Research Article

A Public Twitter Dataset from Pakistan During Covid-19

Syed Tayyab ul Mazhar[1], Hasnain Khan[1], Uzma Afzal[1]*, Shazia Usmani[1] and Tariq Mahmood[2]

[1]Federal Urdu University of Arts, Science and Technology, Karachi, Pakistan; [2]Institute of Business Administration Karachi, Pakistan.

Abstract: In this article, we provide a collection of tweet data from five major cities (provisional capitals and federal capital) of Pakistan for the year 2021. A python's library "Tweepy" was used to collect the data. Tweets were filtered out using a set of related keywords. The dataset has five sub-datasets, i.e., one dataset for a city. Each sub-dataset contains the tweet data on a daily basis and can be analyzed separately. Pakistan observed two Covid waves in the year 2021 so these datasets also have tweets reflecting people's behavior from the 3rd to the 4th wave. These datasets can be used to compare the mental health and sentiments of the people from the different cities. These data sets also attract the attention of researchers from different fields such as data science, sentiment analysis, natural language processing, psychology and others.

Received: May 24, 2022; **Accepted:** October 29, 2023; **Published:** November 25, 2023
***Correspondence:** Uzma Afzal, Federal Urdu University of Arts, Science and Technology, Karachi, Pakistan; **Email:** uzma.afzal@fuuast.edu.pk
Citation: Mazhar, S.T., H. Khan, U. Afzal, S. Usmani and T. Mahmood. 2023. A public twitter dataset from Pakistan during Covid-19. *Journal of Engineering and Applied Sciences,* 42: 31-37.
DOI: https://dx.doi.org/10.17582/journal.jeas/42.1.31.37
Keywords: Covid-19, Pakistan, Twitter, Sentiment analysis, Corona, Dataset

Introduction

Covid-19 started spreading in 2019 from Wuhan, China. It is a respiratory infection that spreads from a person's cough, sneeze or breath. Covid-19 symptoms can be mild, moderate or serious. The first news of Covid-19 breakout in China was reported in December 2019 (Peirlinck *et al*., 2020). A state of emergency was issued by WHO due to the quick spread of the said virus (Soehrabi *et al*., 2020). Until Dec 14, 2021, more than 271M cases were reported worldwide, with a death toll of around 5M. Pakistan reported the first Covid case in Feb 2020, till November 2021, more than 1M cases were reported with 28K deaths (WorldoMeter, 2021). Pakistan has faced four Covid waves with different intensities. Government of Pakistan announced a complete lockdown in March 2020 followed by many small and smart lockdowns to control the rise in Covid cases (Dawn News, 2021). Luckily the recovery rate was very satisfactory and more than 98% of patients recovered.

In April 2020, the number of cases were risen and led the first Covid-19 wave (Shafi *et al*., 2020). Almost 6k positive cases were reported in a single day. After four months, the wave ended in July 2020. In October 2020, Covid-19 cases were risen again and led the second Covid-19 wave in Pakistan. This second wave lasted almost five months and ended in February 2021. The highest 3.7k cases were reported in a day during this wave. Third wave initiated in March 2021 and ended in June 2021. Six Covid-19 waves of different intensities hit Pakistan till the mid of the year 2022.

Covid-19 effected the feeling and attitude of the people, full and partial lockdowns in Pakistan hurt people's mental physical and emotional state. Social distancing increased the use of online gadgets and platforms during the pandemic. People shared their views and feeling on social media platforms such as Twitter and Facebook. This change in people's behavior attracted researchers to gauge the impact of pandemic on people's sentiments. To the best of our knowledge, we did not find any Covid related public dataset from Pakistan that store city-wise and wave-wise tweets.

To fill this research gap and record the views, sentiments and feeling of Pakistani people during the pandemic, we collected data of Covid related tweets from Pakistan. We designed an automated framework based on python scripts, deepnote and Tweepy to record and filter tweets (detailed in Section 2.2., i.e., Data Collection). We collected tweets from Karachi, Lahore, Quetta, Peshawar, and Islamabad for the year 2021. They are the representative cities to collect data (Sadiq and Qureshi, 2010).

The main contributions and novelty of the paper as follows:

- An automated framework is designed to record the tweets from different perspectives.
- A collection of Covid-19 related tweet data from five major cities is recorded for the year 2021.
- Datasets are pre-processed for two different dynamics, i.e., daily tweets collections for all selected cities and wave-wise tweets collection.

