A DATASET OF LOCATION-BASED TWEETS IN AFRICA ON COVID-19 OUTBREAK

Emeka OGBUJU[68]

Oluwatobi BANJO[69]

Ezekiel Adebayo OGUNDEPO[70]

Sakinat FOLORUNSO[71]

Francisca OLADIPO[72]

Abstract
Twitter has proven to be a ready venue for democratizing opinion data even during the COVID-19 pandemic. During the protracted periods of the resultant lockdown, access to the internet allowed citizens of various nations and government agencies to express their opinions online using their Twitter handles. In this data article, a collection of 619,203 tweets posts were provided on COVID-19 in some selected countries in Africa. This data was collected over 180 days, from February 14, 2020, to August 14, 2020. This dataset can attract researchers' attention related to different fields of knowledge such as data science, natural language processing, social science, informatics, tourism, and infodemiology

Keywords: tweets, COVID-19, coronavirus, lockdown, pandemic, africa

JEL Classification: Z00

1. Introduction

Data-driven research and development is crucial to the COVID-19 Pandemic. Opinions mined from the internet especially from social media are very useful to decision makers in a crisis period like the pandemic. This data article provides opinions from five (5) selected African countries about the pandemic to assist data scientists and epidemiologists in providing valuable incident response to the situation. The Specifications Table gives a summary of the dataset, and the rest of the article describes its usefulness.

[68] Senior Lecturer, CS, Federal University Lokoja, Nigeria, emeka.ogbuju@fulokoja.edu.ng
[69] Lecturer, CS, Olabisi Onabanjo University, Nigeria, banjo.tobi@oouagoiwoye.edu.ng
[70] Data Scientist, Data Science Nigeria, gbganalyst@gmail.com
[71] Senior Lecturer, CS, Olabisi Onabanjo University, Nigeria, sakinat.folorunso@oouagoiwoye.edu.ng
[72] Professor and Chair, CS, Federal University Lokoja, Nigeria, francisca.oladipo@fulokoja.edu.ng

Specifications Table

Subject	Infectious Diseases, Epidemiology, Social Science, Health Informatics, Computer Science
Specific subject area	Social Media
Type of data	Tweets
How data were acquired	Querying Twitter REST API using the "GetOldTweets" Python package [1]
Data format	Analysed and Filtered
Parameters for data collection	Tweets matching keywords COVID, COVID19, COVID-19, Corona, Coronavirus, Lockdown, Pandemic
Description of data collection	Tweets matching COVID-19 and Lockdown between February 14, 2020 to August 14, 2020 from Nigeria, South Africa, Algeria, Egypt, and Sudan
Data source location	Nigeria, South Africa, Algeria, Egypt, and Sudan
Data accessibility	Repository name: Mendeley Data Repository: http://dx.doi.org/10.17632/c8x5tpvzmk.3 Project URL: http://bit.ly/COVID-19-tweets-github-repo

2. Value of the Data

● The data collection covers the period of the initial outbreak of the pandemic and the early stages of the lockdown in the selected countries. It can serve as a gauge for a comparative study on the pandemic's perception and the lockdown measures.

● It is useful for qualitative analysis of social media contents to extract personal and corporate opinions on the pandemic.

● The individual tweets may be a valuable resource for analyzing different users' posting patterns; it can reveal the dynamics of each country's citizens on their perceptions about the pandemic and well as the lockdown.

● The tweets may be useful for computational tasks that may provide domain summary and/or thematic analysis that determines main topics in each African country about the COVID-19 pandemic.

● The data is relevant for sentiment analysis tasks to determine Africans' opinions and emotions on the pandemic. Please abide exactly by the above regulations.

