MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220170

Linking Tweets Towards Geo-Localized Policies : COVID-19 Perspective

Sheikh Saifur Rahman Jony[a], Rifat Shahriyar[a], M. Saifur Rahman[a], M. Sohel Rahman[a] , Tanvir Alam[b]

[a] Department of CSE, Bangladesh University of Engineering and Technology, West Palashi, Dhaka 1205, Bangladesh
[b] College of Science and Engineering, Hamad Bin Khalifa University, Doha 34110, Qatar

Abstract

COVID-19 pandemic is taking a toll on the social, economic, and psychological well-being of people. During this pandemic period, people have utilized social media platforms (e.g., Twitter) to communicate with each other and share their concerns and updates. In this study, we analyzed nearly 25M COVID-19 related tweets generated from 20 different countries and 28 states of USA over a month. We leveraged sentiment analysis and topic modeling over this collection and clustered different geolocations based on their sentiment. Our analysis identified 3 geo-clusters (country- and US state-based) based on public sentiment and discovered 15 topics that could be summarized under three main themes: government actions, medical issues, and people's mood during the home quarantine. The proposed computational pipeline has adequately captured the Twitter population's emotion and sentiment, which could be linked to government/policy makers' decisions and actions (or lack thereof). We believe that our analysis pipeline could be instrumental for the policymakers in sensing the public emotion/support with respect to the interventions/actions taken, for example, by the government instrumentality.

Keywords:

COVID-19, Coronavirus, Twitter

Introduction

The first confirmed case of COVID-19 was reported in November 2019 at Wuhan, China [1]. Within two months of the first reported case, the Chinese government started to issue lockdown in several states. COVID-19 started to spread rapidly across the globe and caused thousands of fatalities within a few months [2]. WHO declared COVID-19 a pandemic on March 11, 2020 [3]. As of December 29, 2020, the number of confirmed COVID-19 cases was 81.67M with 1.78M deaths worldwide, affecting 220 countries and USA is the most affected country with 19.78M confirmed cases and 343.18k deaths [4]. Since the transmission of COVID-19 takes place government of different countries started to issue and enforce lockdown directives as soon as COVID-19 confirmed cases started to increase. Offices and educational institutions were closed indefinitely and people were advised to shelter in their homes with a goal to enforce social distancing. As people could not physically meet others due to social distancing, social media platforms became the mainstay for their communication and their activity in social media increased. As a result, people started to express their views, concerns, status, change of interest in lockdown, panic etc. with respect to COVID-19 issues on social media platforms. For example, in a recent survey, carried out over 4500 respondents in North America showed that 72% of the respondents believe that their social media consumption has increased during the pandemic [5].

Twitter is one of the most popular social media platforms having more than 330M active users at the first quarter of 2019 [6]. Many people share their various activity information and express their mood, views etc. towards different events through Twitter. Therefore Twitter has also experienced heightened usage by its users during the pandemic period. Many research works have been reported in the literature that focus on analyzing public sentiments, behaviour etc. based on COVID-19 related tweets. For example, Jain *et al.* developed a method to identify the most influential Twitter users in COVID-19 [7], where methods have been. Similar angle has also been pursued by Haman [8], albeit with the twist that he analyzed the tweets done by the state leaders along with the public's reaction thereto. Similar study has also been carried out on a limited setting - focusing only on G7 leaders - by Rufai and Bunce [9]. The power of Twitter in tracking epidemic outbreaks has also been investigated in the literature. For example, Abouzahra and Tan [10] have investigated the potential of Twitter in augmenting the epidemic tracking systems in the context of Zika epidemic. Potential of Twitter as a crisis communication medium during natural disasters has also been investigated in the literature (e.g., [11] for Hurricane Harvey; [12] for Hurricane Sandy etc.). Since a comprehensive literature review on research work on Twitter is out of the scope of this work, the readers are kindly referred to a recent survey there on [13].