These datasets are very useful for many data applications. Researchers can use them to analyse the behavioural patterns of people, mental and health issues during the pandemic. Rest of the paper is organized as follows. Experimental design, materials and methods are presented in section 2. Section 3 discusses data description and analysis. Word cloud, importance of work and specification table are explained in section 4.

Experimental design, materials and methods

Data source: Covid-19 has caused many global issues with deep impact on a typical human behaviour. Social distancing and other safety measures to counter the spread of the virus isolated people and locked them in their homes. These safety measures hit the normal human communication channels such as meetings, parties and get-togethers. The use of the social media sites such as Twitter, Instagram, WhatsApp and Facebook increased drastically during the pandemic. People used these sites to share and express their views and feeling.

These sites have played a huge role during this pandemic. During the pandemic people of Pakistan have used twitter to exchange information regarding the pandemic and express their sentiments, thoughts and feelings (Batool *et al.*, 2021). Studies have shown that social media sites can be used to understand the response and behaviour of people during the pandemic (Manguri *et al.*, 2020; Nemes and Kiss, 2021). We selected Twitter for the research and data collection because Twitter data is publicly available with certain predefined limits. This dataset can be used to give an insight into the mental condition of different cities of Pakistan, moreover, it can also be used to analyse and compare the people behaviour during the pandemic.

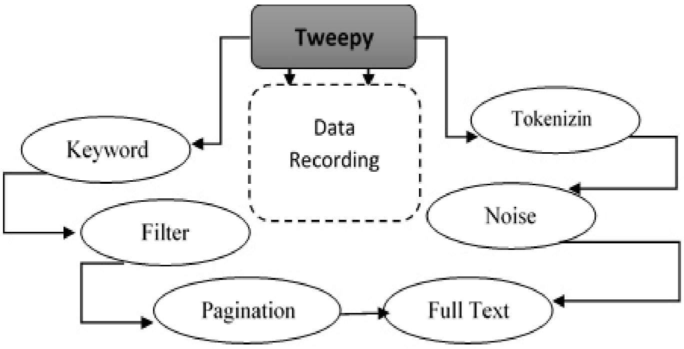

Figure 1: *An automated framework for data collection.*

Data collection

Figure 1 shows a design of an automated framework that was used for data collection. The raw data of tweets were collected on a daily basis for each specified city. A python script based on Tweepy library was run daily to download data (Roesslein, 2009; Bilal *et al.*, 2018). Daily execution of the script to collect the tweets manually was not an efficient way. Deepnote was used to automate the process and schedule the script execution daily at the specified time (Deep Note, 2021). Collection process was started in the end of March 2021 and lasted till October 2021. Data was stored in a separate file for each city. The hash tagged version of search terms were used to filter out the relevant tweets only. Corona virus, Covid19, lockdown were some of terms selected using Google

Trends (Sulyok *et al.*, 2021; Google Trends, 2021). We collected tweets only in English using 'en' as a language option in Tweepy. To specify a city, geo-coordinates of the city were mentioned. API returns a *Status* object with each fetched tweet, a separate class was created to filter the irrelevant information and store the tweet id, tag and tweet text (Roesslein, 2009; Arı, 2018). To avoid the truncation of tweet text get_status method of tweepy was used. Tweepy pagination was used to fetch the maximum number of tweets. tweepy.get_status method is used to fetch the full text of a tweet. Finally, regex was used to smooth the data and remove the noise such as punctuations, numbers and special characters. Each tweet is tokenized and split into words to perform stemming on those words. We used another python script to combine the wave-wise tweets, we ran this script for each city. For, 3rd wave data we combined tweets from march till June 2021 and for 4th wave we combined tweets from July 2021 till October 2021.

Data description and analysis

In this article, we present a collection of Covid related tweets from the five representative cities (Karachi, Lahore, Quetta, Peshawar, and Islamabad) of Pakistan for the year 2021 (Sadiq and Qureshi, 2010). We present data set from two different perspectives.

Table 1: *An overview of data sets.*

Dataset	Data files	Third wave tweets (≈)	Fourth wave tweets (≈)
Karachi	218	122633	27337
Lahore	218	113126	22401
Peshawar	218	11470	4010
Quetta	218	2330	1183
Islamabad	218	81028	16569

First, we have date-wise and time-wise data files for each city. In the data set, a folder with a city name contains 218 files, every file represents a date and contains different number of Covid related tweets with time stamp. Then, we have data collection on waves based, a folder named "third wave" contains five files; one file for a city with a collection of tweets from the starting of 3rd wave till the moderation, i.e., late March till June (Govt. of Pakistan, 2021). Another folder named "fourth wave" also contains five files; one file for a city with a collection of tweets from the starting of 4th wave till its moderation, i.e., July till October (Govt. of Pakistan, 2021). Table 1 presents an overview of the data sets.

