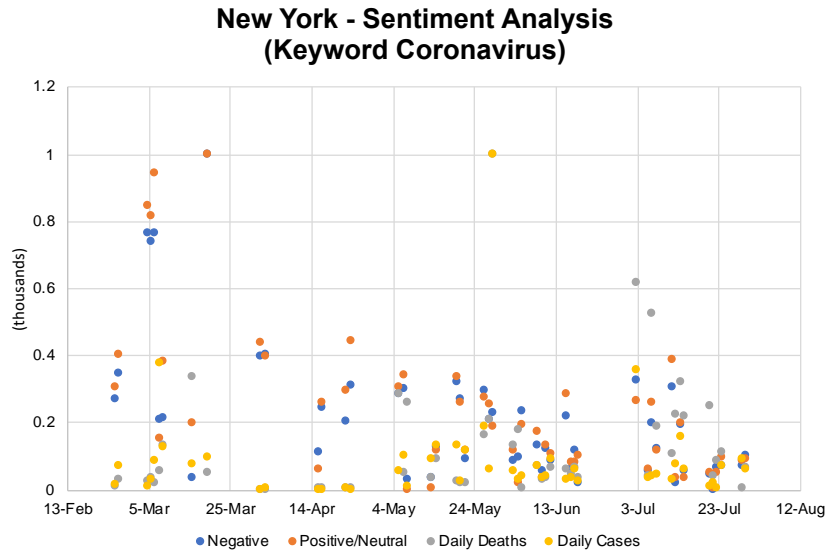


Figure 4. New York - Sentiment Analysis (Keyword Coronavirus)

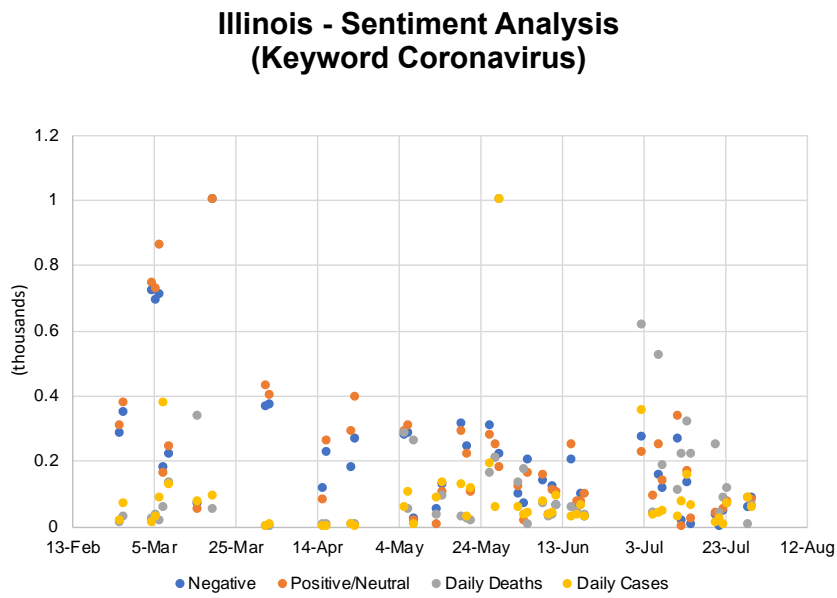


Figure 5. Illinois - Sentiment Analysis (Keyword Coronavirus)

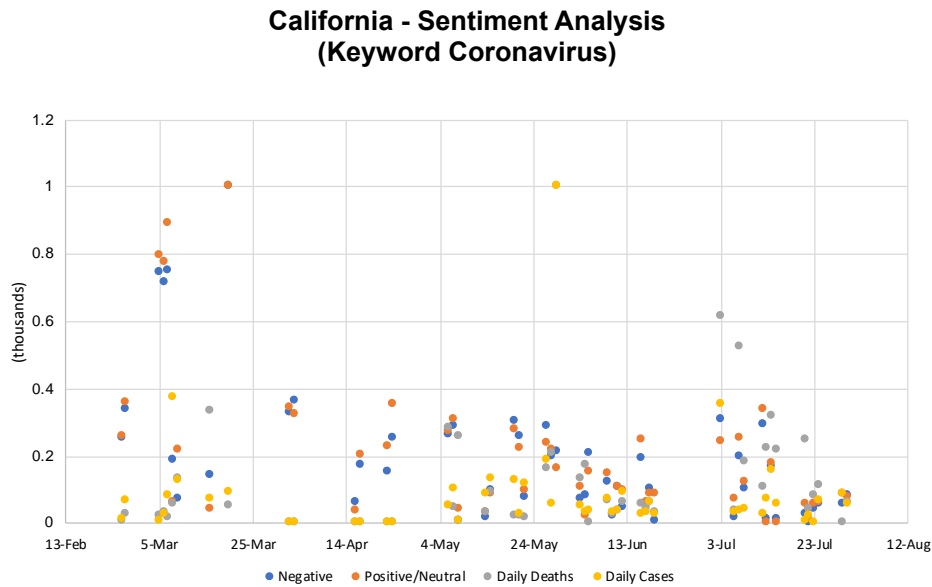


Figure 6. California - Sentiment Analysis (Keyword Coronavirus)

