Annals of Data Science (2022) 9(1):175–186
https://doi.org/10.1007/s40745-021-00358-5

Dictionary Based Global Twitter Sentiment Analysis of Coronavirus (COVID-19) Effects and Response

Elphas Okango[1] · **Henry Mwambi**[1]

Received: 6 December 2020 / Revised: 24 August 2021 / Accepted: 9 October 2021 /
Published online: 20 January 2022
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract
In December 2019, a new pandemic called the coronavirus began ravaging the world. By May 2020, the pandemic had caused great loss of lives and disrupted the way of lives in more ways than one. The nature of the disease saw several strategies to curb its spread rolled out. These strategies included closing of businesses and borders, restriction of movements and working from home, mask mandate among others. With these measures and the effects, many individuals have taken to the social media to express their frustrations, opinions and how the pandemic is affecting them. This study employs dictionary based method for sentiment polarization from tweets related to coronavirus posted on Twitter. We also examine the co-occurrence of words to gain insights on the aspects affecting the masses. The results showed that mental health issues, lack of supplies were some of the direct effects of the pandemic. It was also clear that the COVID-19 prevention guidelines were well understood by those who tweeted. The results from this study may help governments combat the consequences of COVID-19 like mental health issues, lack of supplies e.g. food and also gauge the effectiveness or the reach of their guidelines.

Keywords Coronavirus · Lexicon · Sentiment · Word co-occurrence · Valance shifters

1 Introduction

In recent years, data science has emerged as a new and important discipline which can be viewed as an amalgamation of traditional disciplines like statistics, data mining and distributed systems [1]. Data driven decision making has become ubiquitous in almost all aspects of the society. With the Internet of Things, huge volumes of wide

✉ Elphas Okango
kangphas@gmail.com

[1] School of Mathematics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa

variety of data are generated at high velocity. Real time decision making is central to the Internet of Things [2].

In the world where a lot is bound to happen, what the masses think or feel about these happenings is a concern for governments, businesses and even individuals. Governments would want to know how their policies, interventions etc. are received or perceived by the masses, politicians would want to know if they have a favorable rating, and how their policies are received and implemented while businesses would want to understand the reputation of their brands. Social media presents a great promise for achieving this by analyzing social media posts, product reviews, customer feedback etc. The advent of social media has made available a platform where individuals can freely express their opinions, feelings or judgments. Careful data mining techniques can help unravel valuable information and draw insights which may be hidden in these expressions.

On 31st December 2019, a cluster of pneumonia cases of unknown etiology was reported in Wuhan, Hubei Province, China [3, 4]. About a week later on 9th January 2020, the Chinese center for disease control (CDC) reported a novel coronavirus as the causative agent of this outbreak, corona virus disease 2019 (COVID-19). Covid-19 is spread from person to person through respiratory droplets when an infected person sneezes, coughs or talks [5]. One is also able to contract the COVID-19 by touching a surface or object that has the virus on it and then touching his/her nose, mouth or eyes.

As of April 29th 2020, there were 2,995,758 confirmed cases, 204, 987 deaths in 213 countries, areas or territories [6]. The nature of the disease (highly transmissible even when an infected individual is still asymptomatic) has seen many governments put in place a raft of measures in bid to curb the spread or the disease. Some of these measures include total and partial lockdowns which have seen businesses closed, curfews, advocacy for staying at home, social distancing, wearing of a cloth face covering nose and mouth in public places, regular washing of hands for at least 20 s or by using alcohol based hand sanitizers that contains at least 60% alcohol, quarantine for infected individuals [5].