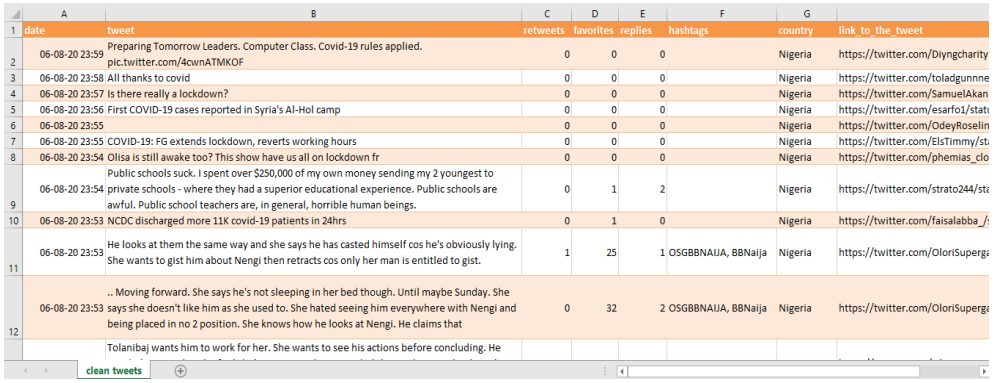
3. Acquirements of the Data

According to a global report on June 2020 in [2], 'Africa has recorded fewer than 6,000 deaths, according to an AFP tally, but just five countries account for 70% of these: South Africa, Algeria, Nigeria, Egypt, and Sudan'. This justifies our choice to collect the pandemic datasets from these countries. The datasets were collected over a period of six (6) months from February 14, 2020, to August 14, 2020, using seven (7) keywords/hashtags as follows: COVID, COVID19, COVID-19, Corona, Coronavirus, Lockdown, and Pandemic. This is not unconnected with the fact that Africa's first case of COVID-19 was confirmed in Egypt on February 4, 2020, and by August 2020, Africa's COVID-19 confirmed cases had increased to over one million [3].

Table 1: Description of Variables

Variable Name	**Description**
date	Date and time (in UTC) in which the tweet was posted
tweet	The text of the post
retweets	The number of times the tweets was reposted by other users
favorites	The number of times other users liked the post
replies	The number of users who commented on the tweet
hashtags	Specific keywords used within the tweet
country	The nation from which the tweets was extracted
link_to_the_tweet	A live URL of the tweet

The tweets are stored in two separated Comma-Separated Values (CSV) files, Figure 1 shown below is the glimpse of the clean tweets.



	A	B	C	D	E	F	G	
1	date	tweet	retweets	favorites	replies	hashtags	country	link_to_the_tweet
2	06-08-20 23:59	Preparing Tomorrow Leaders. Computer Class. Covid-19 rules applied. pic.twitter.com/4cwnATMKOF	0	0	0		Nigeria	https://twitter.com/Diyngcharity
3	06-08-20 23:58	All thanks to covid	0	0	0		Nigeria	https://twitter.com/toladgunnne
4	06-08-20 23:57	Is there really a lockdown?	0	0	0		Nigeria	https://twitter.com/SamuelAkan
5	06-08-20 23:56	First COVID-19 cases reported in Syria's Al-Hol camp	0	0	0		Nigeria	https://twitter.com/esarfo1/stat
6	06-08-20 23:55		0	0	0		Nigeria	https://twitter.com/OdeyRoselin
7	06-08-20 23:55	COVID-19: FG extends lockdown, reverts working hours	0	0	0		Nigeria	https://twitter.com/ElsTimmy/st
8	06-08-20 23:54	Olisa is still awake too? This show have us all on lockdown fr	0	0	0		Nigeria	https://twitter.com/phemias_clo
9	06-08-20 23:54	Public schools suck. I spent over $250,000 of my own money sending my 2 youngest to private schools - where they had a superior educational experience. Public schools are awful. Public school teachers are, in general, horrible human beings.	0	1	2		Nigeria	https://twitter.com/strato244/st
10	06-08-20 23:53	NCDC discharged more 11K covid-19 patients in 24hrs	0	1	0		Nigeria	https://twitter.com/faisalabba_/
11	06-08-20 23:53	He looks at them the same way and she says he has casted himself cos he's obviously lying. She wants to gist him about Nengi then retracts cos only her man is entitled to gist.	1	25	1	OSGBBNAIJA, BBNaija	Nigeria	https://twitter.com/OloriSuperg
12	06-08-20 23:53	.. Moving forward. She says he's not sleeping in her bed though. Until maybe Sunday. She says she doesn't like him as she used to. She hated seeing him everywhere with Nengi and being placed in no 2 position. She knows how he looks at Nengi. He claims that	0	32	2	OSGBBNAIJA, BBNaija	Nigeria	https://twitter.com/OloriSuperg
		Tolanibaj wants him to work for her. She wants to see his actions before concluding. He						

clean tweets

Figure 1. A glimpse of the dataset

raw tweets: This CSV file comprises raw tweets with some duplicates.

clean tweets: This CSV file comprises clean tweets with no duplicate

4. Experimental Design, Materials, and Methods

The raw data of 826,412 Twitter posts were downloaded from Twitter REST API search using the "GetOldTweets" Python package, which was designed to bypass some of the limitations of the official Twitter API [1]. The collection duration is between February 14, 2020, to August 14, 2020, i.e., 6 months. We search for keywords which include COVID, Coronavirus, Lockdown, Pandemic, etc. The overall flow diagram of the work is presented in Figure 2.