In this work, we analyzed tweet data linking public emotions towards geo-localized policymakers' decision. We developed a pipeline where tweet data to capture public emotion to aid the policymakers from multiple geo-locations in assessing the impact of their decisions/actions (or lack thereof). Our work is unique and differs from previous works in multiple directions. We analyzed the tweets for different geolocations (countries and US states). Then, we studied their emotions and cluster the locations that seem to share the same emotional state concurrently, and identify the possible reason therefore. We also analyzed sudden up and down trends and make an effort to explain such phenomena. We identified and analyzed the most prominent topics people discussed and how they showed their emotions towards those.

Methods

Data Collection

We used the twitter dataset (referred to as the Lamsal dataset henceforth) reported in [14]. Lamsal dataset is a collection of English-language tweets related to COVID-19 tweeted by people all over the world. The dataset is continuously being updated with everyday data starting from March 19, 2020. We collected 24.5M tweet IDs from the Lamsal dataset covering tweets related to COVID-19 from March 19, 2020 to April 17, 2020 and were able to extract country information from 9.51M of those. From the corresponding countries, to ensure a meaningful analysis, we only chose countries having at least 500 tweets per day. Based on this threshold, we were able to study tweets from 20 different countries. We also worked on a subset of the tweets originating from the USA only. We found

2.74M tweet IDs from USA during this period. Considering the same threshold we were able to extract tweets form 28 states of the USA. The number of tweets collected from different countries and US states are shown in Supplementary Figure S1. We used Twitter standard search API to collect all the fields related to tweets. We also used the sentiment value for every tweet computed by Lamsal [14], which ranges from -1 (very negative) to 1 (very positive). The tweet fields considered in this study are reported in Supplementary Figure S2. Informatively, fields like coordinates and places that are indicative of the users' locations were found to be null for almost all the records and hence could not be used for location extraction. As an alternative, we used the optional "user_location" field to estimate the location of a user.

Data Preprocessing

We extracted date and time information for each tweet from the *created_at* field. We removed URLs and *user_name* mentioned in the tweet text and derived a new field, *cleaned_text*. Then we removed stopwords and tokenized each tweet. We applied lemmatization on each of the tokens. We used Spacy library for these pre-processing tasks. The *user_location* field of a tweet contains user provided information about his/her location and contains a lot of imaginary and comical locations. We used regular expressions, Geotext, and Flashgeotext libraries for extracting the corresponding city, state and country for each tweet user. In a few cases, manual intervention was needed due to the erroneous nature of this field (*user_location*).

Sentiment based clustering of Geolocations

We grouped tweets from each of the 20 countries (or 28 states of the USA) under consideration by date. At first we calculated average daily sentiment (i.e., arithmetic mean of sentiment values) of each country, albeit with little or no meaningful insight. So, we proceeded along a different path as follows. We treated each country as a 30D (thirty-dimensional) data-point, where each dimension corresponds to the average daily sentiment of the tweets for a particular day of the month. Then we ran K-Means clustering algorithm [15] on the datapoints. We initialized our cluster centroids using K-Means++ algorithm [16] and ran the model 10 times using different centroid seeds to mitigate the effect of local minima [17]. As the distance measure, we used Within-Cluster-Sum-of-Squares (WCSS) measures to calculated distance of all the points within a cluster to the centroid of corresponding cluster.

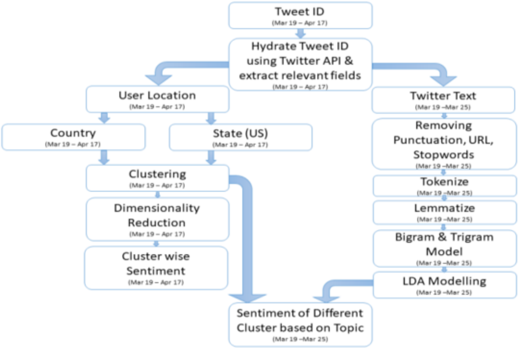


Figure 1– Workflow of data preprocessing

To determine the number of clusters (K), we used the Elbow method [18]. In our experiments on the 20 countries (28 states of the USA), $K = 3$ formed an elbow with a WCSS value of 0.48 (0.186). Hence, we chose 3 as our number of clusters in both cases. For the visualization of the clusters, we reduced 30 dimensional datapoints into two dimensions using Principal Component Analysis (PCA)[19] for the US states and t-SNE [20] for the countries. Figure 1 highlights the computational pipeline that was used in this study.

