# On Public Sentiments Towards COVID–19 Vaccines in South African Cities: An Analysis of Twitter Posts.

Blessing Ogbuokiri, Ali Ahmadi, Nicola Bragazzi, Zahra Nia, Bruce Mellado, Jianhong Wu, James Orbinski, Ali Asgary, Jude Kong

Submitted to: Journal of Medical Internet Research
on: February 02, 2022

**Disclaimer: © The authors. All rights reserved.** This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

## Table of Contents

# On Public Sentiments Towards COVID–19 Vaccines in South African Cities: An Analysis of Twitter Posts.

Blessing Ogbuokiri[1] PhD; Ali Ahmadi[2] PhD; Nicola Bragazzi[3] PhD; Zahra Nia[1] PhD; Bruce Mellado[4*] PhD; Jianhong Wu[3*] PhD; James Orbinski[5*] PhD; Ali Asgary[6*] PhD; Jude Kong[1*] PhD

[1]Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC) York University Toronto CA
[2]Faculty of Computer Engineering K.N. Toosi University Tehran IR
[3]Laboratory for Industrial and Applied mathematics York University Toronto CA
[4]School of Physics, Institute for Collider Particle Physics University of the Witwatersrand Johannesburg ZA
[5]Dahdaleh Institute for Global Health Research York University, Toronto CA
[6]Advanced Disaster, Emergency and Rapid-response Simulation (ADERSIM) York University Toronto CA
[*]these authors contributed equally

**Corresponding Author:**
Jude Kong PhD
Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC)
York University
4700 Keele Street
Toronto
CA

## Abstract

**Background:** While vaccination against the coronavirus (COVID–19) lasts, Twitter has become one of the social media platforms used to generate discussions about the COVID–19 vaccination. These types of discussions most times lead to a compromise of public confidence towards the vaccine. The text-based data generated by these discussions are used by researchers to extract topics and perform sentiment analysis on provincial, country, or continent level without considering the local communities.

**Objective:** The aim of this study is to use geoclustering of Twitter posts to inform city-level variations in sentiments toward COVID–19 vaccine-related topics in the three largest South African cities (Cape Town, Durban, and Johannesburg).

**Methods:** We generated a dataset and processed (n=25,000) COVID–19 vaccine-related tweets in South Africa from January 2021 to August 2021 using the academic researcher Twitter Application Programming Interface (API) with keywords like vaccine, vaccination, AstraZeneca, Oxford-AstraZeneca, VaccineToSaveSouthAfrica, JohnsonJohnson, and pfizer. Tweets were mapped with their geolocation. The Latent Dirichlet Allocation was used to identify frequently discussed topics across the cities. Senti- ments (negative, neutral, and positive) scores were assigned using the Valence Aware Dictionary with Support Vector Machine classification algorithm.

**Results:** The number of new COVID–19 cases significantly positively correlated with the number of Tweets in South Africa (Corr=0.462, P<.001). Out of the 10 topics identified from the tweets, 2 were about the COVID–19 vaccines: uptake and supply, respectively. The intensity of the sentiments score for the two topics was associated with the total number of vaccines administered in South Africa (P<.001). Discussions regarding the two topics showed higher intensity scores for neutral sentiment class (P=.015) than for other sentiment classes. Additionally, the intensity of the discussions for the two topics was associated with the total number of vaccines administered, new cases, deaths, and recoveries across the three cities (P<.001). The sentiment score for the most discussed topic, vaccine uptake, differed across the three cities, with (P=.003), (P=.002), and (P<.001) for positive, negative, and neutral sentiments classes, respectively.

**Conclusions:** The outcome of this research showed that geolocation clustering of Twitter posts can be used to better analyze the sentiments towards COVID–19 vaccines at the local level. This can provide additional city–level information to health policy and decision-making regarding COVID–19 vaccine hesitancy.

(JMIR Preprints 02/02/2022:36796)
DOI: https://doi.org/10.2196/preprints.36796

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

# On Public Sentiments Towards COVID–19 Vaccines in South African Cities: An Analysis of Twitter Posts.

Blessing Ogbuokiri[1,2]. Ali Ahmadi[3], Nicola Bragazzi [1,2], Zahra Movahedi Nia[1,2], Bruce Mellado[1,5,*], Jianhong Wu[1,2,*], James Orbinski[1,6,*], Ali Asgary[1,4,*], and Jude Kong [1,2,*,†]

[1]Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), York University, Canada

[2]Laboratory for Industrial and Applied mathematics, York University, Canada

[3]K.N. Toosi University, Faculty of Computer Engineering, Tehran, Iran.