The measures put in place as a result of the COVID-19 pandemic has affected the way people do things and it would be of interest to know and understand the feelings, opinions or judgment of the masses on various issues. Several studies have used varied approaches and datasets to try and explain the COVID-19 dynamics. Kumar [7] employed cluster analysis in monitoring COVID-19 infections in India. The approach identified areas/clusters that needed more medical facilities (ventilators, testing kits, masks etc.) and those that needed optimization of monitoring techniques (screening, lockdowns, closedowns, curfews etc.). Khakharia et al. [8] used machine learning techniques to predict the outbreak of COVID-19 for 10 densely populated countries. In particular, they compared the performance of 9 machine learning models in predicting the outbreak. The highest prediction accuracy was achieved for Ethiopia using the Autoregressive Moving Average Model. Social contact based analysis has also been employed to study the underlying disease transmission patterns. Liu et al. [9] using this approach showed that the age-groups involving relatively intensive contacts in households and public/communities were dispersedly distributed explaining why the transmission of COVID-19 in the early stage mainly took place in public places

and families in Wuhan. Other data mining techniques that can be employed to study the dynamics of COVID-19 are available in [10, 11]

Sentiment analysis or opinion mining can be defined as the process of identifying and extracting the subjective information that underlies a text [12]. This information can either be an opinion, a feeling about a particular topic or subject matter or a judgment. Sentiment analysis is becoming a field of interest that cannot be ignored. Nguyen et al. [13] employed sentiment analysis on social media to predict stock movement. Their method of incorporating social media data achieved 2.07% better performance than the model using historical prices only. Vincenza et al. demonstrated that Twitter data and sentiment analysis can be used to study disease dynamics [14]. Twitter is a micro blogging and social networking service on which users post and interact with messages known as tweets [15]. Twitter's 321 million active users provide a rich source of data from the tweets they post. In this study we seek to mine opinions and sentiments on the COVID-19 pandemic from Twitter users.

2 Methodology

2.1 Data

This study seeks to provide a framework for real time social media data analysis for actionable intelligence. Data used in this study were tweets relating to the COVID-19 pandemic and these were streamed live from Twitter on 14th -15th April 2020 (from 16:43:09 on 14th to 23:50:53 on 15th) and from 18:24:25 on 17th April 2020 to 16:41:16 the following day using streamR package [16]. The time periods were East African time. In particular only tweets bearing words such as corona, covid-19, sanitizer, virus, lockdown, quarantine, social distance were of interest and thus streamed.

The streaming was broken into two–three hours intervals with about 2 s break between each interval in order to obtain smaller sizes of streamed tweets. The tweet files were then parsed and compiled into a single excel file. We obtained more than 20 million tweets out of which a 91,784 geo-tagged tweets from all over the world were derived.

2.2 Exploratory Data Analysis

Figure 1 was obtained using data from the John Hopkins University and functions from tidycovid19 package [17]. The United States of America, some parts of Europe, Asia and Russia had the highest number of active cases per 100,000 inhabitants.

2.3 Tweets Location

Figure 2 shows COVID-19 related tweets. It is clear there was concentration of tweets around Europe, South America in particular Brazil, Asia in particular India and in