Keyword Search

Following sentiment analysis, our tweet text observations were normalized in format by using the NLP library. This allowed the removal of urls, html code, punctuation and much more. In addition, stemming and lemmatization were conducted to return the root of all keywords. With the data now formatted, we developed a list of common keywords which were deemed related topics shared by individuals. Keywords were chosen which related to both COVID-19 effects to one’s physical health and hospitalization as we used the Kaiser Family Foundation (KFF) website¹³ to develop this sector. The KFF website was used as it provided a reliable and detailed glossary of various topics/terms. In addition, keywords were chosen which closely related to social distancing protocols but in place by both local and federal governments, organizations and businesses throughout the U.S. Thus, Table 1 lists the keywords used within our study. The study’s keywords were a very critical choice as all our algorithms relied heavily on the aggregation and correlation of keyword counts. We observe a summarization of our keyword counts in Figure 8. Between all keyword searches, there was a sum of 30,405,065 tweets. This calculated to be 40.54% of our initial dataset.

Table 1. Keywords Used in Dataset

Keywords
Asymptomatic, Bar, Beach, Case Fatality Rate, Case, CDC, CFR, Church, Close Contact, Community Spread, Community Transmission, Contact Tracking, Corona, Covid, Curve, Droplet, Epidemic, Essential Worker, Fake, Hospital Bed, ICU, Isolation, Mask, Negative, Pandemic, Positive, Quarantine, Recovery, SARS, Self Quarantine, Spread, Stay at Home, Symptomatic, Symptom, Testing, Tracing Ventilator, Airline Flight

Aggregating Keywords

Due to the size of the dataset and the increased number of tweets on certain days, there was an overflow of tweets from the same day within numerous files. Therefore, aggregating keywords on the date feature was completed. This allowed the date to have the correct representation of tweets based on keywords on each day. As seen in Figure 7, we visualize the daily number of tweets mentioning ‘coronavirus for various dates.

¹³ <https://www.kff.org/glossary/covid-19-outbreak-glossary/>

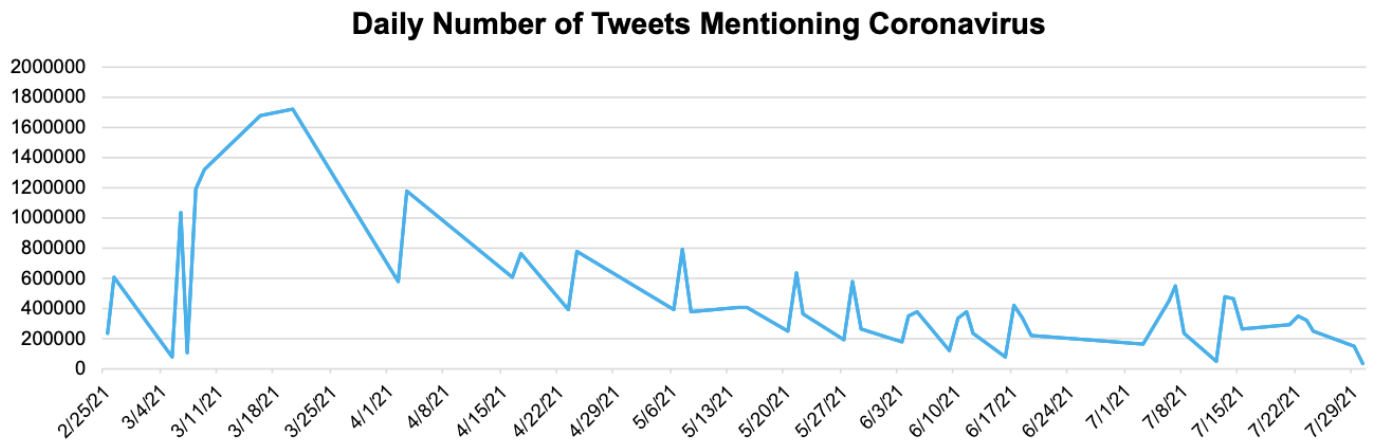


Figure 7. Daily Number of Tweets Mentioning Coronavirus

Scaling Data

As DCCD and the sum of daily keyword tweets being on different scales, it was important to scale our data. This would allow for a clear representation of the data and would ultimately help with the creation of our models. As noted by Figure 3, keywords and COVID-19 statistics were combined in preparation of building a dataset to feed to the model.

Absolute Delta Learning Rule

In an attempt to scale our data, the absolute delta learning rule was used. The absolute delta learning rule is defined as the absolute subtraction of the first observation and its subsequent observation within a feature. Mathematically the absolute delta learning rule is described as, $x_i = |y_i - y_{i+1}|$, where x represents the newly created feature for the absolute delta. And y_i represents the current observation value and y_{i+1} represents the subsequent observation. Here is a review of using the absolute delta learning rule.