[4]Advanced Disaster, Emergency and Rapid-response Simulation (ADERSIM), York University, Toronto, Ontario, Canada

[5]School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, South Africa

[6]Dahdaleh Institute for Global Health Research, York University, Canada

*These authors have contributed equally to this work and share last authorship
†Corresponding Author: jdkong@york.ca

## Abstract

**Background:** While vaccination against the coronavirus (COVID–19) lasts, Twitter has become one of the social media platforms used to generate discussions about the COVID–19 vaccination. These types of discussions most times lead to a compromise of public confidence towards the vaccine. The text-based data generated by these discussions are used by researchers to extract topics and perform sentiment analysis on provincial, country, or continent level without considering the local communities.
**Objective:** The aim of this study is to use geoclustering of Twitter posts to inform city-level variations in sentiments toward COVID–19 vaccine-related topics in the three largest South African cities (Cape Town, Durban, and Johannesburg).

**Method:** We generated a dataset and processed (n=25,000) COVID–19 vaccine-related tweets in South Africa from January 2021 to August 2021 using the academic researcher Twitter Application Programming Interface (API) with keywords like *vaccine*, *vaccination*, *AstraZeneca*, *Oxford-AstraZeneca*, *VaccineToSaveSouthAfrica*, *JohnsonJohnson*, and *pfizer*. Tweets were mapped with their geolocation. The Latent Dirichlet Allocation was used to identify frequently discussed topics across the cities. Senti- ments (negative, neutral, and positive) scores were assigned using the Valence Aware Dictionary with Support Vector Machine classification algorithm.

**Result:** The number of new COVID–19 cases significantly positively correlated with the number of Tweets in South Africa (Corr=0.462, P<.001). Out of the 10 topics identified from the tweets, 2 were about the COVID–19 vaccines: uptake and supply, respectively. The intensity of the sentiments score for the two topics was associated with the total number of vaccines administered in South Africa (P<.001). Discussions regarding the two topics showed higher intensity scores for neutral sentiment class (P=.015) than for other sentiment classes. Additionally, the intensity of the discussions for the two topics was associated with the total number of vaccines administered, new cases, deaths, and recoveries across the three cities (P<.001). The sentiment score for the most discussed topic, vaccine uptake, differed across the three cities, with (P=.003), (P=.002), and (P<.001) for positive, negative, and neutral sentiments classes, respectively.

**Conclusion:** The outcome of this research showed that geolocation clustering of Twitter posts can be used to better analyze the sentiments towards COVID–19 vaccines at the local level. This can provide additional city–level information to health policy and decision-making regarding COVID–19 vaccine hesitancy.

**KEYWORDS:** COVID-19; Vaccine; Vaccination; Sentiment Analysis; Tweets; South Africa; Hesitancy

# Introduction

Despite a few antivirals that have been approved very recently by the US FDA against coronavirus [1], preventive measure(s) against the virus is still very relevant [2, 3]. According to World Health Organization (WHO) [4, 5], vaccination is one of the primary preventive measure against the novel coronavirus, in addition to other measures already in place to curb the spread of the virus such as social distancing, the use of face masks, sanitization, and isolation [6].

To vaccinate or not to vaccinate has become a very important question facing communities in South Africa and Africa at large as the COVID-19 pandemic lasts [7, 8]. As vaccine uptake across South Africa increase, new cases and deaths because of the COVID-19 virus remain [7, 9, 10]. The unvaccinated people with serious illness and fatalities are the most admitted as reported by most hospitals in South Africa [9].

However, public bias or sentiments influenced by some religious leaders, social media influencers or legal restrictions, as reflected in most of the anti-vaccination messages on social media platforms [5, 11, 12], may have a significant impact on the progression toward achieving vaccination against COVID–19 in South Africa, especially in the local communities [10, 13]. Social Media platforms are applications that enable communication amongst users or groups to interact, share, or reshare information on the Internet using different platforms or devices within the comfort of their homes [5, 14]. Information sharing on social media spread very fast even if it is a rumour from an unverified source. The impact of rumours is always dangerous, especially in places where users are not well informed about the subject of discussion [11].