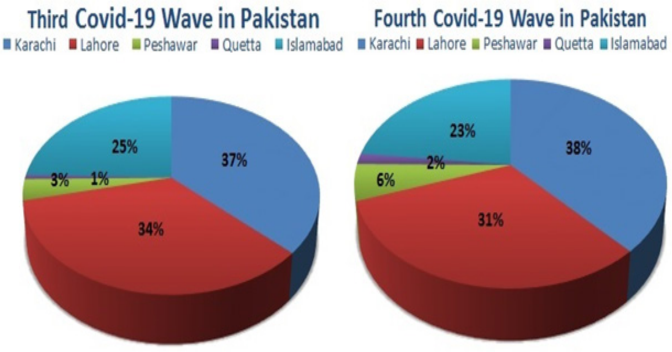

Figure 2: *Pie charts of cities representing Covid-19 Tweets count.*

Figure 2 shows the participation of cities in total number of collected tweets. During 3rd wave people of Karachi and Lahore used Twitter actively to express their thoughts and sentiments. Comparatively, Quetta and Peshawar posted less number of tweets. Similar to 3rd wave, Karachi and Lahore showed their opinion openly in 4th wave. A slight increase in Karachi, Quetta and Peshawar participations, and a slight decrease in Lahore and Islamabad participations are observed.

Figures 3 and 4 show the frequency trend of 3rd and 4th wave of Covid in 2021 for all five cities. Figure 3 shows that the tweets posting varies with the wave intensity. People shared a few hundred tweets in less intense days and more than thousands in peak days. The maximum number of tweets collected in the last 15 days of April 2021. Highest number of tweets was collected from Karachi and the least number of tweets was reported from Quetta. More than 7000 tweets were collected from Karachi in April and that was the highest frequency of tweets during the 3rd Covid wave.

Similar to 3rd wave, Figure 4 also shows that the tweets collection varies with the wave intensity. The maximum number of tweets collected in the first 15 days of July 2021. The highest number of tweets was collected from Karachi followed by Lahore during the 4th Covid wave. Trend shows that by the passage of time people lost their interest in Covid related discussion. Tweets were recorded in hundreds during the 4th wave while in 3rd wave, thousands of tweets were reported regularly.

There is a drastic difference in number of tweets collected in 3rd and 4th wave. Figure 5 shows the comparison. It shows that people started living with the fear factor of Covid-19 with the passage of time and lost their interest in posting Covid related tweets.

