Getting Local and Personal: Toward Building a Predictive Model for COVID in Three United States Cities

1

April Edwards[1], Leigh Metcalf[2], William A. Casey[1], Shirshendu Chatterjee[3], Herralal Janwa[4], Ernest Battifarano[5]

Abstract

The COVID-19 pandemic was lived in real-time on social media. In the current project, we use machine learning to explore the relationship between COVID-19 cases and social media activity on Twitter. We were particularly interested in determining if Twitter activity can be used to predict COVID-19 surges. We also were interested in exploring features of social media, such as replies, to determine their promise for understanding the views of individual users. With the prevalence of mis/disinformation on social media, it is critical to develop a deeper and richer understanding of the relationship between social media and real-world events in order to detect and prevent future influence operations. In the current work, we explore the relationship between COVID-19 cases and social media activity (on Twitter) in three major United States cities with different geographical and political landscapes. We find that Twitter activity resulted in statistically significant correlations using the Granger causality test, with a lag of one week in all three cities. Similarly, the use of replies, which appear more likely to be generated by individual users, not bots or public relations operations, was also strongly correlated with the number of COVID-19 cases using the Granger causality test. Furthermore, we were able to build promising predictive models for the number of future COVID-19 cases using correlation data to select features for input to our models. In contrast, significant correlations were not identified when comparing the number of COVID-19 cases with mainstream media sources or with a sample of all US COVID-related tweets. We conclude that, even for an international event such as COVID-19, social media tracks closely with local conditions. We also suggest that replies can be a valuable feature within a machine learning task that is attempting to gauge the reactions of individual users.

Keywords

COVID-19 pandemic, feature selection, Granger causality, machine learning features, natural language processing, Pearson correlation, predictive modeling, regression analysis, social media mining, Twitter replies.

April Edwards (✉)
[1]United States Naval Academy, Cyber Science Department, Annapolis, MD, USA e-mail:
[2]Carnegie-Mellon University, Pittsburgh, PA
[3]Department of Mathematics, City University of New York, New York, NY (USA)
[4]Department of Mathematics, University of Puerto Rico, Rio Piedras, PR
[5]Finance Department, NYU School of Professional Studies, New York, NY

1 Introduction

Both individuals and organizations have been shown to use social media to manipulate public opinion in an attempt to influence political outcomes [5], [7], [10]. Furthermore, Twitter has been used as a predictor for real world events, such as vulnerability exploits [6]. In this article, we explore the relationship between social media activity on Twitter and a global event, the COVID-19 pandemic. In particular, we are interested in understanding how social media is used and consumed because this can greatly impact our understanding of how information is shared, which can lead to an improved understanding of how we combat the spread of misinformation and improve our ability to spread accurate information organically [20].

COVID-19 is a global pandemic that is experienced locally and regionally as well as globally. Rates of infection vary by country, state, region, and county. In the current study, we are interested in how social media correlates with outbreaks of COVID-19. In particular, we are interested in comparing the volume of COVID-19 social media posts to determine if increased social media activity can predict a potential outbreak in a particular region. Conversely, we are also interested in knowing if social media activity in a region increases in response to a local outbreak. In the current study, we first compare social media activity for three US cities: Miami, FL; Las Vegas, NV; and Seattle, WA, to determine correlations between social media activity and COVID-19 cases. Once features are identified, we explore their usefulness for predicting number of cases in the near term.

2 Data Sources

To determine the relationships between social media activity and COVID-19 cases, we needed multiple data sets. Data on, COVID-19 related tweets with location data was available from [9]. Huang et al. collected and posted COVID-related tweet IDs daily, beginning on Feb 6, 2020 [9]. The collection is regularly updated with additional data (as of the time of writing, data was available through Sept 30, 2022). Huang et al. select all tweets which contain one or more of the following terms in the tweet text or hashtags: coronavirus, wuhan, 2019ncov,sars, mers, 2019-ncov, wuflu, COVID-19, COVID19, COVID, covid-19,

covid19, covid, SARS2, and SARSCOV19. In addition to the tweet-id for each post, the data contains the date, keywords related to COVID-19, and the inferred geolocation, when available (country, state, and city).