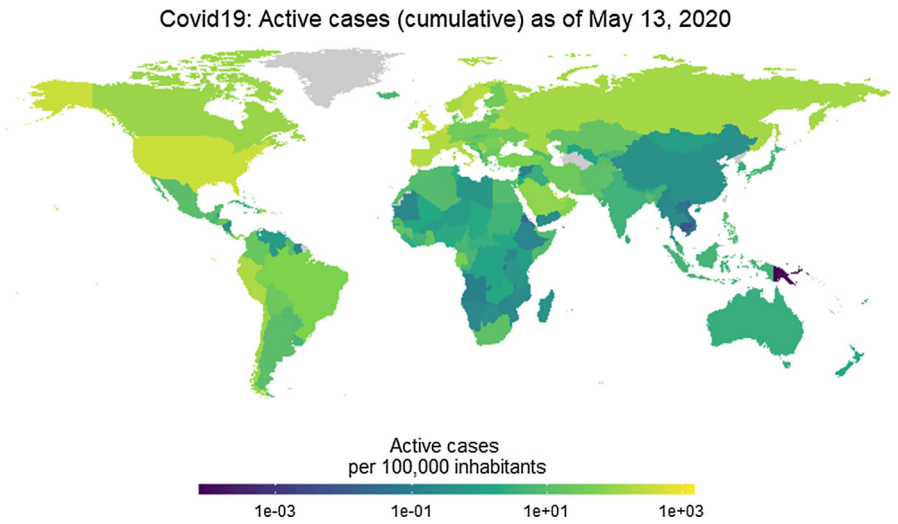

Fig. 1 COVID-19 Cumulative active cases

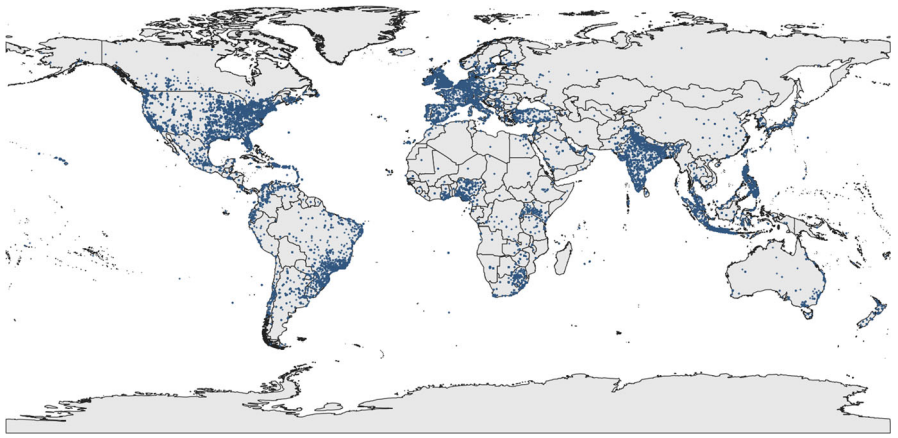

Fig. 2 Location of COVID-19 related tweets

Western and Southern Africa. Countries with high cases of COVID-19 posted more tweets.

Figure 3 displays tweets location by language. English language was the most dominant language in our data set denoted by red dots. In particular there were 63,056 English tweets.

Fig. 3 Tweets by language

2.4 Methods

There exists several methods of sentiment analysis. Sentiment analysis can be done on three levels namely: document-level, sentence level and aspect-level [18]. Document-level sentiment analysis considers the whole document as a basic information unit (talking about one topic) and classifies it as expressing a negative, positive or neutral sentiment. Sentence level sentiment analysis classifies sentiment expressed in each sentence [18]. Sentiment classification techniques can be divided into machine learning approach, lexicon based approach and a hybrid approach that combines machine learning and lexicon approaches [19]. Machine learning approach relies on the machine learning algorithms like the naïve Bayes, support vector machines, neural networks among others together with linguistic features. In lexicon based approach, a collection of known and precompiled sentiment terms known as sentiment lexicon is used. This approach can be divided into dictionary based approach and corpus based approach which employs statistical or semantic methods to find sentiment polarity [18].

In communication, one listens out to an entire sentence and derive meaning that is greater than the sum of individual words. Calculating polarity or sentiment by matching words with those in the dictionary of words classified as positive, negative or neutral leaves out useful information. In many cases valance shifters (negators, amplifiers/intensifiers, de-amplifiers/downtoners, adversative conjunctions) are not taken into account. Negators flip the sign of a polarized word e.g. “I do not like”, An amplifier (intensifier) increases the impact of a polarized word (e.g., “I **r eally** like it.”). de-amplifier (downtoner) reduces the impact of a polarized word (e.g., “I **hardly** like it.”). An adversative conjunction overrules the previous clause containing a polarized word (e.g., “I like it **but** it’s not worth it.”) [20].

Valence shifters affect polarized words and if they do occur frequently, a single dictionary look up may not be the best approach to model the sentiments appropriately.

The entire sentence may be reversed or overruled in the case of negators and adversative conjunctions [20].