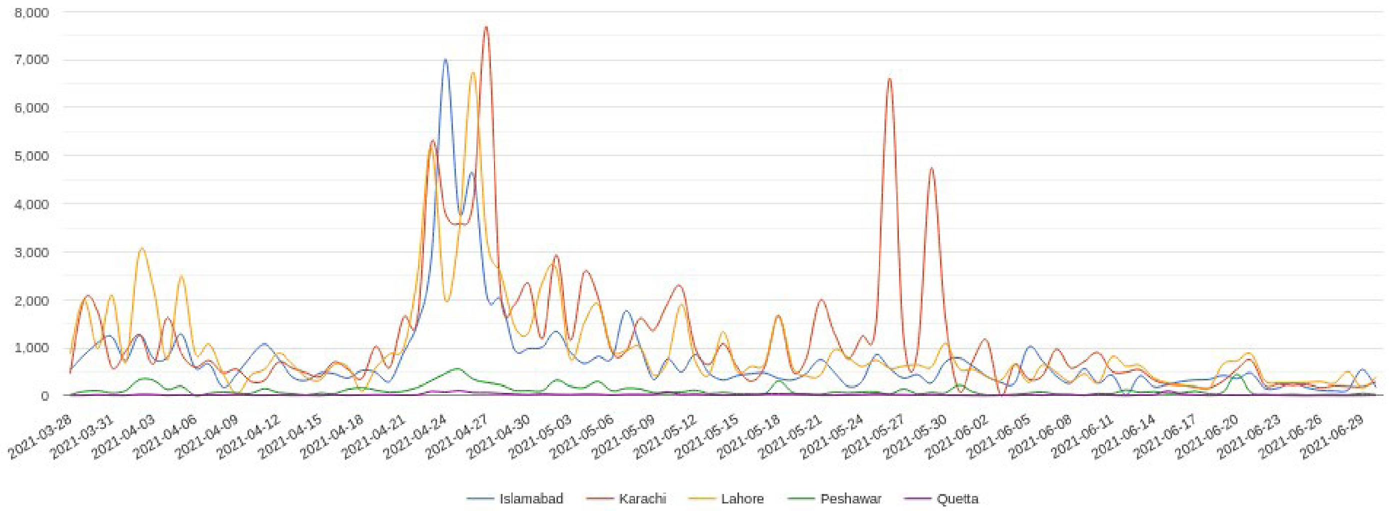

Figure 3: *Frequency trend of Covid related tweets- third wave.*

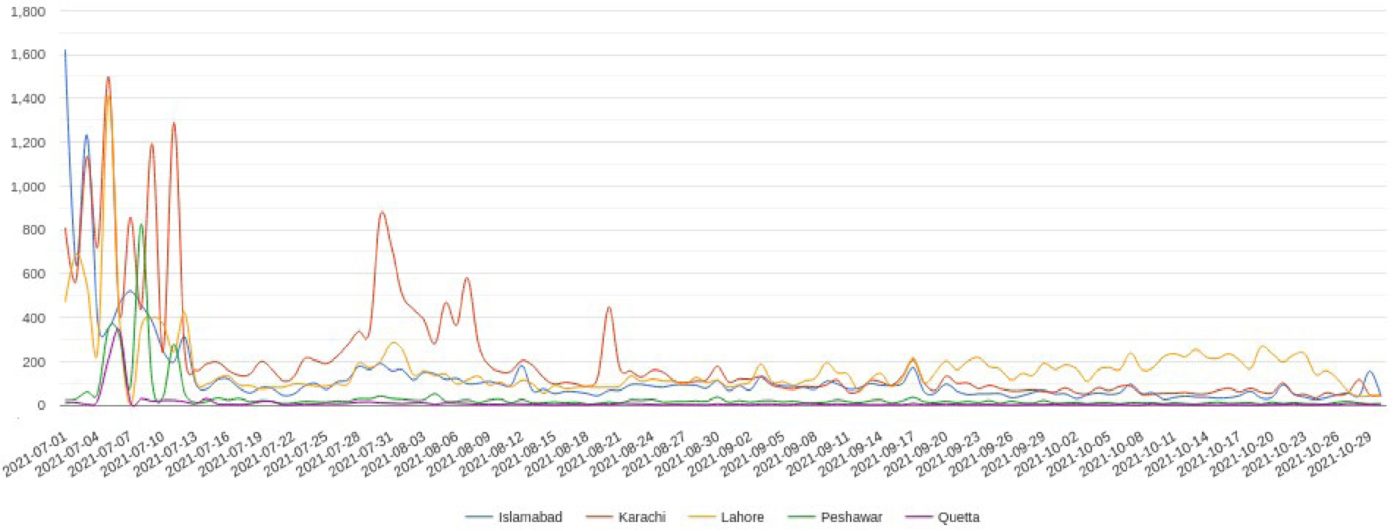

Figure 4: *Frequency trend of Covid related tweets- fourth wave.*

Table 2 describes the descriptive analysis such as minimum, mean and standard deviation. Mean measures the central tendency and standard deviation measures the dispersion. The mean describes a central point without giving information where it comes from. To give more insight data set dispersion is also recorded. 2/3 of the values fall within ± one SD of the mean. Similarly, 95% values are within ± 2SD. Rest of the table is self-explanatory so we are not expounding them in text.