We downloaded COVID-19 case data from USA FACTS [16]. USA Facts provides information on daily COVID-19 cases, hospitalizations, and deaths as reported by local public health agencies. Data is available for the entire US, and also disaggregated by state (including US territories), county, and city.

Mainstream news data was retrieved from the GDELT project [2]. GDELT consolidates data from broadcast, print, and online news sources. For the purposes of this project, we used a compilation of URLs and brief snippets of worldwide English-language news coverage mentioning COVID-19 from GDELT [2]. The normalized data of interest appears in Figure 1. At first glance there appear to be commonalities in graph shape, but these commonalities are not consistent. We therefore undertook a systematic approach toward understanding fluctuations in the data.

3 Correlations for Select US Cities

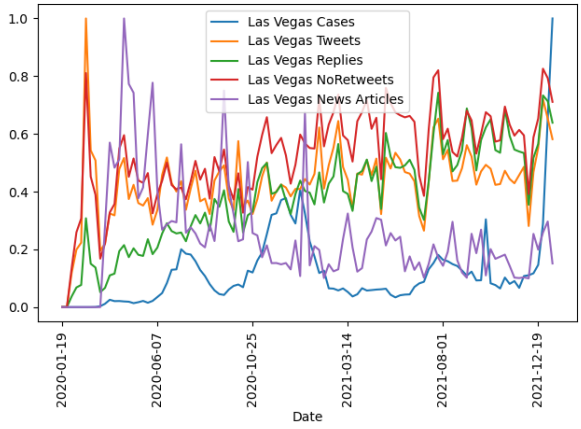
This section describes the methodology and results from our correlation studies.

3.1 Methodology

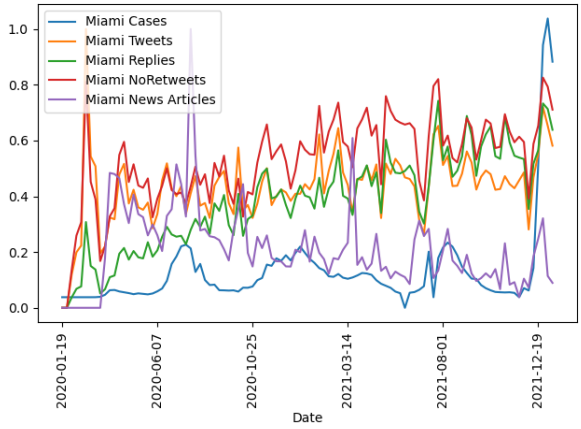
For purposes of analysis, we collated all data (news stories, social media activity, and COVID-19 cases) by week for three selected US cities in different regions and with different demographic and political profiles: Las Vegas, NV; Miami, FL; and Seattle, WA. These cities were selected because they are in different regions of the country and appeared in the top 15 for the number of Twitter posts related to COVID-19 over the period of interest (Table 1). We deliberately chose cities near the bottom of the top 15 under the assumption that data from the largest cities would not necessarily generalize, as their case averages are substantially higher than other cities on the list.

Weekly data was used because public health agencies vary in their reporting habits, with weekend data often reported on the following Monday or Tuesday. We also assume that both mainstream news (GDELT) and social media (Twitter) activity may vary based on the day of the week [1], [14]. For our study, we used data from the week of Jan 19, 2020 (the first US cases were reported in January 2020) to Jan 16, 2022 (the end date for the GDELT data). This is 105 weeks of data (just over 2 years).