1. Where x_i , represents observation 1 within our dataset. Thus,
 $x_i = NULL$ as it is currently a blank observation
2. As $x_i =$ observation 1, this indicates that y_i , is also related to observation 1 where,
 $y_i = \text{observation 1}$
3. As y_{i+1} , is related to the subsequent observation, observation 2. In this case,
 $y_{i+1} = \text{observation 2}$
4. Indicating the following,
 $x_i = |y_i - y_{i+1}|$
5. If $y_{i+1} = NULL$ indicating that there was no subsequent observation. This indicates that y_i was the last observation within the feature. Thus,
 $x_i = |y_i - 0|$

Min-Max Normalization

To further scale our data into the same range, min-max normalization was utilized. Min-Max Normalization linear transformed the values between the range of 0 – 1. As seen in Figure 8, the range of the graph surpasses 1 because the graph is a representation of the sum of observations for each feature for the datasets date range. However, min-max normalization was done on a daily level and not on an overall span level. Mathematically the min-max normalization is described as, $X_{scaled} = \frac{x_n - x_{min}}{x_{max} - x_{min}}$ (Mustaffa and Yusof, 2010)

where,

X_{scaled} = new value for variable X

X_n = current value for variable X

X_{min} = minimum value in the dataset

X_{scaled} = maximum value in the dataset

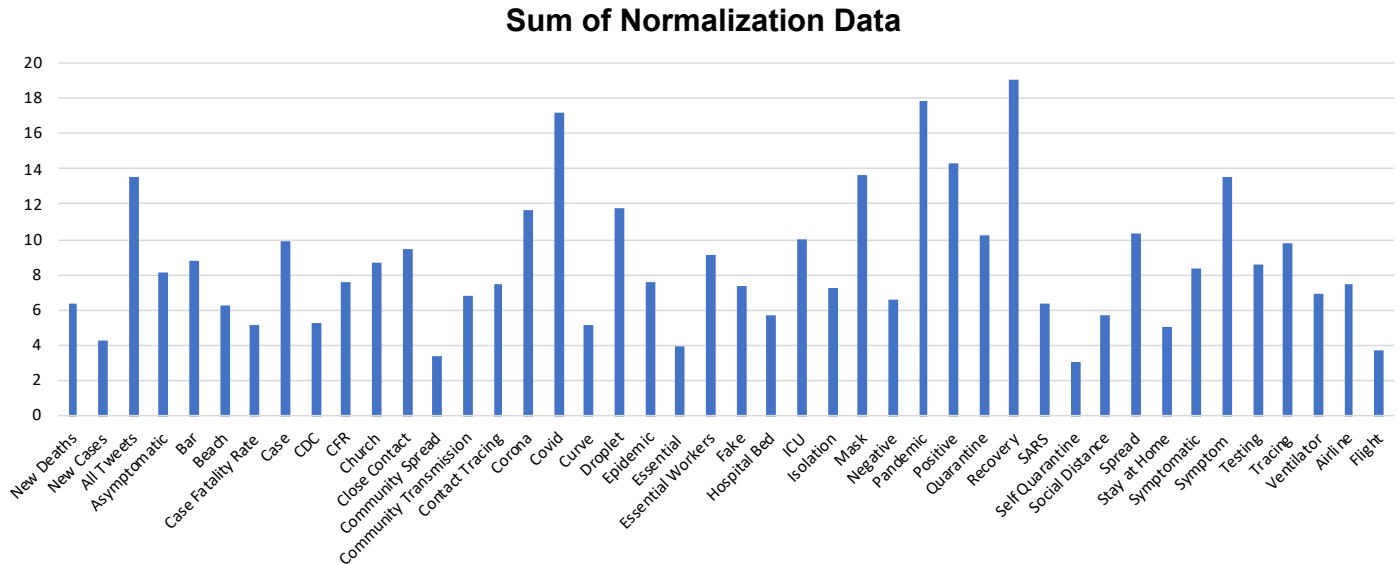


Figure 8. Aggregated Count of Min-Max Normalization of Data

Spearman Correlation

Methods of correlation summarize the relationship between two variables in a single number called the correlation coefficient. This study uses Spearman correlation to calculate the correlation between the DCCD and each of the 41 keywords used within the filtering tweets section. Mathematically spearman correlation is described as, $\rho =$

$$1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2-1)} \text{ (Mukaka, 2012)}$$

where,

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

In the instance of finding the correlation between the entirety of the dataset, there were no strong correlations found. However, as we began to get more high level with our calculations and began to break correlations down by months. Using a significance level of $\rho > 0.40$ and $\rho < -0.40$, we found 36 correlations between keywords and daily deaths (DD). This correlation threshold was used to find all correlations which were deemed moderate or stronger in accordance with¹⁴ as noted in Table 2. In addition, we found there to be 32 keywords correlated with DC as seen in Table 3.

Table 2. Correlation Thresholds and Rankings

Correlation Threshold	Ranking
0.00 – 0.19	“Very Weak”
0.20 – 0.39	“Weak”
0.40 – 0.59	“Moderate”
0.60 – 0.79	“Strong”
0.80 – 1.00	“Very Strong”

¹⁴ <https://www.statstutor.ac.uk/resources/uploaded/spearman.pdf>

Table 3. Keywords Correlated with Target Keyword

Keywords Correlated to Death	Keywords Correlated to Cases
Asymptomatic, Bar, Beach, Case Fatality Rate, Case, CDC, CFR, Close Contact, Community Spread, Community Transmission, Contact Tracing, Corona, Curve, Droplet, Epidemic, Essential, Hospital Bed, ICU, Isolation, Mask, Negative, Pandemic, Positive, Quarantine, Recovery, SARS, Self Quarantine, Spread, Stay at Home, Symptomatic, Symptom, Testing, Tracing, Ventilator, Airline	Asymptomatic, Bar, Beach, Case Fatality Rate, Case, CDC, Church, Close Contact, Community Transmission, Contact Tracing, Corona, Covid, Droplet, Epidemic, Essential, Essential Worker, Fake, ICU, Isolation, Mask, Negative, Quarantine, Recovery, SARS, Self Quarantine, Spread, Stay at Home, Symptomatic, Symptom, Tracing, Airline, Flight

Modeling and Evaluation

Dataset

With our data now being scaled, a dataset was created as seen in Table 4. To summarize the dataset covers dates spanning between February 25, 2020 to July 30, 2020. In addition, all data within the dataset with the exception of the date observations are represented using min-max normalization. As seen in Table 4, there are 41 keywords within the dataset where {41} represents the keywords listed in Table 1.