Table 2: *Descriptive analysis of Covid-19 data in Pakistan.*

	Minimum	Maximum	Sum	Mean	Std. deviation
Third wave	2330	122,633	330,587	66,117.4	56,305.6
Fourth wave	1183	27,337	71,500	14,300.0	11,387.2
Total	3513	149,970	402,087	80,417.4	67,640.4

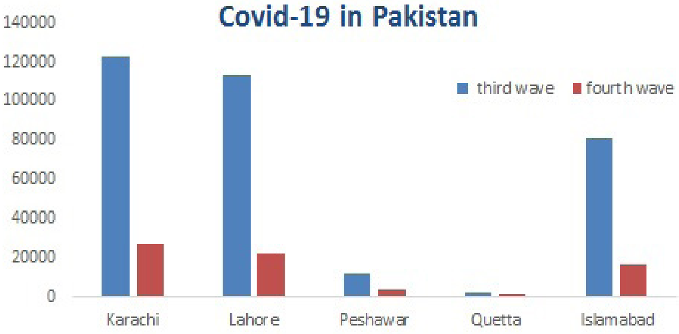

Figure 5: *A comparative analysis.*

Table 3 presents a comparative analysis of proposed work with the existing research works. Year column stores the specific year of tweets collection, Duration describes the total time frame of collection. Domain explains the generalization/specialization of tweets, Geo-Cordinates stores the location coordinates of tweets. DS (Dataset) and Exp (eriments) describe the

Table 3: *A comparative analysis research communities perspective.*

Paper	Year	Duration	Domain	Geo-co-ordinates	Pub-lic?	Tweets	DS?	Exp?
Banda *et al.*, 2021	2020-2021	≈ 18 Months	Covid related-Generic	Global	✓	≈ 1.12 Billion	✓	
Lamsal, 2021	2020	≈ 01 Month	Covid related-Generic	Global	✓	≈ 310 Million	✓	
Naseem *et al.*, 2021	2020	≈ 01 Month	Covid related-Generic	Global	✓	≈ 90 Thousand	✓	✓
Qazi *et al.*, 2020	2020	≈ 03 Months	Covid related–Social distancing, symptoms, supplies	Global		≈ 524 Million	✓	
Gupta *et al.*, 2021	2020-2022	≈ 28 Months	Covid related - vaccine	Global	✓	≈ 252 Million	✓	✓
Lopez, 2020	2020	≈ 02 Months	Covid related - policies	Global	✓	≈6.4 Million	✓	
Alqurashi *et al.*, 2020	2020	≈ 3.5 Months	Covid related – Arabic	Global	✓	≈3.9 Million	✓	
Haouari, 2020	2020-2021	≈ 12 Months	Covid related – Arabic	Global	✓	≈2.7 Million	✓	
Mubarak and Hassan, 2020	2020	≈ 02 Months	Covid related – Arabic	Global	✓	≈30 Million	✓	✓
This Paper	2021	≈ 08 Months	Covid related - generic	Defined	✓	≈0.4 Million	✓	

scope of the work. (Banda *et al.*, 2021; Lamsal, 2021; Naseem *et al.*, 2021) collected tweets which discussed the Covid related post without specifying a special filter. Qazi *et al.* (2020) presents a collection of tweets which discusses the social distancing, symptoms of the virus and shortage of the Covid supplies. Gupta *et al.* (2020), Lopez *et al.* (2020) recorded tweets from vaccine and policies perspective. Alqurashi *et al.* (2020), Haouari *et al.* (2020), Mubarak and Hassan (2020) collected and filtered only Arabic tweets related to Covid pandemic. Naseem *et al.* (2021), Gupta *et al.* (2020), Mubarak and Hassan (2020) present a collection of datasets along with the experimental results. Unlike other existing research work, we defined geo-cordinates during the tweets collection to store tweets originated from Pakistan. Moreover, we discuss the potential research domains to experiment with the given datasets.

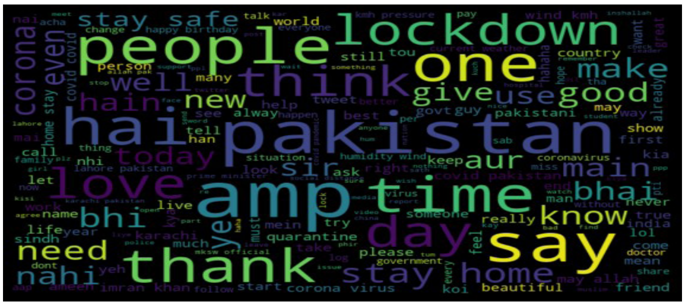

Figure 6: *Covid 3rd wave-wordcloud.*

Wordcloud

Wordcloud is a popular text analysis tool that provides a visualization of word frequency in the source text while giving more prominence to words that occur more often. Figures 6 and 7 show word cloud visualizations- of the most frequently encountered words in the dataset. It provides a general overview of the dominant terms related to the COVID-19 topic.