From Tinker's methodology [20], tweet S_j is a sentence composed of words $W_1, W_2, \ldots, W_n$. Each tweet is broken down into an ordered bag of words. With the exception of pause punctuations (commas, colons, semicolons) which are considered words within a sentence, other punctuations are removed. The words are indexed as W_{ij} indicating the jth word in the ith tweet. The words in each tweet are searched and compared to a dictionary of polarized words, with positive words W_{ij}^{+} assigned $+1$ and negative ones W_{ij}^{-1} -1 or other positive and negative weighting depending on the sentiment dictionary used.

Denote polarized words by $\boldsymbol{pw}$, these will form a polar cluster c_{ijl} which is a subset of a tweet i.e.$c_{ikl} \subset s_{ij}$. The polarized cluster of words c_{ijl} is pulled from around the polarized word $\boldsymbol{pw}$ and defaults to 4 words before and two words after $\boldsymbol{pw}$ to be considered as valance shifters. The cluster is represented as$c_{ijl} = pw_{ij} - nb, \ldots, pw_{ij}, \ldots, pw_{ij} - na$. Here nbandna are parameters n-before and n-after set by the user. The words c_{ijl} are labeled neutralw_{ij}^{0}, negatorw_{ij}^{n}, amplifier/intensifier w_{ij}^{a} or deamplifier of downtonerw_{ij}^{d}. Neutral words only contribute to the number of words in the equation. Each polarized word is then weighed by some function and the number of valence shifters surrounding the positive or negative word. Pause locations denoted by cw (i.e. punctuations that denote a pause including commas, colons and semicolons) are indexed and incorporated in calculating the upper and lower bounds in the polarized context cluster. The polarized word in the cluster is acted upon by the valence shifters. Amplifiers increase polarity by 1.8 (0.8 is the default weight) and they become de-amplifiers if the context cluster contains an odd number of negators (two negatives equal a positive and 3 negatives equal a negative). De-amplifiers decrease polarity. Adversative conjunctions (AC) (e.g. but, however, although) also weight the cluster. AC before a polarized word up-weights the cluster by $1 + z(n_{AC})$ (with 0.85 being the default weight for $z_2$andn_{BAC} is the number of ACs before the polarized word. An AC after the polarized word down weights the cluster by$1 + \{n_{AAC} - 1\} * z_2$. The weights z may be provided by the use with the default being 0.8. Lastly, these weighted context clusters c_{ijl} are summed and divided by the square root of the word count (W_{ijn}) yielding the polarity score δ_{ij} for each tweet i.e. $\delta_{ij} = \sum c_{ij} / \sqrt{W_{ijn}}$

For the co-occurrence of words the study used the udpipe package [21]. The study considered only 63,056 tweets that were written in English.

3 Results

3.1 Tweet Polarity

Figure 4 shows the spatial location of all positive tweets. USA, Europe, Western and Southern Africa and India had high number of positive tweets.

Regions with high numbers of positive tweets also posted high number of negative tweets. This could indicate that individuals had opposing views on different issues (Fig. 5, 6).