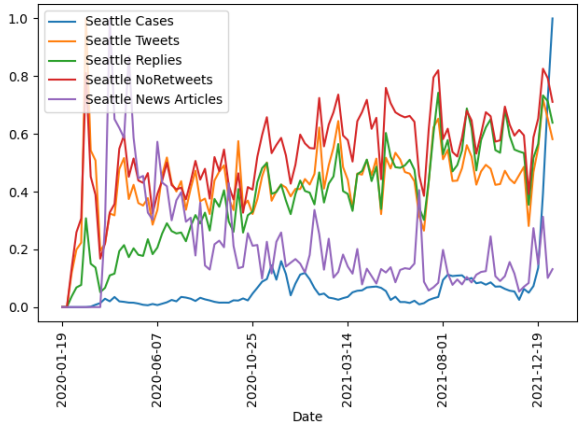
As noted above, the three selected cities also have different demographic profiles (Table 2). Miami is by far



(a) Cases, Social Media, and News for Las Vegas



(b) Cases, Social Media, and News for Miami



(c) Cases, Social Media, and News for Seattle

Fig. 1: Graphical Depiction of Fluctuations in Data by Week

Table 2: Demographic profile for selected cities as of July 1, 2021

City, State	Total Population	Age 18-64	White (not Hispanic)	Black	Hispanic	Bachelors or Higher
Las Vegas NV	646,790	61.4%	42.9%	12.1%	33.2%	25.2%
Miami FL	439,890	66.2%	11.5%	16.0%	72.5%	31.5%
Seattle WA	733,919	73.0%	62.6%	7.1%	7.1%	65.0%

Table 4: COVID-19 Twitter Data

Dataset	Num Tweets	Num Rehydrated	Num No Retweets	Num Replies Only
Miami	1,010,235	733,743	295,469	77,100
Seattle	1,197,313	904,401	338,047	146,217
Las Vegas	962,992	663,309	277,245	99,848
All US	2,700,000	2,027,919	790,502	308,278

Table 1: Top 15 Cities by Number of COVID-19 Tweets as of Jan 2022

City, State	Region	Avg daily COVID-19 Tweets
Los Angeles CA	West	5,404
New York NY	Northeast	5,325
Washington DC	Mid-Atlantic	5,301
Chicago IL	Midwest	3,653
Houston TX	South	2,471
Atlanta GA	South	2,298
San Francisco CA	West	1,793
Dallas TX	South	1,749
Boston MA	New England	1,745
Seattle WA	Northwest	1,741
Philadelphia PA	Mid-Atlantic	1,556
Austin TX	South	1,475
San Diego CA	West	1,473
Miami FL	South	1,467
Las Vegas NV	West	1,398

Table 3. Political profile for selected cities

County	%Republican (Trump)	%Democrat (Biden)
Clark County NV	44.3%	53.7%
Miami-Dade County FL	46.0%	53.3%
King County WA	22.2%	75.0%

the most diverse in terms of race/ethnicity, with only 11.5% of the population identifying as non-Hispanic white. Seattle has over double the percentage of adults 25 or older with a bachelor's or advanced degree in comparison with the other two cities. Las Vegas has more families with children (population age 18 or younger is 23.7%). (Demographic data from US Census Bureau [4].) Voting data were available at the county level, and, like most urban centers, all three cities favored the democratic candidate (Biden) in the 2020 US presidential election over the republican candidate (Trump). Seattle voted heavily democratic, while Las Vegas and Miami were more balanced politically (Table 3) [19].

For this project, we rehydrated all the tweets for Miami, Seattle, and Las Vegas, as well as a random sample of 100,000 tweets per month from over 137 million available US-based tweets. Some of the tweets could not be re-hydrated because they had been removed and some contained only partial geolocation data (country and state, or just country) or none at all. Table 4 shows the number of available tweets by location. Of those rehydrated, between 58 and 63 percent were retweets, and between 10 and 16 percent were replies.

We segregated the Twitter data into three data sets for each sample: All Tweets, No Retweets (i.e., all Tweets with retweets removed), and Replies Only. The Replies Only collection was of interest because it appeared that, while much of the Twitter data that was retrieved was from "corporatized" sources (i.e., news agencies, public health/governmental agencies, etc.), the replies appeared to be from individual users who were reacting to information about the pandemic.