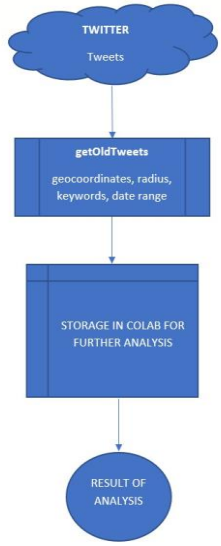


Figure 2. Overall Flow Diagram of the Work

It is worth mentioning that the procedure to search tweets was not a one-off exercise, network breaks and server timing out led to us having to run the process several times, creating 66 files, which summed up to 826,412 records. We eliminated duplicated ones, and the final number of retained tweets was 619,203 unique tweets. All the collected data are multilingual since we did not enforce any form of language restriction. Each location's data were collected by querying each location's center and providing its geo coordinates with the radius being set to 100km for each of the geo-coordinates queried. There are countries where we had to query twice based on the few tweets retrieved at the initial trial. Table 2 provides details of the geo-coordinates queried for each country and the number of tweets retrieved.

Table 2. Number of data retrieved from each country

Country	Geo Coordinates (Latitude, Longitude)	Total number of tweets retrieved	Distinct tweets by user ID
Nigeria	9.064331,7.489297	361,368	231,719
South Africa	-29.116395, 26.215496	170,561	169,509
Algeria	34.671359, 3.254037	119,889	106,105
	27.194077, 2.481557		
Egypt	26.547748, 31.699264	109,396	101,940
Sudan	15.644554, 32.477731	65,198	9,930
	15.624521, 32.58819		

5. Application and Limitation of the Data

The frequency of daily tweets retrieved per country is shown in Figure 3. It can be seen that fewer daily tweets about COVID-19 pandemic were received in all the countries in March while the number increased in April but decreased in May through to July. However, more tweets were recorded in August in Nigeria and South Africa, but Sudan maintains very low daily tweets throughout the period.

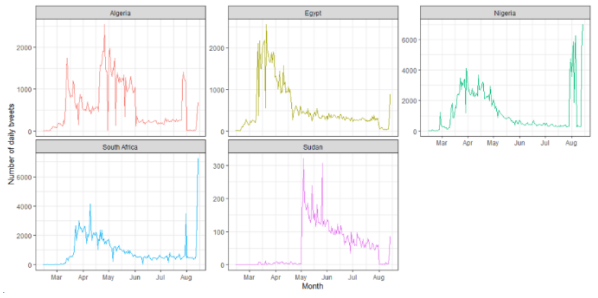


Figure 3. Frequency of tweets per country

Figure 4 shows that COVID-19 and lockdown were the most frequently used words tweeted across the countries during COVID-19 pandemic.