Twitter being one of the most influential social media plat- forms, has become a good tool for sharing news, information, opinions, and emotions about COVID–19 vaccine-related discussions [6, 11, 15, 16]. As Twitter users remain connected while observing COVID–19 restrictions, misinformation, un- confirmed rumours, vaccination and anti-vaccination messages regarding COVID–19 continue to spread [3, 14, 17, 18]. These messages, which are mostly text-based, spread in the form of users' posts or retweets, without confirming their sources. These types of discussions keep weakening the confidence level of the public well before they are vaccinated [15, 19, 20]. Given a large amount of text-based data from Twitter, a lot of research has leveraged on it to draw insight and make predictions on the users' sentiment of the COVID–19 vaccines at a continent, country, or province level while neglecting the local communities [12, 16, 17, 21].

In this study, we used Twitter posts to inform location clustering in city-level variations in sentiments toward COVID–19 vaccine-related topics in the three largest South African cities (Cape Town, Durban, and Johannesburg). We started with an analysis of Twitter posts from the South African context from January 2021 to August 2021 to understand the popular topics that are being discussed within the period. Then, an exploration of users' sentiments toward the vaccines and how they inform vaccine uptake was conducted. Finally, we performed a comparison of the popular topics and sentiments across the three cities. The approach used in this research can inform geolocation clustering of Twitter posts, which can be used to better analyze the sentiments towards COVID–19 vaccines related topics at a local level.

# Methods

## Data Collection

With an existing Twitter account, we applied for Developer Access and was granted access to Twitter Academic Researcher API, which allows for over 10 million tweets per month. Then, we created an application to generate the API credentials (access token) from Twitter. The access token was used in Python (v3.6) script to authenticate and establish a connection to the Twitter database. To get goetagged vaccine-related tweets, we used the python script we developed to perform an archive search[1] of vaccines related keywords with place country South Africa (ZA) from January 2021 to August 2021. These vaccine-related keywords include but are not limited to *vaccine, anti-vaxxer, vaccination, AstraZeneca, Oxford-AstraZeneca, IChooseVaccination, VaccineToSaveSouthAfrica, JohnsonJohnson, and pfizer.* A complete list of the keywords used can be found in Appendix A. The preferred language of the tweet is English.

A total of ~45,000 tweets was pulled within the period of discussion using the archive search. Each Tweet contains most of the following features: TweetText, TweetID, Create- Date, RetweetCount, ReplyCount, LikeCount, GeoId, GeoCityProvince, GeoCountry, GeoCoordinate(bbox), and Location. Others are AuthorID, UserName, Hashtags, CreatedAccountAt, FollowerCount, FollowingCount, and TweetCount.

Additionally, daily statistics of new COVID-19 cases, vaccinations, deaths, and recoveries in Cape Town, Durban, and Johannesburg were used. These statistics were obtained from the South African coronavirus official website [21] and the COVID-19 South Africa dashboard [22] which are primarily updated every day.

## Data Preprocessing

User tweets contain a lot of information about the data they represent. That means, raw tweets that have not been processed are highly unstructured and contain a lot of redundant information. To overcome these issues, preprocessing of raw tweets have become increasingly paramount. In this study, the tweets, date created, location,

[1] Twitter archive search allows a user to get all Tweets (including Retweets) going back to the beginning.

country, and geocoordinate (bbox)[2] were extracted from the dataset into a dataframe using pandas (v1.2.4) [24].

We prepared the tweets for Natural Language Processing (NLP) by removing the URLs, duplicate tweets, tweets with incomplete information, punctuations, special and non alphabetical characters, emojis, non–English words, and stopwords using the tweets-preprocessor toolkit (v0.6.0) [25], Natural Language Toolkit (NLTK; v3.6.2) [26], and Spacy2 toolkit (v3.2) [27, 28]. The Spacy2 toolkit was used to perform lowercase and to tokenize the tweets. This process reduced the tweets in the dataset to ~25,000. The word-cloud generated from the dataset shows vaccine as one of the most frequent words, see Figure 1.

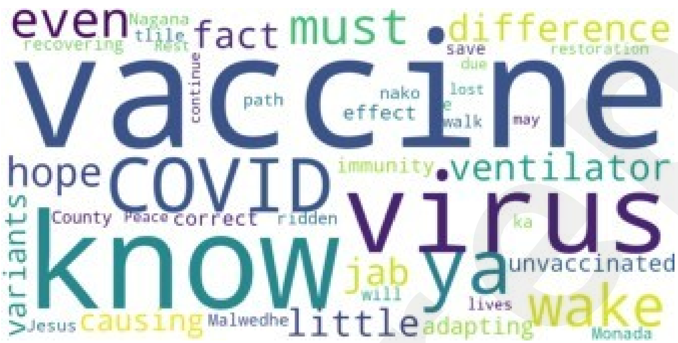


**Figure 1.** The most frequently used words of our dataset

### Sentiment Analysis

We applied VADER (Valence Aware Dictionary for Sentiment Reasoning) [29] available in the NLTK package to the tweets to get compound scores and assign labels to the tweets. The compound score was used to determine when the label (positive, negative, or neutral) can be assigned to a tweet. A compound score ≥ 0.5, < 0, and *x*, where *x* satisfies the inequality 0.5 > $x$ ≥ 0 are assigned the label positive, negative, and neutral, respectively. Further, we randomly selected 2,500 (10%) of the tweets and manually labeled them as positive, negative, or neutral (See Figure 2) to ensure VADER accurately labeled the tweets.