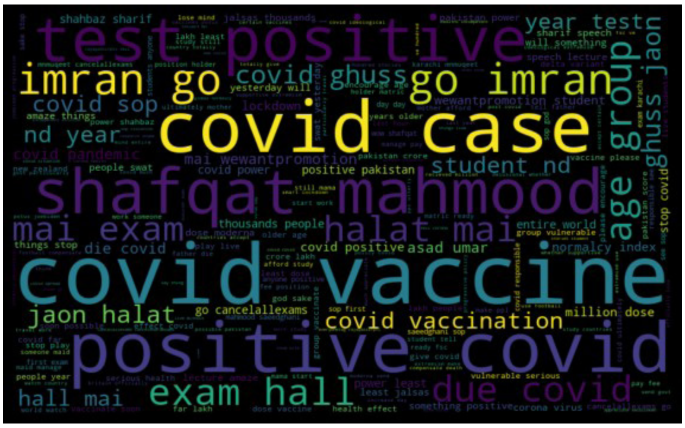

Figure 7: *Covid 4th wave-wordcloud.*

Importance of the work

- These datasets are useful and important because they are the very first public available twitter datasets from five main cities of Pakistan discussing the sentiments on corona virus.
- Researchers who want to analyse the sentiments of people during the pandemic can use these datasets. Also, researchers can use these datasets to conduct the sentiments comparisons on city basis. Government organizations can benefit from these datasets by developing appropriate policies to tackle negative sentiments of people.
- Researchers can also use these datasets to analyse the change in people psyche during different Covid waves.
- These datasets contain the Covid related tweets for year 2021 in Pakistan. In this year Pakistan observed two Covid waves. The data were collected from the starting of the third wave till the moderation of the fourth wave daily. Thus, these are time-based datasets.

Specification table

Type of data	Text (CSV-formatted)
Data collection	A python library (Tweepy library) was used to retrieve the tweets. Irrelevant tweets were filtered out using a set of keywords.
Description of data collection	Data were collected on a daily basis using the Tweepy library. A python script with language ('en') and location parameters of the cities was used. In order to collect the relevant tweets, a set of Covid related keywords were used. To remove the irrelevant information from a fetched tweet, a separate tweet class was created. Tweepy pagination was used to fetch the maximum possible tweets. Regex was used to remove the noise from tweets such as mentions and url. Pakistan observed two waves of corona in the year 2021. The collection process was started with the introduction of 3rd wave (March 2021) till the moderation of the 4th wave (October 2021). Another python code was used to combine the tweets for 3rd and 4th waves.
Data source origination	Tweets: Institution: Twitter.com Cites: Karachi, Lahore, Islamabad, Peshawar and Quetta Country: Pakistan
URL	Repository name: Mendeley DataData identification number: 10.17632/n98bxc4vjh.1 Direct URL to data: https://data.mendeley.com/datasets/n98bxc4vjh/1

Novelty Statement

An automated framework is proposed to record the tweets from different perspectives. Covid-19 related tweet data from Pakistan is recorded for the year 2021. Datasets are public and pre-processed for two different dynamics, i.e., daily tweets collections and wave-wise tweets collection.

Author's Contribution

Syed Tayyab Ul Mazhar-wrote the manuscript and recorded the data
Hasnain Khan -reviewed the manuscript and added automation, and recorded the data
Uzma Afzal-supervised the data collection and involved in data processing
Shazia Usmani-conceptualized and created the research setup and reviewed the final version of the paper
Tariq Mahmood-conceptualized and created the overall research setup and paper reviewed

Ethics statements

Data was collected and distributed under Twitter's developer policy. It was anonymised, and no personal information of user's was exposed to cater the copyright infringement and privacy issue. Therefore, the authors ensure data is fully compliance with Twitter policies (Twitter, 2021).

Conflict of interest

The authors have declared no conflict of interest.

References

Alqurashi, S., A. Alhindi and E. Alanazi. 2020. Large arabic twitter dataset on covid-19. arXiv preprint arXiv: 2004.04315.

Arı, E., 2018. Trangling weratedogs twitter data to create interesting and trustworthy explosatory/predictive anaylses and visulation using different machine learning algorithms.