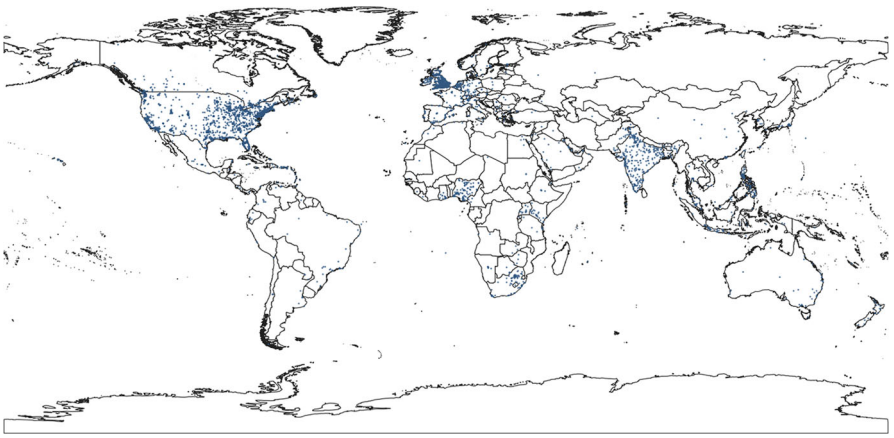

Fig. 4 Location of positive tweets

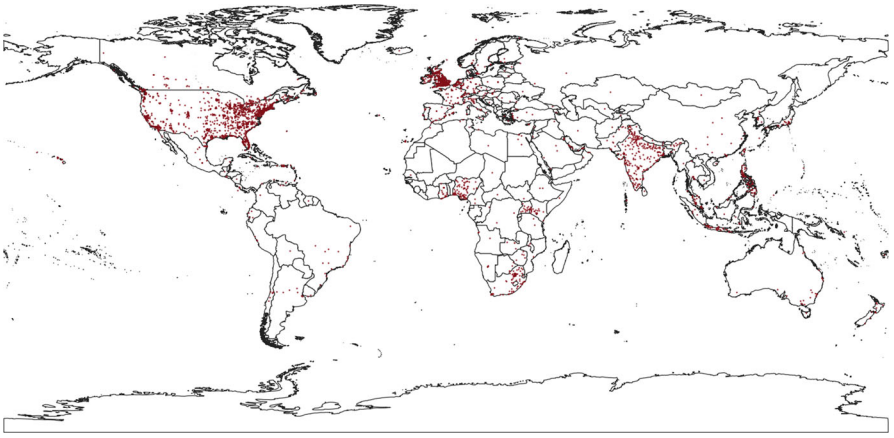

Fig. 5 Location of negative tweets

Figure 6 show the location of positive, negative and neutral tweets.

3.2 Word Co-occurrence

Spatial distribution of tweets: negative, positive and neutral is not that informative. A look at word co-occurrence may supply more insights on why tweets were negative, positive or neutral.

Figure 7 shows which words co-occurred with negative words. The thicker the path the more the co-occurrence. The blue dots depict the negative words while the red ones the words they occurred with. The strongest co-occurrence was mental and health. With many individual's routine lives altered, there is a risk of mental health problems. Qui j

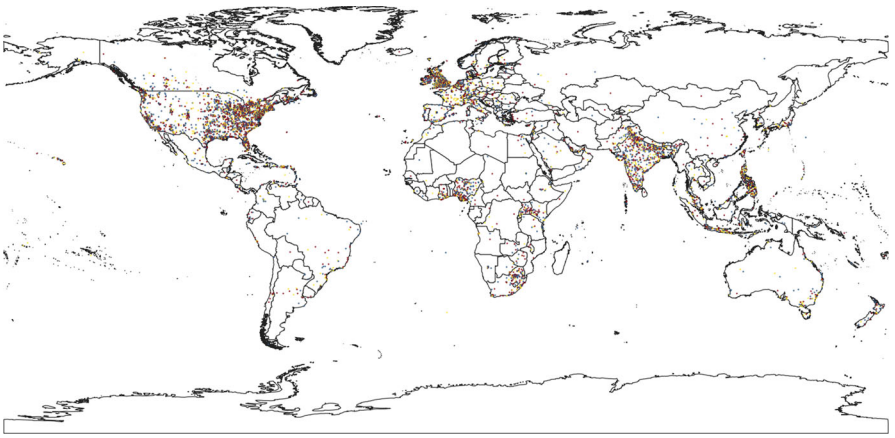

Fig. 6 Location of Positive (Blue), Negative (Red) and Neutral tweets (Yellow). (Color figure online)