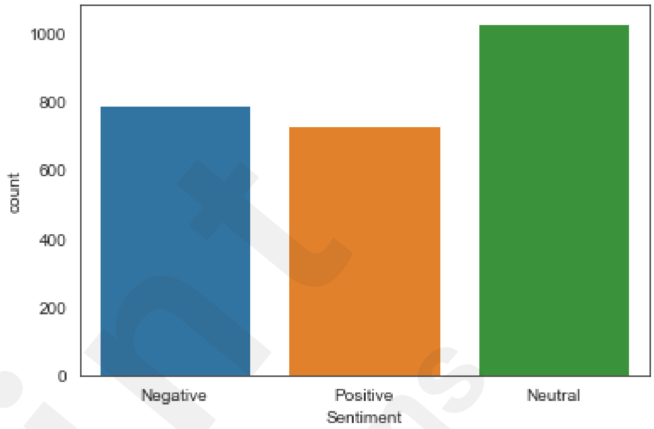


**Figure 2.** Distribution of manually labeled tweets.

The dataset was divided into two parts, namely, training set (0.80) and test set (0.20) which was fitted into the Support Vector Machines (SVM) [30, 31] classification algorithm for evaluation and prediction. The Root Mean Square Error (RMSE) [32] was used to compare the model outputs with the true values.

### Topic Modeling

The Latent Dirichlet Allocation (LDA) [33] model was used for the topic modeling of the dataset through the Gensim (v4.0.1) [34] package in python. The LDA was used because it is assumed to be one of the popular models for this type of analysis, is easy to use, and has been successfully used and implemented in recent studies such as [15] and [12]. The LDA models was created for 1 to 10 topics to optimize the number of topics. Jaccard similarity [35] and coherence measure [36] tests were administered and calculated across the topics. The Jaccard similarity (*sklearn.metrics.jaccard_score*) package in python

[2] A bounding box (usually shortened to bbox) is an area defined by two longitudes and two latitudes, where: Latitude is a decimal number between -90.0 and 90.0. Longitude is a decimal number between -180.0 and 180.0. [23]

was used for a test of the uniqueness of the topics while the coherence (*gensim.models.coherencemodel*) package using the c_v[3] option in python was used for the measure of the degree of similarity between high scoring words in the topic.

### Test Statistics

The trends in time between vaccine-related tweets and new COVID–19 cases in South Africa were compared using the Granger causality test[4] from the *statsmodels.tsa.stattools import grangercausalitytests* package in python. The correlation coefficient was calculated using the Pearson correlation from the *scipy.stats.pearsonr* package in python. Further, the intensity of the sentiments of each vaccine-related topic was also compared using the Mann–Whitney *U* test[5] from the *scipy.stats.mannwhitneyu* package in python.

Similarly, the time series trends for the intensity of the vaccine-related topics for each of Cape Town, Durban, and Johannesburg were compared to the total vaccinations, new cases, deaths, and recoveries using the Granger causality test. The Mann–Whitney *U* test was used to compare the distribution of the sentiment intensity for each vaccine-related topic for each city.

Finally, the sentiment intensity distribution for the vaccine trending topic across the three cities was compared using the Kruskal Wallis *H* test[6] from the *scipy.stats.kruskal* package in python.

# Results

### Our Dataset in South African Context

The Figure 3 shows the summary statistics of our dataset in the South African context with time. As shown in Figure 3, there is upward growth in the number of tweets in the first, second, and third weeks of January and February. However, there are some levels of consistency in growth in the number of vaccine- related tweets for every other week of the month.