Topic Modeling

Firstly, we trained bigram and trigram models on the tokens. These tokens were subsequently used for topic modeling. We trained a Latent Dirichlet Allocation (LDA) topic model on the dataset, albeit on a smaller time range (due to computational resource constraints): from March 19, 2020 to March 25, 2020. We used the coherence score to find the suitable number of topics. This helps to distinguish topics that are semantically interpretable. For the visualization of LDA based results, we created an Intertopic Distance Map of the topics using multidimentional scaling using pyLDAvis [21]. Additionally, we merged the topics into different exciting themes and consequently inferred every tweet to a specific theme. Finally we analyzed topical theme wise sentiment for the clusters that we had determined for different states of USA.

Results

Sentiment based Geo-clustering

We discovered three different geo-clusters (country- and US state-based) according to our analysis. The first country cluster (C-A) comprises only two Asian countries, namely, UAE and Indonesia. The second country cluster (C-B) is mainly formed by some African countries, namely, Ghana, Kenya, Nigeria and Uganda; but Uruguay and Portugal are also in this cluster. The third country cluster (C-C) is diverse and it covers 12 countries across multiple continents including Europe, Asia, Australia, North America and South America. The clustering of the 28 states of the USA (Clusters S-A, S-B and S-C) found under consideration are depicted in Table 1.

Table 1– Country/States Assigned to different Cluster

Cluster	Country/States
C-A	UAE, Indonesia
C-B	Ghana,Kenya,Nigeria,Portugal,Uganda,Uruguay
C-C	Australia, Brazil, Canada, Germany, France, UK, India, Malaysia, Philippines, Pakistan, USA, South Africa
S-A	California,Pennsylvania,Ohio,Michigan,Arizona ,Nevada, Alabama, Indiana, Missouri, South Carolina,Connecticut,Oklahoma,Kentucky
S-B	Texas,Florida,Illinois, Georgia, Maryland, North Carolina, New Jersey, Louisiana, Tennessee
S-C	New York, Washington, Massachusetts, Oregon, Minnesota, Colorado

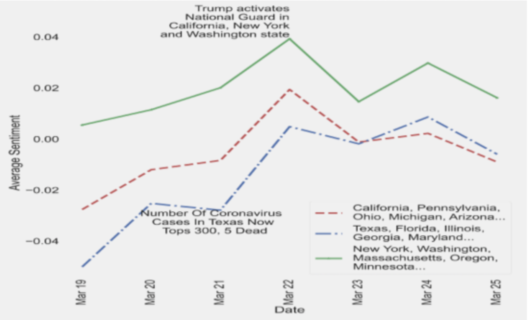
Topic Modeling and Thematic Analysis

We found 15 topics having the best coherence score of 0.43. These topics were grouped into three different thematic areas representing, respectively, people's sentiment towards different government actions, contemporary medical issues and home quarantine.

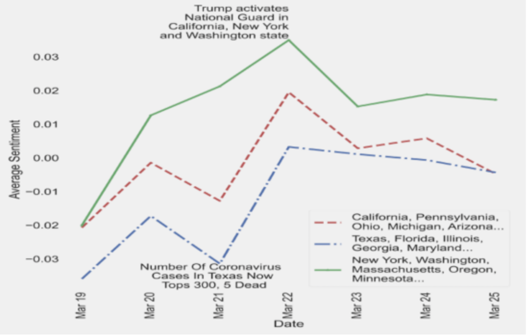
Theme 1: Government Actions : Some of the discovered topics were heavily related to different keywords about governmental issues. The word cloud and a few tweets under this theme are reported in Supplementary Figure S3. Figure 2a presents cluster-wise average sentiment values of the US states considering only the tweets related to this theme. As can be seen from the figure, for the green curve (corresponding to Cluster S-C), there is a sharp rise in the overall sentiment till the mid of the week following which there is a slight decline albeit with a slightly upward trend again at the end. Interestingly, all three curves (i.e., representing three clusters of the states of the USA) register a negative sentiment on March 19, which changes to a positive sentiment on March 22.