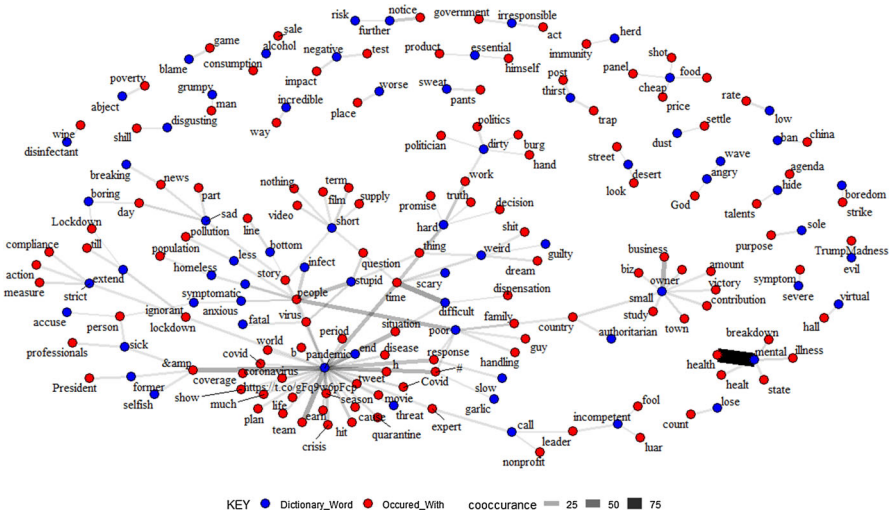

Fig. 7 Negative

et al. [22] in their nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic rightly captures the aftermath of the COVID-19 pandemic: "*The implementation of unprecedented strict quarantine measures in China has kept a large number of people in isolation and affected many aspects of people's lives. It has also triggered a wide variety of psychological problems, such as panic disorder, anxiety and depression*". These aspects are supported by Fig. 7.

Other strong co-occurrences were small-business which have been adversely affected by the lock down. In their survey on the effects of COVID-19 on small

businesses, Alexander W et al. [23] report that 43 percent of businesses are temporarily closed, and businesses have—on average—reduced their employee counts by 40 percent relative to January 2020.

The strongest co-occurrence was face and mask followed by hand and sanitizer (Fig. 8). These are some of the WHO recommended steps to curb the spread of COVID-19 [24]. This indicates an effective campaign to help curb the spread. Other concerns were food supplies, insecurity, school etc.

From Fig. 9, most tweets conveyed negative sentiment, understandably so because of the pandemic.