Table 4. Features in Dataset

Feature name	Description
Date	The date of COVID-19 statistics and tweets
Daily Deaths	The daily number of COVID-19 deaths within the U.S.
Daily Cases	The daily number of COVID-19 cases within the U.S.
Total Number of Daily Tweets	The overall number of coronavirus tweets posted within the U.S.
{41} Keywords	The number of times a keyword was mentioned within the related date. Note: {41} is a representative of 41 unique variables

APPROACH AND RESULTS

In this study, we used SLR. SLR models study the relationship between a single dependent variable Y and one or more independent variables X it is called SLR (Bangdiwala, 2018). In addition to SLR, SVM was used, it is a classification and regression prediction tool that uses ML theory to maximize predictive accuracy while automatically avoiding overfitting to the data (Jakkula, 2006).

We used SLR to build a base model. Given that SLR models are seen as lightweight and basic models in ML and are built upon. This allowed us to see how our model would perform over the data with little to no computational power. Following the creation of our base model, we used a SVM to find patterns within our data. SVMs are used in time series modeling which allows this algorithm to use past patterns to predict future trends from the given data.

As our model uses the correlation between the keywords and DCCD reported, we leveraged this information when creating keyword groups for feature engineering. We used a significance level of $\rho > 0.40$ and $\rho < -0.40$ for keywords selected in both groups in relation to the target variable. We then selected keywords at random to build two groups for both target variables as seen in Table 5.

As the goal of our paper is geared towards finding the best keyword(s) to use when modeling DCCD. To find the best keywords to use, we conducted feature engineering. Our feature engineering was broken down into 6 categories – baseline using either the DCCD depending on the target variable; baseline and total number of daily tweets; DCCD and feature group (FG) 1; DCCD and FG 2; DCCD and all 41 keywords; and DCCD, all 41 keywords and total number of daily tweets.

We then ran both of our algorithms over all six categories. Following this step, we evaluated each category using the Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R-Square (RSQ). With the use of 2 algorithms and 3 evaluators, we had the ability to see the average of how different categories and clusters of keywords performed. This allows for a double verification on the values produced.

Our data cleaning, exploration and scaling was done in Jupyter Notebook 6.0.3. Our predictive modeling was done in R 4.0.2. Algorithms used within our modeling were SLR using the `lm()` function; in addition to svm using the `e1071` library and `svm()` function.

Table 5. Target Variable Feature Groups

DC FG 1	DC FG 2	DD FG 1	DD FG 2
Asymptomatic	Beach	Asymptomatic	Bar
Case	Church	Case	Beach
Case Fatality Rate	Essential Worker	Case Fatality Rate	ICU
Epidemic	Flight	CDC	Positive
Isolation		Curve	Recovery
Symptomatic		Droplet	SARS

Table 6. DC Feature Evaluation

Feature Engineering	Evaluation	Model	Score
DD + 41 Keywords + Total Number of Daily Tweets	RMSE	SLR	0.0425504
DD + 41 Keywords + Total Number of Daily Tweets	RMSE	SVM	0.1268647
DD + 41 Keywords + Total Number of Daily Tweets	MAE	SLR	0.03295105
DD + 41 Keywords + Total Number of Daily Tweets	MAE	SVM	0.04142392
DD + 41 Keywords + Total Number of Daily Tweets	RSQ	SLR	0.9187351
DD + 41 Keywords + Total Number of Daily Tweets	RSQ	SVM	0.5237014

Table 7. DD Feature Evaluation

Feature Engineering	Evaluation	Model	Score
DC + FG 2	RMSE	SLR	0.07346836
DC + FG 2	RMSE	SVM	0.119796
DC + FG 2	MAE	SLR	0.0491152
DC + FG 2	MAE	SVM	0.06255359
DC + FG 2	RSQ	SLR	0.8307047
DC + FG 2	RSQ	SVM	0.374529

As seen from Table 6 and 7, we were informed of the values produced by each evaluator. When determining which keyword(s) produced the best model, we used the average of how input features did over all evaluators. From Table 6, we notice that the input feature of all 41 keywords, daily count of tweets and DD produced the best evaluation scores over all evaluators. This input feature has a RMSE score of 0.0425504, MAE score of 0.03295105 and RSQ score of 0.5237014 between both algorithms. We note from Figure 9 and 11 that our model graphs the data very accurately using this input feature. From Table 7, we can note that the input feature of DC and FG 2 produced the best average evaluator score over all evaluators. This input feature has a RMSE score of 0.07346836, MAE score of 0.0491152 and RSQ score of 0.374529 between both algorithms.

We note from Figure 10 and 12 that our model graphs the data moderately accurately using this input feature. We can further note that both cases and deaths have different input keyword features which produced the best modeling graphs. This can be expected as we noticed a difference in correlations between keywords over both DCCD.