Theme 2: Medical Issues : Figure 2b presents cluster-wise average sentiment values of the US states considering only the tweets related to the current theme. As can be seen from the figure, the green curve corresponding to states like New York, Washington etc. remained on the top of other curves for the whole time period. For the other two curves, there was a sharp decline of sentiment on March 21, albeit followed by an immediate rise and subsequent fluctuations. Interestingly, all three curves registered a negative sentiment on March 21, which changed to a positive sentiment on March 22 with some mild fluctuations afterwards. The word cloud and a few tweets under this theme are reported in Supplementary Figure S3.

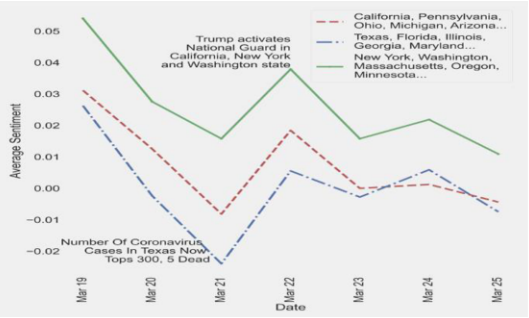
Theme 3: Home Quarantine : Some topics we discovered were related to people's mood swing during home quarantine. The word cloud and a few tweets under this theme are reported in Supplementary Figure S3. Figure 2c presents cluster-wise average sentiment values of the US states considering only the tweets related to the current theme. Here as well, the green curve remained on the top of other curves for the whole time period. Similar to the situation in Theme 2 (Medical Issues), all three curves registered a negative sentiment on March 21, which changed to a positive sentiment on March 22 with some fluctuations afterwards.



(a) Government Issues



(b) Medical Issues



(c) Home Quarantine

Figure 2 – Average Sentiment of different themes.

To test whether our pipeline generalizes, we collected another month of data (OCT 16, 2020 - NOV 15, 2020) keeping US election the middle. Social media is supposed to show a great amount of fluctuation during this time period. Figure 3 is a plot of average sentiment of different states of the USA. Here, without any complicated clustering, we just plotted average sentiment of top 5 states. We can see that the plot show highly uptrend in sentiment on 9 November when BioNTech said that their vaccine is more than 90% successful [22]. This clearly indicates that our pipeline was able to capture the mood swing from tweets.

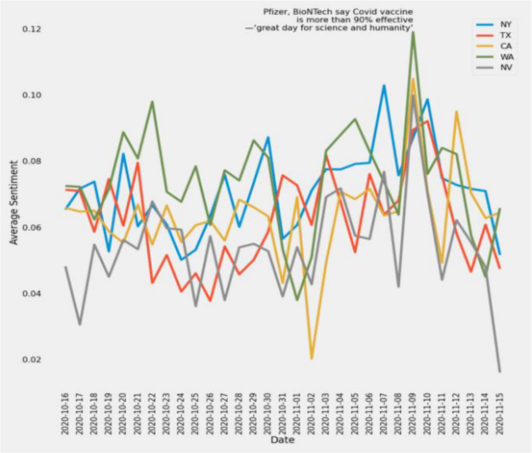


Figure 3 – Average Mood Swing of Different States of US.

Discussion

The three country clusters (Clusters C-A, C-B and C-C) found in our analysis provide us with some interesting insight when we analyze the confirmed case counts (Figure 4) as well as measures taken by the respective countries during the period under consideration (referred to as the *Analysis Period* henceforth) as follows. Cluster C-B consists of some African countries (i.e., Nigeria, Kenya, Uganda and Ghana), Uruguay and Portugal. Nigeria, Kenya, Uganda and Ghana did not suffer severely from COVID-19 during the Analysis Period which was reflected in the sentiment values of the corresponding tweets and was well-captured by our model. Uruguay was also not very much affected during that period. However the inclusion of Portugal in this cluster by our model is apparently surprising as Portugal was already well-affected by COVID-19 during the Analysis Period. Interestingly however, we found that Portugal took some exemplary measures to fight COVID-

19 [23,24]. These good measures, as it seems, have swayed the sentiments of the Twitter users.