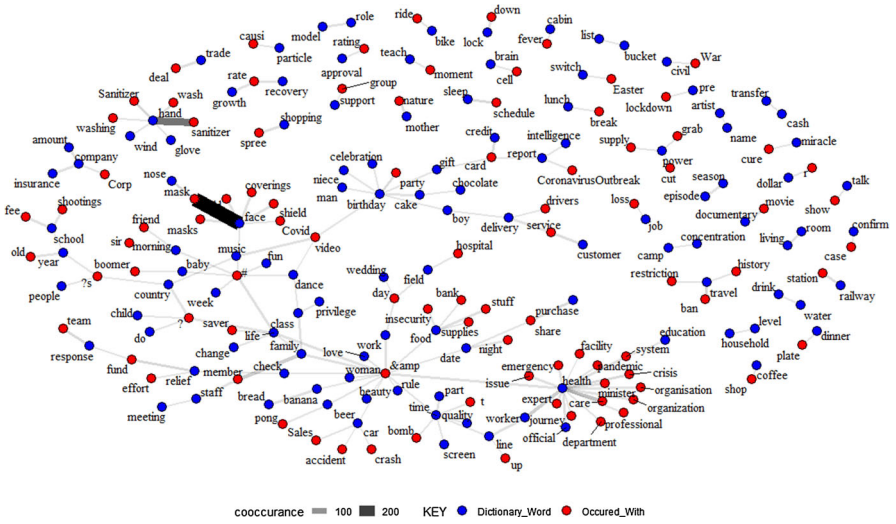

Fig. 8 Positive word co-occurrence

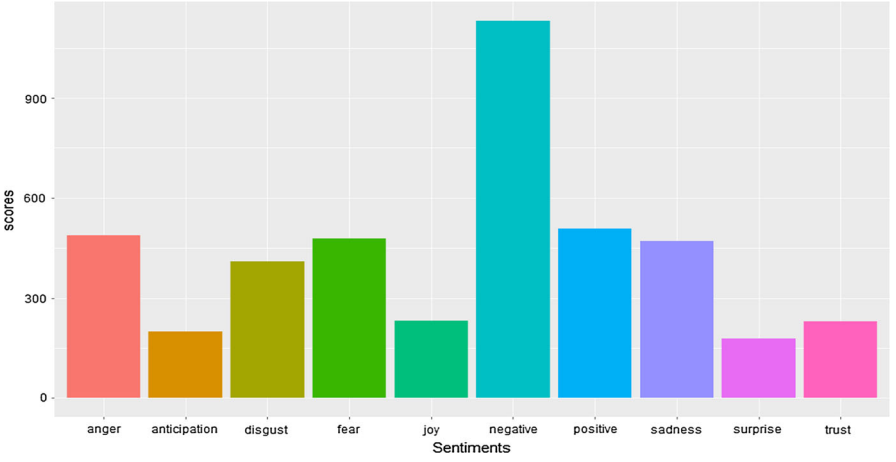

Fig. 9 General feel of the tweets

4 Discussions

This paper has demonstrated the wealth of information that is contained in sentiments expressed on social media, in this case Twitter. The direct effect of a great pandemic like the corona virus is death which can easily be measured. The indirect effects which range from loss of jobs [25], mental issues [22] to closing down of countries need other methods of quantification. Sentiment analysis is particularly useful in gauging the uptake of directives, emerging issues relating to the topic of interest among others, fake news, misinformation that may lead to fear and panic.

Results from sentiment analysis may help the government or relevant authorities relax, tighten or change approach altogether. Mental health was among the key concern among individuals, fear and panic was also evident Fig. 7 and Fig. 9. Studies [26, 27] have indicated that domestic violence is on the rise during this period of the coronavirus pandemic, a clear indication of mental anguish faced by the masses. Face-mask and hand-sanitizers also had high number of co-occurrence indicating that the sensitization efforts were working.

As it is with all studies, this one too has some shortfalls and limitations. Some of the shortfalls is that global Twitter data comes in various languages and as such methodologies to handle multilingual sentiment analysis are still in development. Our study focused on tweets written in English. The 2019 global multidimensional poverty index report indicates that 1.3 billion people or 23.1% are multidimensionally poor (in terms of health, education, standards of living) [28], This makes Twitter not a good platform to get insights from this group of people making it another downside of this study. Analysis of Twitter data for over a long period of time is computationally expensive as a whole day's tweets may be few hundred gigabytes. The study results relied on the data from a two day live stream, further work can be dedicated towards live streaming for a longer period or using historic data for over a longer period. These limitations however do not invalidate the results.

The results from this study may help governments combat the consequences of COVID-19 like mental health issues, lack of supplies e.g. food and also gauge the effectiveness or the reach of their guidelines. Li et al. [29] motivate the need for multifaceted approach in combating the COVID-19 pandemic. They stress that there is a need for more global collaboration to effectively combat the COVID-19 pandemic. They outline five pillars for achieving this including: Cross cultural collaboration and communication, strengthening of data and information sharing system, Adopting early experiences learned in other countries, evaluation and strengthening of public health systems and promoting of virtual communities to help improve mental health and well-being issues.

Acknowledgements The authors also thank the University of KwaZulu-Natal for its continued support for research and publication.

Author Contributions Elphas Okango: Conceptualization, Methodology, Writing-Original draft preparation, Software. Henry Mwambi: Conceptualization, Editing and Reviewing.

Availability of Data and Material The data set used for this study is available upon request.

Code Availability The code used for this study is available upon request.

Funding The authors received no specific funding for this work.

Declarations

Conflict of interest The author declares that no conflict of interest exists.