Banda, J.M., R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding and G. Chowell. 2021. A large-scale COVID-19 Twitter chatter dataset for open scientific research. An international collaboration. Epidemiologia, 2(3): 315-324. https://doi.org/10.3390/epidemiologia2030024

Batool, S.H., W. Ahmed, K. Mahmood and A. Sharif. 2021. Social network analysis of Twitter data from Pakistan during COVID-19. Inf. Discovery Delivery, 50(4): 353-364. https://doi.org/10.1108/IDD-03-2021-0022

Bilal, M., S. Asif, S. Yousuf and U. Afzal. 2018. 2018. Pakistan general election: understanding the predictive power of social media. In 2018 12th international conference on mathematics, actuarial science, computer science and statistics (MACS). pp. 1-6. IEEE. https://doi.org/10.1109/MACS.2018.8628445

Dawn News, 2021. Corona virus, Available at: https://www.dawn.com/trends/coronavirus (Accessed Nov 2021).

Deep Note, 2021. Scheduling- deep note docs, Available at: https://docs.deepnote.com/features/scheduling (Accessed Nov 2021).

Google Trends, 2021. Available at: https://trends.

google.com/trends/?geo=PK (Accessed Nov 2021).

Govt. of Pakistan, 2021. Corona virus Pakistan cases details Available at: http://covid.gov.pk/stats/pakistan (Accessed Nov 2021).

Gupta, R.K., A. Vishwanath and Y. Yang. 2020. COVID-19 Twitter dataset with latent topics, sentiments and emotions attributes. arXiv preprint arXiv: 2007.06954.

Haouari, F., M. Hasanain, R. Suwaileh and T. Elsayed. 2020. The first Arabic COVID-19 Twitter dataset with propagation networks. arXiv preprint arXiv: 2004.05861.

Lamsal, R., 2021. Design and analysis of a large-scale COVID-19 tweets dataset. Appl. Intel., 51: 2790-2804. https://doi.org/10.1007/s10489-020-02029-z

Lopez, C.E., M. Vasu and C. Gallemore. 2020. Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset. arXiv preprint arXiv: 2003.10359.

Manguri, K.H., R.N. Ramadhan and P.R.M. Amin. 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. Kurdistan J. Appl. Res., pp. 54-65. https://doi.org/10.24017/covid.8

Mubarak, H. and S. Hassan. 2020. Arcorona: Analyzing arabic tweets in the early days of coronavirus (covid-19) pandemic. arXiv preprint arXiv: 2012.01462.

Naseem, U., I. Razzak, M. Khushi, P.W. Eklund and J. Kim. 2021. COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE Trans. Comput. Soc. Syst., 8(4): 1003-1015. https://doi.org/10.1109/TCSS.2021.3051189

Nemes, L. and A. Kiss. 2021. Social media sentiment analysis based on COVID-19. J. Inf. Telecommun., 5(1): 1-15. https://doi.org/10.1080/24751839.2020.1790793

Peirlinck, M., K. Linka, F.S. Costabal and E. Kuhl. 2020. Outbreak dynamics of COVID-19 in China and the United States. Biomech. Model. Mechanobiol., 19(6): 2179-2193. https://doi.org/10.1007/s10237-020-01332-5

Qazi, U., M. Imran and F. Ofli. 2020. GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. SIGSPATIAL Special, 12(1):6-15. https://doi.org/10.1145/3404820.3404823

Roesslein, J., 2009. Tweepy documentation, Available At: http://tweepy.readthedocs.io/en/v3 (Accessed Nov 2021).

Sadiq, N. and M.S. Qureshi. 2010. Climatic variability and linear trend models for the five major cities of Pakistan. J. Geogr. Geol., 2(1): 83-92. https://doi.org/10.5539/jgg.v2n1p83

Shafi, M., Liu, J. and Ren, W., 2020. Impact of COVID-19 pandemic on micro, small, and medium-sized Enterprises operating in Pakistan. Res. Globaliz., 2: 100018.

Sohrabi, C., Z. Alsafi, N. O'neill, M. Khan, A. Kerwan, A. Al-Jabir and R. Agha. 2020. World health organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Int. J. Surg., 76: 71-76. https://doi.org/10.1016/j.ijsu.2020.02.034

Sulyok, M., T. Ferenci and M. Walker. 2021. Google trends data and COVID-19 in Europe: Correlations and model enhancement are European wide. Transb. Emerg. Dis., 68(4): 2610-2615. https://doi.org/10.1111/tbed.13887

Twitter Terms of Service, 2021. Available At: https://twitter.com/en/tos (Accessed Nov 2021).

WorldoMeter, 2021. Covid Live-Corono Virus Statistics. Available At: https://www.worldometers.info/coronavirus/ (Accessed Nov 2021).