Simple Linear Regression

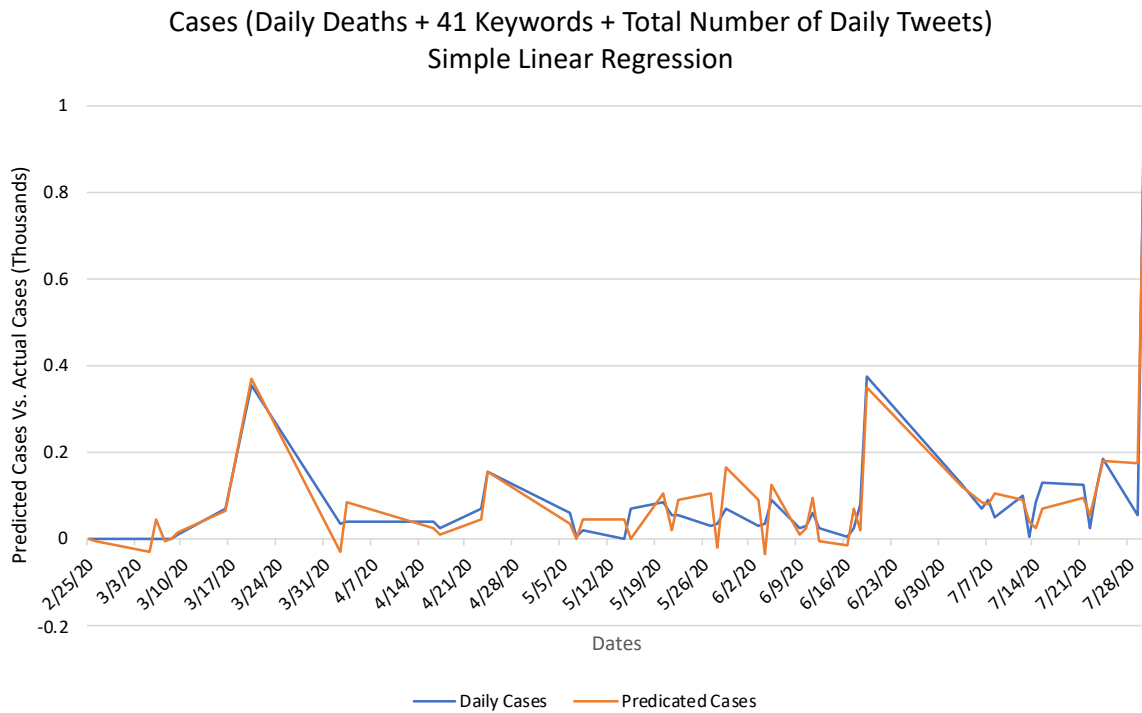


Figure 9. SLR – Modeling DC using DD, 41 Keywords and Total Number of Daily Tweets