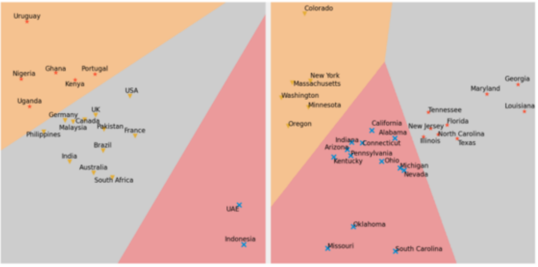
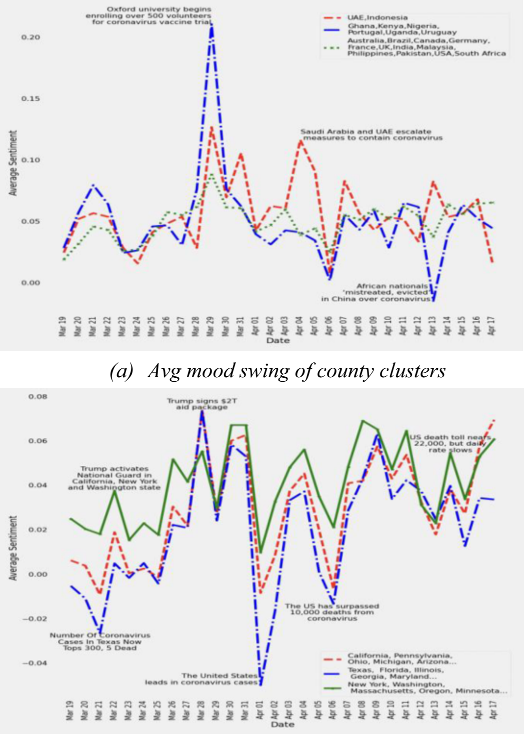


Figure 4 – Clustering of Countries and US States

On the other hand, Cluster C-C consists of the countries that were extremely affected by COVID-19 during the Analysis Period. Thus we hypothesize that this affected the mental condition of the people which got reflected in their tweet sentiments and our model was able to capture that well. Considering the cumulative confirmed case, United Arab Emirates (UAE) and Indonesia were somewhere in between, i.e., they had quite a good number of confirmed COVID-19 cases. However, they also took some strong interventions (e.g., [25-26]). Our model thus indicates that the strong intervention measures taken by these two countries have made them unique and put them in a separate cluster (Figure 4).



(a) Avg mood swing of county clusters

(b) Avg mood swing of US state clusters

Figure 5 – Average Mood Swing

In general, public sentiment changed according to how their country was affected and the relevant intervention measures taken (Figure 5a) and that has been well-captured in our model. On March 29, their is a sharp peak for Cluster C-B who were not affected much or took good measures during the Analysis Period. Also, on that day, Oxford University announced the starting of enrolling volunteers for vaccine trials [27], which, as a globally positive news, might have contributed towards that peak. But it didn't please Cluster C-C countries, where people were severely affected (Figure 5a). There were some different random fluctuations too depending on the concurrent situations. Noticeably, on April 13, Cluster C-B showed a highly negative sentiment (Figure 5a) as African nationals were "mistreated, evicted" in China over coronavirus issues [28].The three state clusters (Clusters S-A, S-B and S-C) found in our analysis are also quite interesting as follows (Figure 5b). Cluster S-B consists of states like Texas, Florida, Illinois and so on. For these states, we notice a very high fluctuation in overall sentiments during the Analysis Period. For example, on March 21, 2020, we notice an overall negative mood, which could be attributed to the fact that, on that day, 300 people were newly infected with 5 registered deaths [29]. But states belonging to cluster S-C (i.e., New York, Washington etc.) exhibited an overall positive mode during the same period which could be attributed to the strong measures taken by the respective state authorities (e.g., [30]). Finally, Cluster S-A consists of states like California, Pennsylvania, Ohio and so on. This cluster shows a sentiment value that is almost an average of the other two clusters.