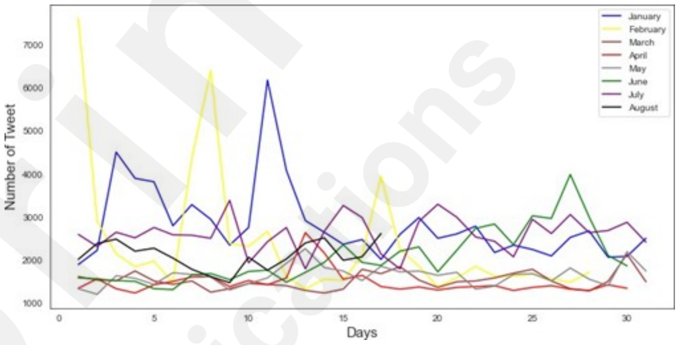


**Figure 3.** South Africa vaccine-related Tweets from January to August 2021.

In Figure 4, the trend in the growth of the daily tweets and daily COVID–19 cases proved to be consistent with time. For instance, the upward growth in the number of daily tweets correlated with growth in the number of daily new COVID-19 cases for January and July. Similarly, the decline in the numbers of daily tweets and a daily number of cases demonstrated a similar trend over time. Therefore, the number of new COVID-19 cases significantly positively correlated with the number of Tweets (*corr* = 0.462, p<.001, and 95% CI) as shown in Figure 5.

[3] ***C_v*** measure is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity [37]

[4] The Granger causality test is a statistical hypothesis test for determining whether one time series is a factor and offer useful information in forecasting another time series [38].

[5] Mann-Whitney *U* test is used for comparing differences between two independent groups. It tests the hypothesis that if the two groups come from same population or have the same medians [39].

[6] Kruskal-Wallis test used for comparing the differences between two or more groups. It is an extension to the Mann Whitney *U* Test, which is used for comparing two groups. It compares the mean ranks (medians) of groups. [40]

**Figure 4.** Number of new COVID-19 cases in South Africa and the number of daily Tweets.

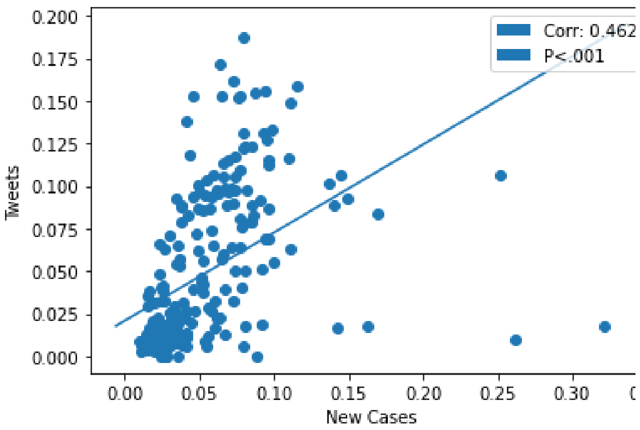


**Figure 5.** The number of new COVID–19 cases significantly positively correlated with the number of Tweets.

### Sentiment in the South African Context

The SVM model classified the tweets as positive (0.313) negative (0.304), and neutral (0.383) and achieved an accuracy of 0.79 on our dataset with a Mean Absolute Error (MAE): 0.04, Mean Squared Error (MSE): 0.06, and Root Mean Squared Error: 0.07. This shows that the error rate of our prediction is low. Figure 6 summaries the distribution of the sentiment classes over time.

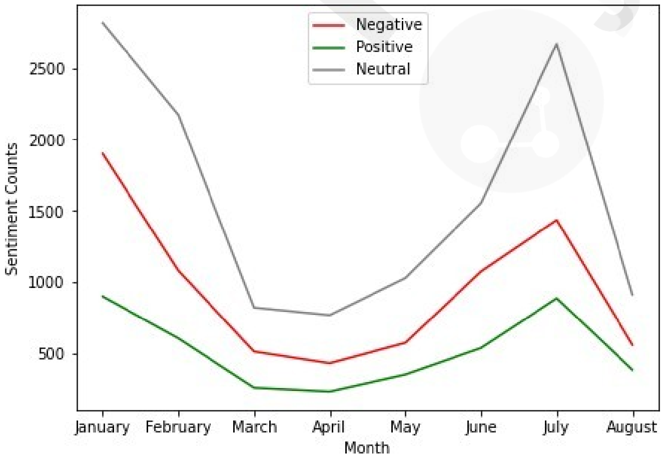
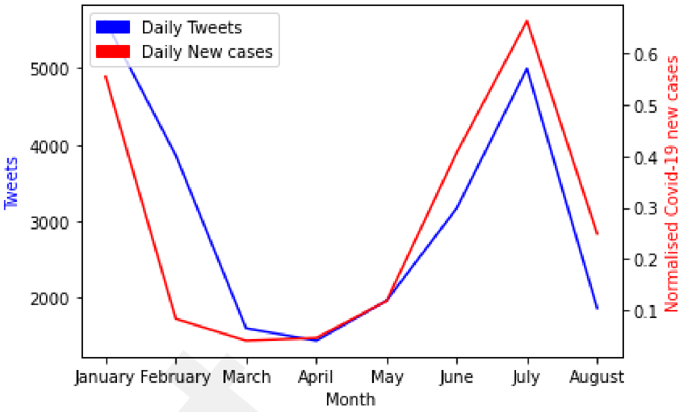