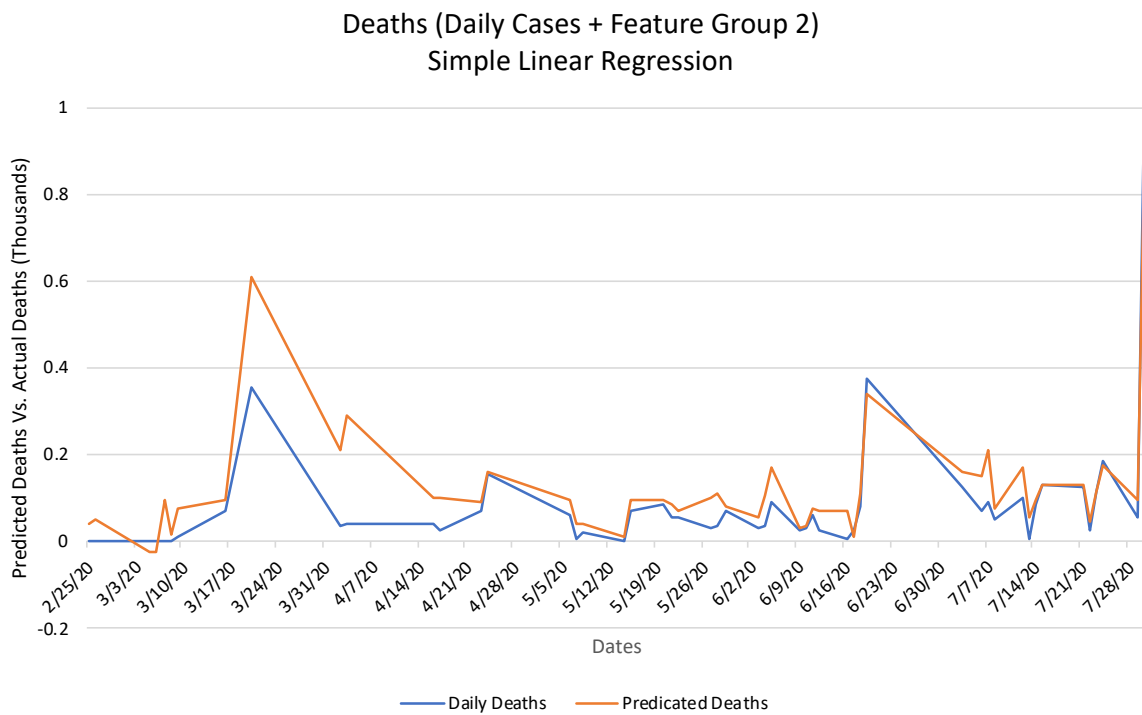


Figure 10. SLR – Modeling DD using DC and FG 2 Keywords

Support Vector Machine

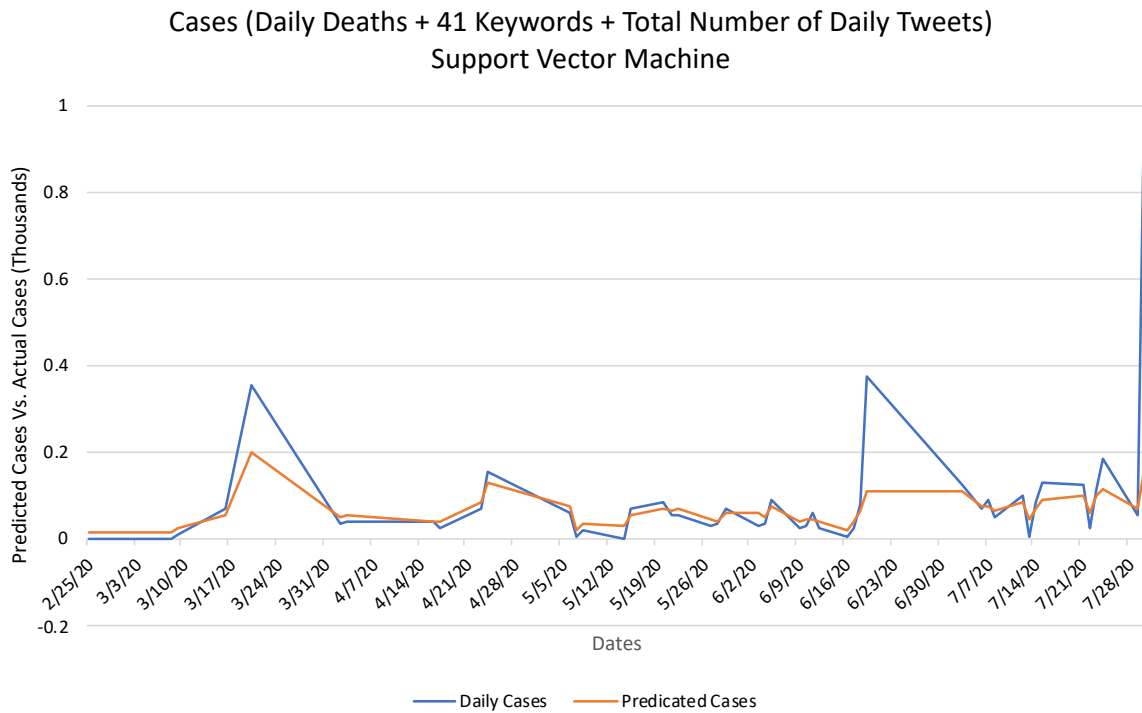


Figure 11. SVM – Modeling DC using DD, 41 Keywords and Total Number of Daily Tweets