Noticeably, overall positive mood can be noticed on March 28, 2020 in all clusters follows (Figure 5b), most likely in response to a country-wide strong measurement initiative on that day: a $2 Trillion bill was signed by President Trump to fight against corona virus [31]. On the other hand, on April 1, 2020, the White House briefed that the projected COVID-19 deaths of USA would be within the range 100K to 240K [32] and USA had surpassed all nations on COVID-19 case count but not in response against the pandemic [33]; all the states showed very large negative spike on that day. There is also a negative spike in all clusters on April 06, 2020 when the USA surpassed the 10k death mark [34].

Limitations

Our study has the limitation that we have used dataset for small timeframe (over a month). However, this suits our main goal, i.e., to establish the fact that such a tweet analysis pipeline is worth-adopting to gauge the public sentiment and emotions. Also, one could argue that the dataset is outdated. In that regard, we rebut that this was a deliberate choice as various events happened during that period that were responsible for public mood swing. And we have briefly shown that our pipeline is equally effective on data from a different and more recent timeline.

Conclusions

Our analysis reveals that people's emotions during the COVID 19 pandemic period are largely affected by different actions/measures taken by the government and news published about COVID 19 fatality and these emotions are reflected through their tweets. Evidently, the approach and methodology employed in this research to analyze the tweets was able to capture the true emotion of the people geo-locally. This suggests that such an approach and methodology could aid the policy makers to analyze and monitor the effect of different interventions/actions or lack thereof on the general mass.

References

[1] 1st known case of coronavirus traced back to November in China; https://www.livescience.com/first-case-coronavirus-found.html

[2] WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020;. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

[3] World Health Organization. 2020. Rolling updates on a coronavirus disease (COVID-19);. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen [accessed 2020-04-10].

[4] worldometer. COVID-19 Coronavirus Pandemic; 2020. https://www.worldometers.info/coronavirus/

[5] COVID-19 is changing how, why and how much we're using social media; https://www.digitalcommerce360.com/2020/09/16/covid-19-is-changing-how-why-and-how-much-were-using-social-media/

[6] Number of monthly active Twitter users worldwide from 1st quarter 2010to 1st quarter 2019;. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

[7] S. Jain, A. Sinha, Identification of influential users on twitter: A novel weighted correlated influence measure for covid-19, Chaos, Solitons & Fractals. 139 (2020) 110037.

[8] M. Haman, The use of twitter by state leaders and its impact on the public during the covid-19 pandemic, Heliyon. 6 (2020) e05540.

[9] S.R. Rufai1, C. Bunce, World leaders' usage of twitter in response to the covid-19 pandemic: A content analysis, J Public Health.

[10] M. Abouzahra, J. Tan, Twitter vs. Zika—the role of social media in epidemic outbreaks surveillance, Health Policy and Technology. (2020).

[11] C.M. Vera-Burgos, D.R. Griffin Padgett, Using twitter for crisis communications in a natural disaster: Hurricane harvey, Heliyon. 6 (2020) e04804.

[12] B. Wang, J. Zhuang, Crisis information distribution on Twitter: a content analysis of tweets during Hurricane Sandy, Natural Hazards: Journal of the International Society for the Prevention and Mitigation of Natural Hazards. 89 (2017) 161–181.

[13] D. Antonakaki, P. Fragopoulou, S. Ioannidis, A survey of twitter research: Data model, graph structure, sentiment analysis and attacks, Expert Systems with Applications. 164 (2021) 114006.

[14] R. Lamsal, Coronavirus (covid-19) tweets dataset, (2020). https://doi.org/10.21227/781w-ef42.