For purposes of comparing social media with mainstream media, we retrieved news data from the GDELT project that contained references to our target cities between January 1, 2022 and January 16, 2022. We retrieved 68,307 instances for Las Vegas; 105,995 for Miami; and 70,801 for Seattle. Seattle recorded its first case in January 2020 and by April 17, 2022 had 388,271 total cases had been reported and recorded in the USA facts dataset. Las Vegas' first case was reported in early March 2020, and it had 537,776 cases as of April 17, 2022. Miami's reported its first case on March 12, 2022, and over 1.2 million cases had been reported as of June 2022. Reporting for Miami also switched to weekly mid-2021 and in one case the case count for Miami was altered due to reporting of a negative number of cases (on June 11, 2021).

Our motivating task was to predict the number of cases based on social media and/or news data. To identify features of interest, we first conducted a statistical analysis to determine features that appeared to be promising for use in the learning algorithms. As this is a predictive task with numeric output, we focused our machine learning experiments on linear, multilinear, and polynomial regression.

Statistical analysis was conducted in Python3 using the *pearsonr* function in *scipy* [17], and the *grangercausalitytests* function in *statsmodels* [13]. Regression function calculations relied on the

statmodels.api library, with preprocessing using the *sklearn Polynominal* function [3], [11]. We used *pandas* for efficient data storage and retrieval [18].

3.1 Results

3.1.1 Pearson Correlation Coefficient

Table 5 shows the results when social media and news data were correlated with COVID-19 cases for each of the three cities. The *lag* column was introduced to incorporate temporal components. A *lag* of 0 indicates the case data and social media/news data from the same week; a *lag* of 1 indicates that the tweet data preceded case data by 1 week; 2 is tweet data two weeks in advance and so on. Pearson coefficients between .5 and .7 are moderately correlated and coefficients above .7 are highly correlated. Cases with moderate or high correlation are indicated in red in Table 5. From the data it appears that *RepliesOnly* with a lag of 1 or 2 weeks appears to be the most promising feature for prediction. We also see a peak in the correlation coefficient at one week lag time, with correlation diminishing each week afterward. Interestingly, the GDELT had no correlation with the number of cases and none of the cities had a strong correlation between the case counts and the Twitter data from the all US sample.

3.1.2 Granger Causality

Granger causality [8] is used to determine if one time series data trend can forecast another. For this study, we use Granger Causality to determine if the number of cases are correlated using the social media or mainstream news data. We used the *grangercausalitytests* of the *statsmodels* [13] library in Python to determine if there is a granger causal relationship that may be used to predict the number of cases. The *statsmodels* documentation states [12]: "The Null hypothesis for *grangercausalitytests* is that the time series in the second column, *x2*, does NOT Granger cause the time series in the first column, *x1*. Granger causality means that past values of *x2* have a statistically significant effect on the current value of *x1*, taking past values of *x1* into account as regressors. We reject the null hypothesis that *x2* does not Granger cause *x1* if the *p*-values associated with the hypothesis tests corresponding to different lags are small (below a desired size of the test)."

The *grangercausalitytests* method accepts a maximum number of lags as an input parameter (along with the predictor, *x2* and the resultant, *x1*, data to be compared) and returns a *p*-value for each lag. The *p*-value measures the likelihood of the observed test statistic assuming that the null hypothesis is true. So, in our case, having a small *p*-value allows us to reject the null hypothesis and provides evidence that the ability of *x2* to predict *x1* is statistically significant. Table 6 shows the *p*-values for each input type and *lag*. Statistically significant results at the 90% level or above are shown in red.