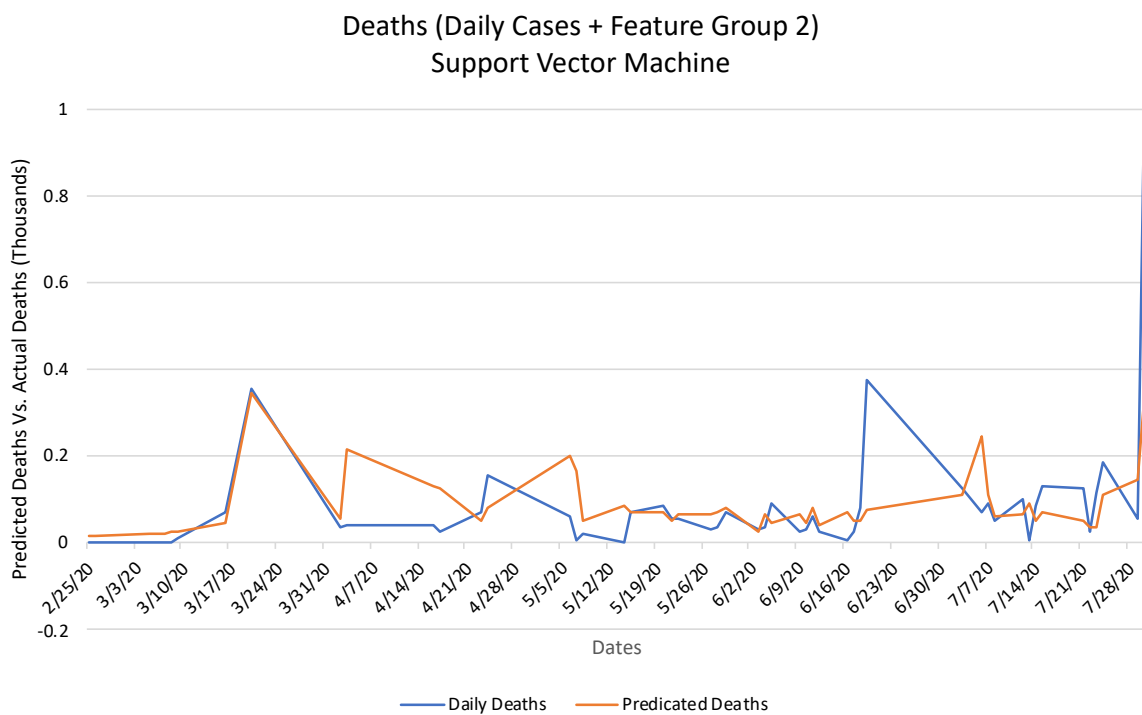


Figure 12. SVM – Modeling DD using DC and FG 2 Keywords

DISCUSSION

As we can see from our above Figures 9 – 12, with the use of simple ML algorithms we were able to accomplish an accurate modeling against that of DCCD. The study demonstrates a correlation between keywords noted in Table 3 and DCCD. As observed from this study with the use of rich data such as Twitter, we found unique ways to predict modeling in an effective way which does not have lag time. With the use of these simple keywords, we were able to successfully model DCCD. We observed graphs and evaluation scores that SLR performed better than SMV. This can be seen in our figures on 07/30/2020 as SMV undershot in its predictive modeling on this day.

Moreover, this analysis supports our hypothesis that there would be a negative correlation between cases and tweets such as “wear mask” and “cases”. As businesses and governments continued to enforce wearing a mask, we expected cases to decrease and tweets about wearing a mask to increase. In addition, there would be a negative correlation between contact tracing and deaths. As contact tracing increased, we expected deaths to decrease. Our data supports this as we note keywords such as “mask” having a -0.69 correlation with DC. We note there to be a correlation of -0.47 between the keyword “community transmission”. A negative correlation indicates that as tweets surrounding these keywords increase then DCCD decreases and vice versa. On the other hand, there is seen to be a correlation of 0.56 between DC and “essential”. And a correlation of “essential worker” and DD, this indicates there to be a positive correlation. A positive correlation indicates as tweets surrounding these keywords increases DCCD also increase and vice versa.

Furthermore, we analyzed state rules and regulations on March 20th, 2020 New Orleans issued its stay-at-home order; Alabama activated its national guard; Miami ordered hotels to shut down and implemented a curfew¹⁵. On April 30th, 2020 California closed down beaches within Orange County; and New York shutdown the subway overnight¹⁶. On May 6th, 2020 the Supreme Court rejected to lift Pennsylvania's coronavirus restrictions and Latinos in Illinois were reported to test at a higher rate for the virus than any other group¹⁷. Oregon required mask wearing outside and limited gatherings to 10 or less and Louisiana limited gatherings to 50 or less, closed bars to indoor dining and mandated wearing a mask on July 17th, 2020¹⁸. Finally, on July 30th, 2020 Indiana extended its state of emergency and the District of Columbia announced students would learn remotely during the fall semester¹⁹.

During our study, there were several limitations which we encountered. Firstly, due to our initial tweet dataset not containing daily tweets, we were unable to have a full month's worth of data to work with. In addition, while conducting our experiments we were unable to create more FGs as we did not have enough computational power. Moreover, we were unable to search for all possible keywords when conducting this study. With more FG and more combinations of different keywords, better scores may be found. We note that several of our limitations could be solved if we used a distributed file system similar to the architecture of (Lee and Fox, 2017).

CONCLUSION

From our study, we can conclude that social media has played a significant role throughout the pandemic. Our study while using a significance level of $\rho > 0.40$ and $\rho < -0.40$ was able to find 36 correlations between keywords and DD. In addition, there was a correlation found between 32 keywords and DC. Moreover, after breaking our feature engineering 6 categories we were able to find the best features for our target variables. For the input feature – DC there was a RMSE score of 0.0425504, MAE score of 0.03295105 and RSQ score of 0.5237014 between both algorithms found. In addition, with the input feature – DD, there was a RMSE score of 0.07346836, MAE score of 0.0491152 and RSQ score of 0.374529 between both algorithms found.

ACKNOWLEDGES

Thank you, Dr. Lee for your support from the inception of this research process. Dr. Buxton for giving me the tools to confidently conduct and build all models needed. Dr. Yang for your great feedback and insights while developing this paper. Thank you all.

¹⁵ <https://www.cnn.com/world/live-news/coronavirus-outbreak-03-20-20-intl-hnk/index.html>

¹⁶ <https://www.cbsnews.com/live-updates/coronavirus-updates-2020-04-30/>

¹⁷ <https://www.cnn.com/us/live-news/us-coronavirus-update-05-06-20/index.html>

¹⁸ <https://www.natlawreview.com/article/covid-19-us-state-policy-report-july-13-2020>

¹⁹ <https://www.natlawreview.com/article/covid-19-us-state-policy-report-july-30-2020>