**Figure 6.** Tweets sentiments classes in South Africa from January to August 2021.



As shown in Figure 6, there is growth in sentiment with time. In January and July, the neutral sentiment class maintained an upward growth followed by the negative and positive sentiment classes, respectively. Additionally, there is a decline in growth in April and August for all the sentiment classes. The difference in sentiment classes between January and the other months is statistically significant (p<.001 and 95% CI).

### Identifying COVID–19 Vaccine Topics in the South African Context

After the application of the LDA model on the tweets, about 45 topics were generated. Some of which are the same and incoherent by observation. We applied the Jaccard similarity test to ascertain the uniqueness of each topic. The Jaccard similarity counts the number of similar words in two topics and divides it by the total number of words from the two topics combined. If the Jaccard similarity value is 1 it shows that the topics are the same and 0 otherwise. We considered topics whose Jaccard similarity value is less than 0.5. However, to be sure that the topics identified are semantically acceptable, we performed a coherence measure test on the topics. High coherence measure value shows that the topic could be meaningful, hence we chose topics that yielded high coherence values.

This process reduced the generated topics to 10 unique topics. The topics are vaccine uptake (topic 1), social distancing (topic 2), xenophobic attack (topic 3), travel restrictions (topic 4), alcohol ban (topic 5), religion (topic 6), sports (topic 7), border closer (topic 8), politics (topic 9), and vaccine supply (topic 10). Two topics were identified to be relevant to this study, which is, vaccine uptake (topic 1) and vaccine supply (topic 10). The first ten top-scoring representative words for topics 1 and 10 and their possible interpretations are shown in Table 1.

**Table 1:** Selected LDA generated topics and their interpretations.

| **Topic Number** | **Representative Word** | **Possible Interpretations** |
|---|---|---|
| 1 | jab, vaccin, pfizer, get, first, dose, jampj, second, got, done | Got my first jab.<br>Done with my first pfizer vaccine jab.<br>Recieved second a dose of the johnson and johnson vaccine. |
| 10 | Vaccin, covid, people, govern, countri, money, world, supply, sa, virus | This topic focuses on the need for the South African government to pay for the supply of more COVID–19 vaccine. |

Next, we compared the level of the vaccine-related discussions to the number of people vaccinated (see Figure 7).

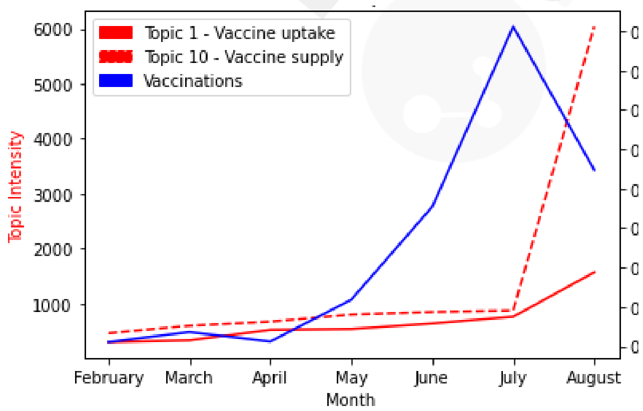
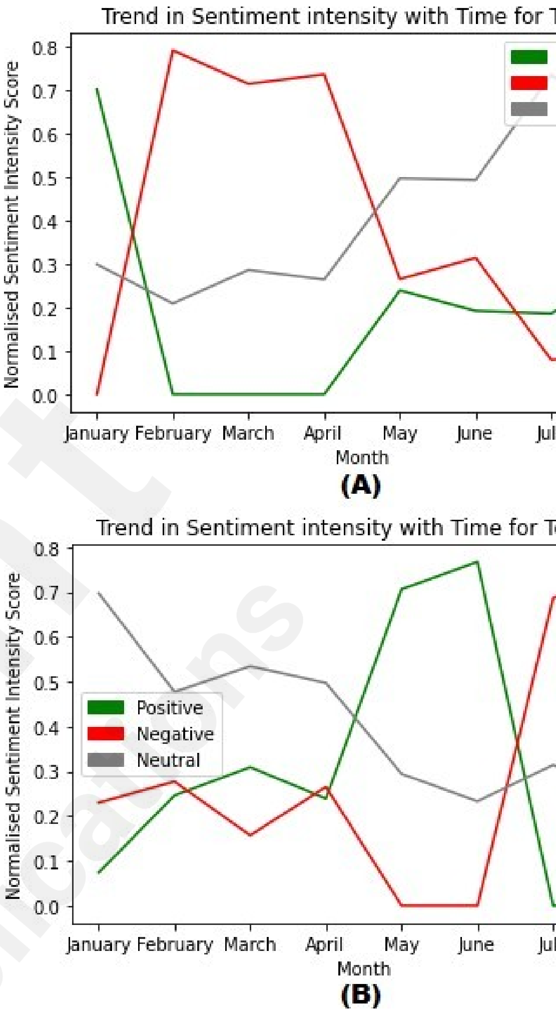