[15] S.P. Lloyd, Least squares quantization in pcm, Information Theory, IEEE Transactions. (1982).

[16] S. Arthur D.; Vassilvitskii, K-means++: The advantages of careful seeding, (2007).

[17] P.S.B.U.M. Fayyad, Refining initial points for k-means clustering, (1998).

[18] R.L. Thorndike, Who belongs in the family?, (1953).

[19] K.P.F.R. S., LIII. On lines and planes of closest fit to systems of points in space, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 2 (1901) 559–572.

[20] L.J.P. van der Maaten, G.E. Hinton, Visualizing high-dimensional data using t-sne, (2008).

[21] pyLDAvis;. https://github.com/bmabey/pyLDAvis

[22] CNBC. Pfizer, BioNTech say Covid vaccine is more than 90% effective |'great day for science and humanity';. https://www.cnbc.com/2020/11/09/covid-vaccine-pfizer-drug-is-more-than-90percent-effective-in-preventing-infection.html

[23] Swift action kept portugal's coronavirus crisis in check, says minister, https://www.theguardian.com/world/2020/apr/19/swift-action-kept-portugals-coronavirus-crisis-in-check-says-minister

[24] COVID-19: Portugal approves extraordinary measures for employees and employers;. https://www.lexology.com/library/detail.aspx?g=dce2fe1b-fb3a-4658-a07e-5e7e17cec637

[25] Firstpost. Coronavirus Outbreak: UAE extends curfew as COVID-19 cases increase, warns of reviewing labour ties with countries refusing to take back citizens;. https://www.firstpost.com/health/coronavirus-outbreak-uae-extends-curfew-as-covid-19-cases-increase-warns-of-reviewing-labour-ties-with-countries-refusing-totake-back-citizens-8254491.html

[26] COVID-19 in Indonesia is Underreported and Under-Addressed; https://www.futuredirections.org.au/publication/covid-19-in-indonesia-is-underreported-and-under-addressed/

[27] Covid-19 vaccine trial on humans starts as UK warns restrictions could stay in place until next year; https://edition.cnn.com/2020/04/23/health/coronavirus-vaccine-trial-uk-gbr-intl/index.html

[28] African nationals 'mistreated, evicted' in China over coronavirus;. https://www.aljazeera.com/news/2020/4/12/african-nationals-mistreated-evicted-in-china-over-coronavirus.

[29] CBS. Number Of Coronavirus Cases In Texas Now Tops 300, 5 Dead;. https://dfw.cbslocal.com/2020/03/21/number-of-coronavirus-cases-in-texas-now-tops-300-5-dead/

[30] CNN. Cuomo orders all nonessential New York workers to stay home; https://edition.cnn.com/2020/03/20/politics/new-york-workforce-stay-home/index.html

[31] Today U. Coronavirus updates: Trump signs $2T aid package; US tops 100K cases; Disney parks close indefinitely;. https://www.usatoday.com/story/news/health/2020/03/27/coronavirus-live-updates-stimulus-vote-us-deaths-china-population/2922314001/

[32] ALJAZEERA. White House projects 100,000-240,000 US deaths from coronavirus;. https://www.aljazeera.com/news/2020/03/white-house-projects-100000-240000-deaths-coronavirus-200331222101557.html

[33] Magazine S. The United States leads in coronavirus cases, but not pandemic response;. https://www.sciencemag.org/news/2020/04/united-states-leads-coronavirus-cases-not-pandemic-response

[34] Today U. The US has surpassed 10,000 deaths from coronavirus. That's more than the total battlefield fatalities from six combined wars;. https://www.usatoday.com/story/news/health/2020/04/06/coronavirusdeath-toll-us-reaches-10000-six-wars-combined/2949285001/

Supplementary File

Supplementary Files are shared in GitHub:
https://github.com/tanviralambd/COVID-Geo-Cluster

Address for correspondence

Tanvir Alam, College of science and engineering, Hamad Bin Khalifa University, Doha, Qatar.

Email: talam@hbku.edu.qa; Phone: +974-4454-2277