Table 5: Pearson Correlation Coefficient, with lag

Miami					
Lag (weeks)	**All Tweets**	**No Re-tweets**	**Replies Only**	**Re-tweets Only**	**GDELT**
0	0.45	0.47	0.57	0.43	0.06
1	0.54	0.55	0.61	0.51	0.08
2	0.44	0.44	0.51	0.42	0.09
3	0.26	0.26	0.35	0.25	0.00
4	0.10	0.10	0.21	0.09	-0.08
Seattle					
Lag (weeks)	**All Tweets**	**No Re-tweets**	**Replies Only**	**Re-tweets Only**	**GDELT**
0	0.17	0.25	0.34	0.12	-0.12
1	0.36	0.47	0.55	0.30	-0.09
2	0.32	0.41	0.52	0.26	-0.07
3	0.23	0.31	0.43	0.18	-0.06
4	0.11	0.19	0.33	0.07	-0.11
Las Vegas					
Lag (weeks)	**All Tweets**	**No Re-tweets**	**Replies Only**	**Re-tweets Only**	**GDELT**
0	0.41	0.49	0.49	0.32	0.01
1	0.63	0.69	0.70	0.53	-0.07
2	0.61	0.67	0.68	0.52	-0.06
3	0.52	0.58	0.62	0.43	-0.08
4	0.38	0.45	0.51	0.29	-0.09

Table 6: Granger Causality p-Value with Lag

Miami					
Lag (weeks)	**All Tweets**	**No Re-tweets**	**Replies Only**	**Retweets Only**	**GDELT**
1	0.050	0.081	0.095	0.046	0.773
2	0.120	0.127	0.099	0.146	0.519
3	0.283	0.3 03	0.082	0.325	0.303
4	0.307	0.397	0.200	0.325	0.406
Seattle					
Lag (weeks)	**All Tweets**	**No Re-tweets**	**Replies Only**	**Retweets Only**	**GDELT**
1	0.109	0.108	0.036	0.130	0.746
2	0.184	0.040	0.003	0.334	0.736
3	0.370	0.087	0.006	0.578	0.816
4	0.262	0.074	0.032	0.449	0.823
Las Vegas					
Lag (weeks)	**All Tweets**	**No Re-tweets**	**Replies Only**	**Retweets Only**	**GDELT**
1	0.005	0.007	0.000	0.010	.709
2	0.004	0.001	0.000	0.022	.678
3	0.003	0.000	0.000	0.021	.852
4	0.008	0.001	0.000	0.045	.926

We conclude from these data that there are strong relationships between local Twitter activity and COVID-19 cases for these three cities. As with the Pearson correlations, the strongest predictor is at the individual (RepliesOnly) level, although it is clear that original tweets also play a role, at least for Seattle and Las Vegas. What is interesting is that retweets (which are a measure

of engagement) do not play a strong role, with the exception of Las Vegas. Once again, mainstream news sources have no Granger causal relationship with the COVID-19 case counts. While aspects of the All US Sample data do show Granger causal relationships with case counts, there is no commonality across cities (data not shown due to space limitations).

In analyzing these data, a question arises regarding the impact of cases on news/social media activity. To test the hypothesis that case counts may be Granger causal related to mainstream news, we reversed the data inputs to the Granger Causal Test functions (i.e., used case counts to predict news stories and social media activity). These experiments showed an even stronger Granger causal relationship between social media and COVID-19 cases (which makes sense, as both organizations and individuals respond to local conditions). Unsurprisingly, the All US Sample did not show a pattern of Granger causal relationships. Surprisingly, however, mainstream news sources for individual cities also did not demonstrate a Granger causal relationship. We believe this occurred because news sources, even those originating in the selected cities, tend to focus on the national and international news. Perhaps the phrase "all news is local" is, in fact, is more true for social media than mainstream news, even in a global pandemic!

3.1.3 Predictive Model

A particular item of interest was the usefulness of the correlation results for determining the best features to use to predict the number of cases based on Twitter activity at a local level. For the discussion below we restrict our discussion to Las Vegas because the correlation and Granger causality results were exceptionally promising. The results were similar for all cities.

Figures 2, 3, and 4 show a typical set of results for multilinear and polynomial regression using different features. In all cases we trained on the first 80 weeks of available data and tested using the subsequent five weeks (roughly trying to predict a month in advance). The results did not change dramatically as long as we used at least 60 weeks of training data (and always predicting 5 weeks). Figure 2 shows the results of the multi-linear model using Replies Only, No Retweets, and All Tweets as our predictors and lag of 1, as these were the best values for our correlations in Table 5. In Figure 3 we approximate the model based on the Granger causality in Table 6 by building a multilinear regression model using four weeks of prior data for both cases and Replies Only. In Figure 4 we reverted back to using Replies Only, No Retweets, and All Tweets as our predictive variables in a polynomial regression model with a lag of 1 and a degree of 2. The polynomial regression was by far the most promising among the models tested. Increasing the degree did not improve the predictive outcomes.