REFERENCES

- Alagoz, O., Sethi, A. K., Patterson, B. W., Churpek, M., and Safdar, N. (2020) Effect of Timing of and Adherence to Social Distancing Measures on COVID-19 Burden in the United States: A Simulation Modeling Approach, *Annals of internal medicine*.
- Bangdiwala, S. I. (2018) Regression: simple linear, *International Journal of Injury Control and Safety Promotion*, 25:1, 113-115, DOI: 10.1080/17457300.2018.1426702
- Bilinski, A., Mostashari, F., Salomon, J. A. (2020) Modeling Contact Tracing Strategies for COVID-19 in the Context of Relaxed Physical Distancing Measures, *JAMA Netw Open*, 3, 8.
- Chiu, N. C., Chi, H., Tai, Y. L., Peng, C. C., Tseng, C. Y., Chen, C. C., and Lin, C. Y. (2020) Impact of wearing masks, hand hygiene, and social distancing on influenza, enterovirus, and all-cause pneumonia during the coronavirus pandemic: Retrospective national epidemiological surveillance study, *Journal of Medical Internet Research*, 22, 8, e21257.
- Fang, Y., Nie, Y., and Penny, M. (2020) Transmission dynamics of the COVID-19 outbreak and effectiveness of government interventions: A data-driven analysis, *Journal of medical virology*, 92, 6, 645-659.
- Gakidou, E. (2020) Global projections of lives saved from COVID-19 with universal mask use.
- Husnayain A, Fuad A, Su E. C. (2020) Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan, *Int J Infect Dis.*, 95, 221-223.
- Jakkula, V. (2006). Tutorial on support vector machine (svm), School of EECS, Washington State University, 37.
- Jorden, M. A., Rudman, S. L., et al. (2020) Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020, *MMWR Morb Mortal Wkly Rep* 2020, 69, 680–684.
- Lee, S., and Fox, E. A. (2017) Archiving and Analyzing Tweets and Webpages with the DLRL Hadoop Cluster, *Bulletin IEEE Technical Committee Digital Libraries*, 13, 1.
- Lee, S., Elsherbiny, N., and Fox, E. A. (2012) A digital library for water main break identification and visualization, *12th ACM/IEEE-CS joint Conference on Digital Libraries*, 335-336.
- Lee, S., Farag, M., Kanan, T., and Fox, E. A. (2015) Read between the lines: A Machine Learning Approach for Disambiguating the Geo-location of Tweets, *15th ACM/IEEE-CS Joint Conf. on Digit. Libraries*, 273-274.
- Li, C., Chen, L. J., Chen, X., Zhang, M., Pang, C. P., and Chen, H. (2020) Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, *China*, 2020, *Eurosurveillance*, 25, 10, 2000199.
- Lin, Y.H., Liu, C.H., Chiu, Y.C. (2020) Google searches for the keywords of "wash hands" predict the speed of national spread of COVID-19 outbreak among 21 countries, *Brain, Behavior, and Immunity*, 87, 30-32.
- Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J. T., and Santillana, M. (2020) A machine learning methodology for real-time forecasting of the 2019-2020 COVID-19 outbreak using Internet searches, news alerts, and estimates from mechanistic models.
- Lyu, W., and Wehby, G. L. (2020) Community Use of Face Masks And COVID-19: Evidence from A Natural Experiment of State Mandates in the US: Study examines impact on COVID-19 growth rates associated with state government mandates requiring face mask use in public, *Health affairs*, 39, 8, 1419-1425.
- Mohamadou, Y., Halidou, A., and Kapen, P. T. (2020) A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19, *Applied Intelligence*, 50, 11, 3913-3925.
- Mukaka, M. M. (2012) Statistics corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research, *Malawi Med, J.* 24, 69–71.
- Mustaffa, Z., and Yusof, Y. (2010) A Comparison of Normalization Techniques in Predicting Dengue Outbreak.
- Nausheen, F., and Begum, S. H. (2018) Sentiment analysis to predict election results using Python, 2018 2nd International Conference on Inventive Systems and Control (ICISC), Coimbatore, 1259-1262.
- Nemes, L., and Kiss, A. (2020) Social media sentiment analysis based on COVID-19, *Journal of Information and Telecommunication*, 1-15.
- Oyebode, O., Ndulue, C., Adib, A., Mulchandani, D., Suruliraj, B., Orji, F. A., and Orji, R. (2020) Health, Psychosocial, and Social issues emanating from COVID-19 pandemic based on Social Media Comments using

- Natural Language Processing, 2007, 12144.
- Reuter, C., Backfried, G., Kaufhold, M., Spahr, F. (2018) ISCRAM turns 15: A Trend Analysis of Social Media Papers 2004-2017, Proceedings of the Information Systems for Crisis Response and Management (ISCRAM), 1-14, Rochester, New York, USA.
- Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., and Hassanien, A. E. (2020) Forecasting models for coronavirus disease (COVID-19): a survey of the state-of-the-art, SN Computer Science, 1, 4, 1-15.
- St. Denis, L. A., Hughes, A. L., Diaz, J., Solvik, K., Joseph, M. B., and Balch, J. K. (2020). 'What I Need to Know is What I Don't Know!': Filtering Disaster Twitter Data for Information from Local Individuals. In Amanda Hughes, Fiona McNeill, and Christopher W. Zobel (Eds.), ISCRAM 2020 Conference Proceedings – 17th International Conference on Information Systems for Crisis Response and Management, 730–743, Blacksburg, VA.
- Tang, L., Bie, B., Park, S. E., and Zhi, D. (2018) Social media and outbreaks of emerging infectious diseases: A systematic review of literature, American journal of infection control, 46, 9, 962-972.
- Thompson, C. N., Baumgartner, J., Pichardo, C., et al. (2020) COVID-19 Outbreak — New York City, February 29–June 1, 2020, MMWR Morb Mortal Wkly Rep, 69, 1725–1729.
- Yang, S., Chung, H., Lin, X., Lee, S., Chen, L., Wood, A., Kavanaugh, A. L., Sheetz, S. D., Shoemaker, D. J., Fox, E. A. (2013) PhaseVis: What, When, Where, and Who in Visualizing the Four Phases of Emergency Management Through the Lens of Social Media, Proceedings of the 10th International ISCRAM Conference, Baden-Baden, Germany, May 2013.