**Figure 7.** Comparing vaccine-related discussions with the number of people vaccinated.



We observed that the intensity of topics 1 (red line) and 10 (dashed red line) started increasing almost at the same pace from February[7] to June with topic 10 slightly higher. In July, the two topics showed high intensity than the other months. Topic 10 started to grow upward from July to August. However, comparing this outcome with the total number of people vaccinated (blue line) within the period. We observed that, while the vaccine-related discussions increase from February to July, the total number of vaccinations also increased. July to August showed a sharp decrease in vaccination while the vaccine-related discussions increased. An evaluation of the impact of vaccine-related discussions on the total vaccinations using the Granger test for causality showed that an increase in vaccine-related discussions correlates with the number of people vaccinated, P=.004 for the two topics.

[7] We ignored January because the rollout of vaccines started from February in South African as shown in the data available to us, see [21].

Further, we present the distribution and compare the differences in sentiment intensity scores between the two topics in Figure 8.

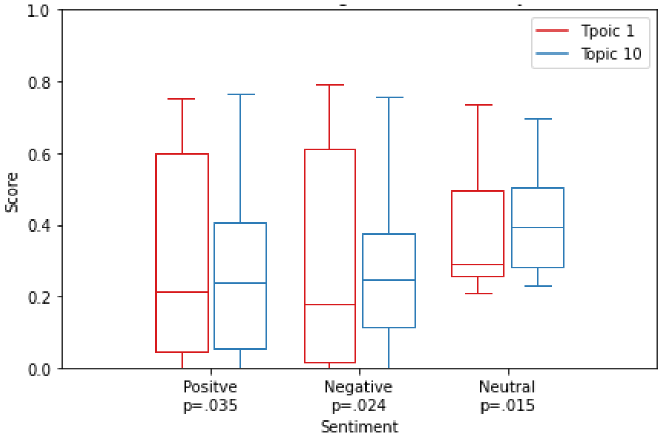


**Figure 8.** Distribution and comparison of sentiment intensity scores for vaccine-related discussions in South Africa using the Mann–Whitney *U* test.

We observed that the sentiment intensity score for both topics had significantly higher scores for neutral class (P=.015) than the negative sentiment (P=.024) and the positive sentiment (P=.035) classes respectively as shown in Figure 8.

We further investigated the differences in trends in sentiment intensity scores between the two topics with time (see Figure 9 (A) and (B)). In January, the sentiment intensity for negative sentiment started to trend upward and downward for the positive sentiment. Both flattened between February and April. There was a sharp decline between April and May for the negative sentiment intensity and an increase for the positive sentiment within the period. However, the intensity for the neutral sentiment trended upward with time for Topic 1 (see Figure 9 (A)).

Similarly, for topic 10, the sentiment intensity started to trend upward for the positive and negative sentiments in January. While the positive sentiment intensity score continued to trend upward until June when it started to decline, the negative sentiment intensity score trended downward until June when it trended upward. Additionally, the neutral sentiment intensity continued to trend downward with time for topic 10 (see Figure 9(B)).

**Figure 9.** Trends in sentiment intensity with time for vaccine related discussions.

### City- Level Analysis of Vaccine Discussions

To investigate the city-level analysis of vaccine-related discussions, we selected tweets for three major cities namely, Cape Town (n = 2, 484), Durban (n= 1, 020), and Johannesburg (n = 2, 898) from the South African dataset that we preprocessed and labeled. We chose these cities because they are the largest cities by population in South Africa [41]. Figure 10 shows the distribution of the selected tweets according to the location.