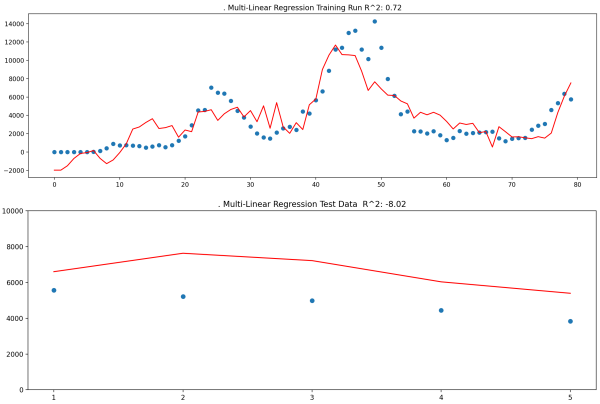


Fig. 2: Multilinear Regression with Replies Only, No Retweets, and All Tweets for Las Vegas. Training Model (top) vs. Test Results

4 Conclusions

In [15], the authors identified 81 articles that analyzed social media communication surrounding the COVID-19 pandemic. The authors concluded that there was a lack of machine learning applications that use social media data for prediction of cases during the COVID-19 pandemic. Tsao, et al. also concluded that there was little evidence that social media data was used for real-time surveillance. In this article we fill a crucial gap in our collective understanding of Twitter activity in the context of a global event, such as COVID-19. Specifically, we show that responses from individual users in a geographic area are strongly correlated with local case counts for COVID-19. We also show that Twitter was a better source for understanding the impact of COVID-19 on a community than mainstream news data. These are novel findings that identify potential new features, such as Replies Only that should be explored when conducting machine learning research using social media data. Use of these features in multilinear and polynomial regression models shows promise for predicting COVID-19 case counts for a month in advance.

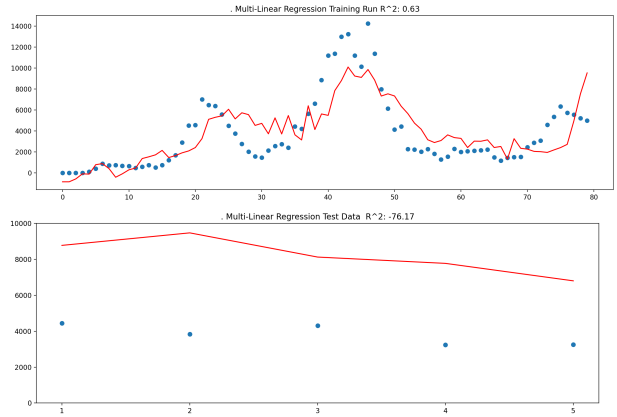


Fig. 3: Multilinear Regression mimicking Granger Causality for Las Vegas. Training Model (top) vs. Test Results (bottom)

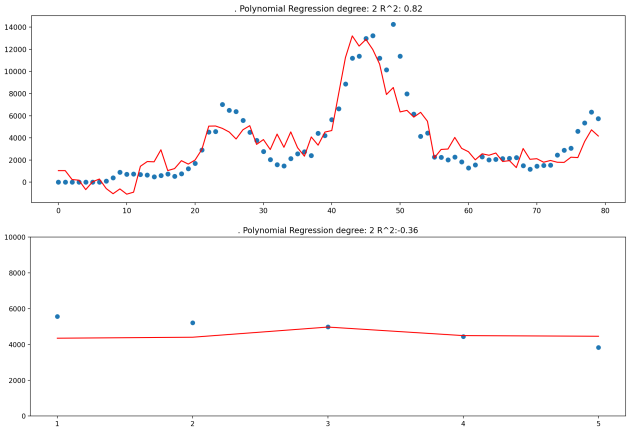


Fig. 4: Polynomial Regression with Degree 2 and Replies Only, No Retweets, and All Tweets for Las Vegas. Training Model (top) vs. Test Results (bottom)

With the prevalence of mis/disinformation on social media [20], it is critical to develop a deeper and richer understanding of the relationship between social media and real-world events in order to detect and prevent future influence operations. This project contributes to our understanding of how social media activity within a geographic region tracks with real-world events and identifies both novel feature sets and new approaches to applying machine learning to these tasks.