Ethical Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. van der Aalst W (2016) Data science in action. In: van der Aalst W (ed) Process mining: data science in action. Springer, Berlin, pp 3–23
2. Tien JM (2017) Internet of things, real-time decision making, and artificial intelligence. Ann Data Sci 4:149–178
3. Anon. Disease background of COVID-19.European Centre for Disease Prevention and Control. Available at https://www.ecdc.europa.eu/en/2019-ncov-background-disease.
4. Anon. WHO | Pneumonia of unknown cause – China.WHO. Available at http://www.who.int/csr/don/05-january-2020-pneumonia-of-unkown-cause-china/en/.
5. Anon. 2019-ncov-factsheet.pdf.
6. Anon. Coronavirus disease 2019.Available at https://www.who.int/emergencies/diseases/novel-coronavirus-2019.
7. Kumar S (2020) Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. Ann Data Sci 7:417–425
8. Khakharia A, Shah V, Jain S, Shah J, Tiwari A, Daphal P, Warang M, Mehendale N (2021) Outbreak prediction of COVID-19 for dense and populated countries using machine learning. Ann Data Sci 8:1–19
9. Liu Y, Gu Z, Xia S, Shi B, Zhou X-N, Shi Y, Liu J (2020) What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization. EClinicalMedicine 22:100354
10. Olson DL, Shi Y, Shi Y (2007) Introduction to business data mining, vol 10. McGraw-Hill/Irwin, New York
11. Shi Y, Tian Y, Kou G, Peng Y, Li J (2011) Optimization based data mining: theory and applications. Springer Science & Business Media, Berlin
12. Anon. (2020). What is Sentiment Analysis?MonkeyLearn Blog. Available at https://monkeylearn.com/blog/what-is-sentiment-analysis/.
13. Nguyen TH, Shirai K, Velcin J (2015) Sentiment analysis on social media for stock movement prediction. Expert Syst Appl 42:9603–9611
14. Carchiolo V, Longheu A, Malgeri M (2015). Using twitter data and sentiment analysis to study diseases dynamics. In: International conference on information technology in bio-and medical informatics pp 16–24. Springer.
15. Anon. (2020). Twitter. *Wikipedia*.
16. Barbera P, Barbera MP, Roauth S (2018). Package 'streamR.'
17. Anon. joachim-gassen/tidycovid19: {tidycovid19}: An R Package to Download, Tidy and Visualize Covid-19 Related Data.Available at https://github.com/joachim-gassen/tidycovid19.
18. Medhat W, Hassan A, Korashy H (2014) Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J 5:1093–1113
19. Maynard D, Funk A (2011). Automatic detection of political opinions in tweets. In: Extended semantic web conference pp 88–99. Springer

20. Rinker T(2017). Package 'sentimentr.' *Retrieved* **8** 31.
21. Wijffels J, Straka M, Strakov J (2018). udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the'UDPipe"NLP'Toolkit. *R package version 0.5*.
22. Qiu J, Shen B, Zhao M, Wang Z, Xie B, Xu Y (2020) A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. Gen Psychiatry 33:e100213
23. Bartik AW, Bertrand M, Cullen ZB, Glaeser EL, Luca M, Stanton CT (2020). How are small businesses adjusting to COVID-19? Early evidence from a survey. National Bureau of Economic Research.
24. ANON. Advice for public.Available at https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public.
25. Adams-Prassl A, Boneva T, Golin M, Rauh C (2020) Inequality in the impact of the coronavirus shock: new survey evidence for the UK. J Publ Econ 189:104245
26. Taub A(2020) A new Covid-19 crisis: domestic abuse rises worldwide. New York Times **6**.
27. Bradbury-Jones C, Isham L (2020) The pandemic paradox: the consequences of COVID-19 on domestic violence. J Clin Nurs 29:2047–2049
28. ANON. The 2019 Global Multidimensional Poverty Index (MPI) | Human Development Reports.Available at http://hdr.undp.org/en/2019-MPI.
29. Li J, Guo K, Viedma EH, Lee H, Liu J, Zhong N, Gomes LFAM, Filip FG, Fang S-C, Özdemir MS, Liu X, Lu G, Shi Y (2020) Culture versus policy: more global collaboration to effectively combat COVID-19. The Innovation 1:100023

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.