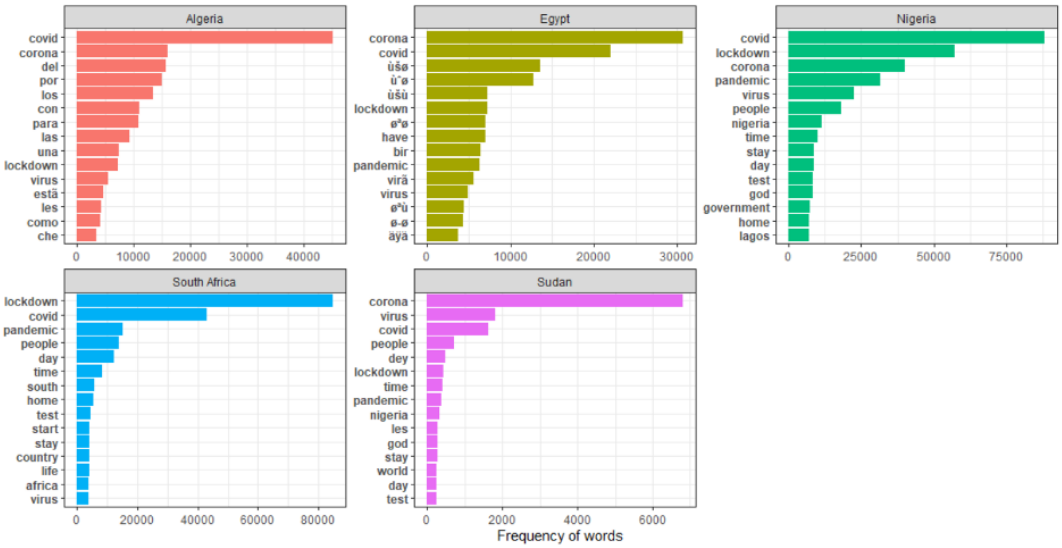


Figure 4. Most common words used during COVID-19 by countries

The sentiment expressed by residents of the five (5) countries were analyzed to uncover the active feedback of the people according to [4]. The sentiment analysis was performed using Valence Aware Dictionary for sEntiment Reasoning (VADER), a lexicon and rule-based sentiment analysis that is designed to detect sentiments expressed on social media [5], and the methodology of Elbagir and Yang [6] which is a more dynamic sentiment classification above the regular binary oriented classification of sentiments scores expressed into 5 groups of Neutral, Positive, Highly Positive, Negative and Highly Negative based on the sentiment intensity.

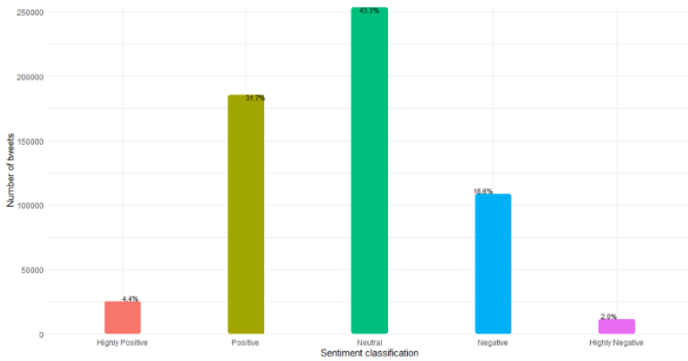


Figure 5: Sentiment classification of the overall tweets

The sentiment analysis of the entire tweets across the six countries is shown in Figure 5, and its distribution is as follows: Highly Positive (26,007; 4.4%), Positive (185,799; 31.7%), Neutral (254,033; 43.3%), Negative (108,939; 18.6%), and Highly Negative (11,954; 2%). The distribution of sentiment across the tweets per county is shown in Figure 6. In Nigerian based tweets, the most dominant sentiment is Positive (85,472; 38.1%). In South Africa, the most dominant sentiment is Neutral (58,645; 36.2%). In Algeria, the most dominant sentiment is Neutral (67,673; 67.9%). Egypt also had its most prevalent sentiment Neutral (51,664, 56.7%) while Sudan equally had Neutral as its most predominant sentiment (3,751; 39.3%).

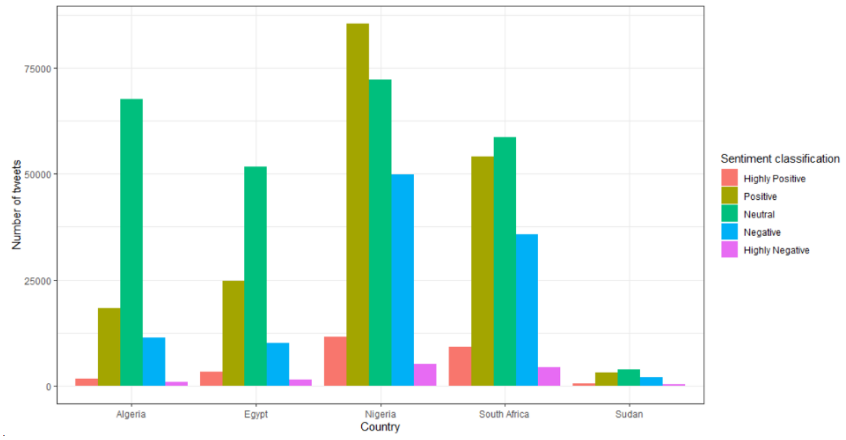


Figure 6: Sentiment classification of tweets by country

Figure 7 also reveals a similar pattern across the entire dataset regarding the dominant sentiment expressed each month being Neutral. Since Africa has 11 cases within February to March 4, 2020 [7], we have fewer tweets and sentiment about COVID-19 in February.