Acknowledgements. Janwa's research was supported in parts by the NASA grants 80NSSC21M0156 and 80NSSC22M0248. Edwards and Casey were funded in part by the Office of Naval Research. The views expressed are those of the authors and do not reflect the official policy or position of the United States Naval Academy, United States Navy, United States Marine Corps, the Department of Defense or the United States Government.

References

1. Avraam, E., Veglis, A., Dimoulas, C. News article consumption habits of Greek internet users. In 6th Annual International Conference on Communication and Management (ICCM2021), Athens, Greece, August, pages 1–5, 2021.
2. The GDELT Project Blog. Now live updating & expanded: A new dataset for exploring the coronavirus narrative in global online news.
3. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pages 108–122, 2013.
4. U.S. Census Bureau. Quickfacts. https://www.census.gov/quickfacts/fact/table/US/PST045221, 2021.
5. Cadwalladr, C., Graham-Harrison, E. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17:22, 2018.
6. Chen, H., Liu, R., Park, N., Subrahmanian, V.S. Using Twitter to predict when vulnerabilities will be exploited. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '19, page 3143–3152, New York, NY, USA, 2019. Association for Computing Machinery.
7. Dandekar, A., Narawade, V. Twitter sentiment analysis of public opinion on COVID-19 vaccines. In Computer Vision and Robotics, pages 131–139. Springer, 2022.
8. Granger, C.W. Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society, pages 424–438, 1969.
9. Huang, X., Jamison, A., Broniatowski, D., Quinn, S., Dredze, M. Coronavirus Twitter Data: A collection of COVID-19 tweets with automated annotations, March 2020. http://twitterdata.covid19dataresources.org/index.
10. Isaak, J., Hanna, M.J. User data privacy: Facebook, Cambridge Analytica, and privacy protection. Computer, 51(8):56–59, 2018.
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D. Brucher, M., Perrot, M., Duchesnay, E.. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.
12. Perktold, J., Seabold, S., Taylor, J. statsmodels.tsa.stattools.grangercausalitytests.https://www.statsmodels.org/dev/generated/statsmodels.tsa.stattools.grangercausalitytests.html.
13. Seabold, S.,Perktold, J. statsmodels: Econometric and statistical modeling with Python. In 9th Python in Science Conference, 2010.
14. Singh, C. What is the best time to post on Twitter in 2022? SocialPilot, 2022.
15. Tsao, S.F., Chen, H., Tisseverasinghe, T., Yang, Y., Li, L., Butt, Z.A.. What social media told us in the time of covid-19: a scoping review. The Lancet Digital Health, 3(3):e175–e194,2021.
16. USAFACTS. US COVID-19 cases and deaths by state. https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/.
17. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, I., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods, 17:261–272, 2020.
18. McKinney, W.. Data Structures for Statistical Computing in Python. In Stefan van der Walt and Jarrod Millman, editors, Proceedings of the 9th Python in Science Conference, pages 56 – 61, 2010.
19. Wikipedia. 2020 United States Presidential Election. https://en.wikipedia.org/wiki/2020 United States presidential election, 2020.
20. Chittari, R., Nistor, M.S., Bein, D., Pickl, S., Verma, A. (2022). Classifying Sincerity Using Machine Learning. In: Latifi, S. (eds) ITNG 2022 19th International Conference on Information Technology-New Generations. Advances in Intelligent Systems and Computing, vol 1421. Springer, Cham. https://doi.org/10.1007/978-3-030-97652-1_31