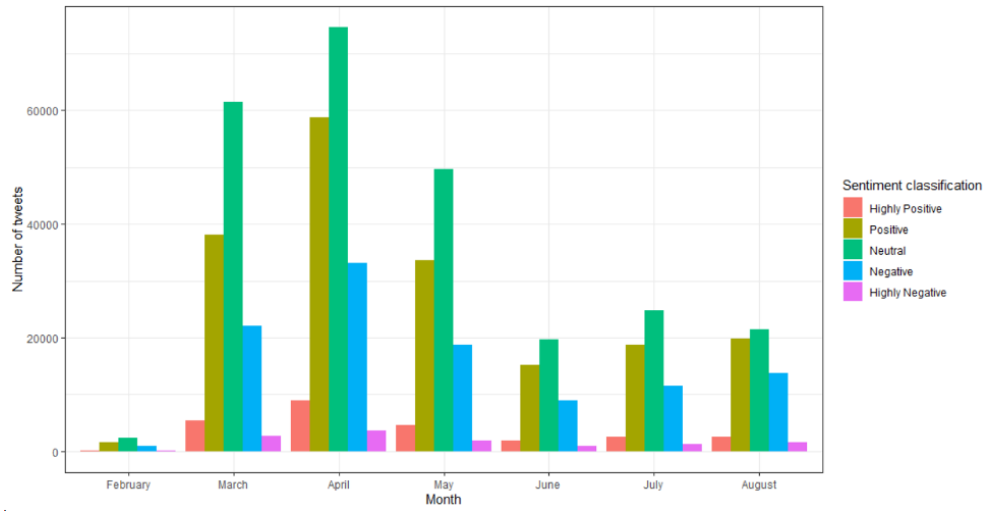


Figure 7: Sentiment classification of tweets by month

6. Project Implementation

Python and R programming languages were used for the analysis of the text data. Packages such as dplyr and ggplot2 for data analysis and visualization [8, 9] and tidytext for text mining [10] were used in R while Pandas and vaderSentiment were used in Python for data analysis and sentiment analysis respectively. The project scripts have been uploaded into a GitHub repository which can be accessed via http://bit.ly/COVID-19-tweets-github-repo.

Acknowledgements

Our gratitude goes to the Virus Outbreak Data Network (VODAN Africa & Asia) for the training and workshops in creating the science of FAIR data in Africa. We also sincerely appreciate its Nigerian Chapters, Federal University Lokoja, Data Science Nigeria (DSN), and Olabisi Onabanjo University for providing data stewardship.

References

[1]. Jefferson Henrique. GetOldTweets. Retrieved from https://github.com/Jefferson-Henrique/GetOldTweets-python on August 13, 2020

[2]. Global report: WHO warns of accelerating COVID-19 infections in Africa.

Retrieved from https://www.theguardian.com/ on June 12, 2020

[3]. BBC News. Coronavirus in Africa tracker. Retrieved from https://www.bbc.co.uk/news/ on 30 August, 2020

[4]. Salinca, A. (2015). Business reviews classification using sentiment analysis. 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 247–250. https://doi.org/10.1109/SYNASC.2015.46

[5]. Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. 8th International AAAI Conference on Weblogs and Social Media, 216–225.

[6]. Elbagir, S., & Yang, J. (2019). Twitter sentiment analysis using natural language toolkit and Vader sentiment. Proceedings of the International MultiConference of Engineers and Computer Scientists, IMECS 2019, 12–16.

[7]. COVID-19 Situation update for the WHO African Region. (2020, March 4). ReliefWeb. https://reliefweb.int/report/nigeria/covid-19-situation-update-who-african-region-external-situation-report-1-4-march-2020

[8]. H. Wickham, F. Romain, H. Lionel, M. Kirill, dplyr: A Grammar of Data Manipulation, 2020 R package version 1.0.0 https://CRAN.R-project.org/package=dplyr

[9]. H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016

[10]. Silge, J., & Robinson, D. (2016). tidytext: Text mining and analysis using tidy data principles in R. Journal of Open-Source Software, 1(3), 37.