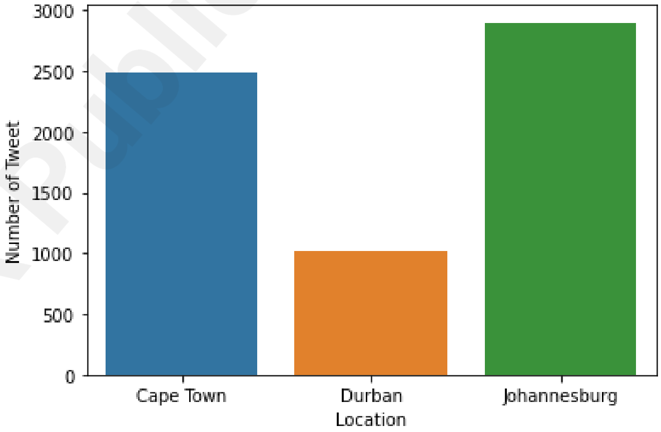


**Figure 10.** South Africa vaccine-related tweets according to selected location.

We also present the distribution of the sentiments of the preprocessed tweets according to each selected city (see Figure 11).

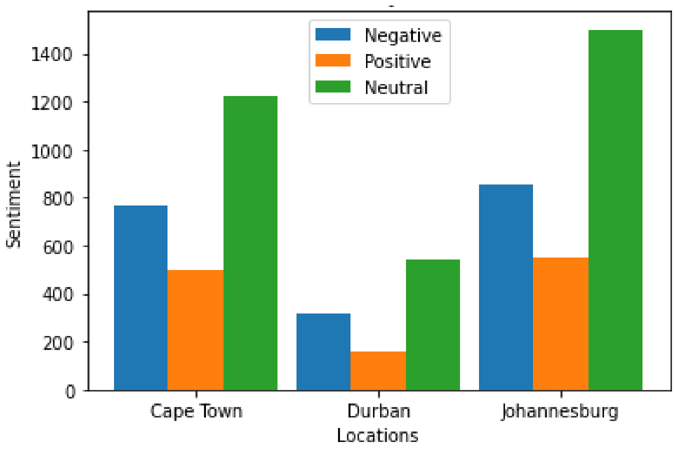


**Figure 11.** City-Level COVID-19 vaccine-related tweet sentiments classification.

The reason for this is to enable us to identify the city-specific discussions and to analyze the intensity of their sentiments. We applied the LDA model on the preprocessed tweets at city- level. Then, Jaccard similarity and coherence tests were applied to the topics. These processes enabled us to identify two unique topics that are relevant to our research across each city. These topics are vaccine uptake (Topic 1) and vaccine supply (Topic 3).

### Cape Town Specific Analysis

We compared the level of the vaccine-related discussions in Cape Town to the number of people vaccinated, new cases, deaths, and recoveries in the Western Cape Province (see Figure 12). This is because there is no South African city- level COVID–19 Data [21, 22] accessible to us at the time of this research. We chose to compare the Western Cape Province data to the Cape Town city vaccine-related discussions because Cape Town is the largest city in the Western Cape Province [41].

As the intensity score for topics 1 and 3 trends upward, total vaccinations also increased from February to August. The evaluation of the impact of vaccine-related discussions on the total vaccinations in the Western Cape province using the Granger test for causality showed a strong statistically significant correlation P<.001 for the two topics. However, as the intensity of topics 1 and 3 increase, the number of new cases and recoveries also increase but at a slow pace from February to August. There is a sharp decline in the number of COVID–19 related deaths in April than other months. The impact of vaccine-related discussions on the new cases, deaths, and recoveries showed a weak correlation (P =.07), see Figure 12.

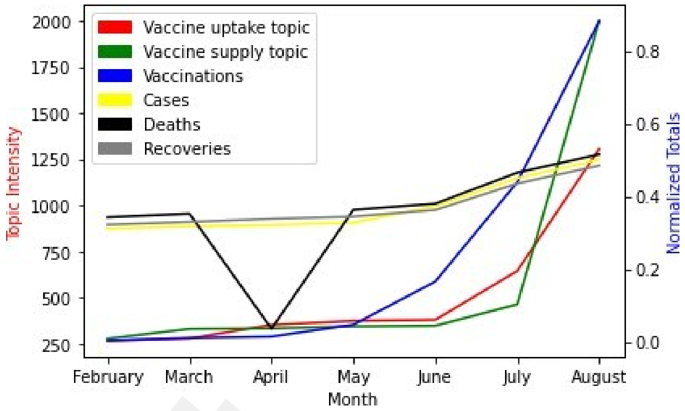


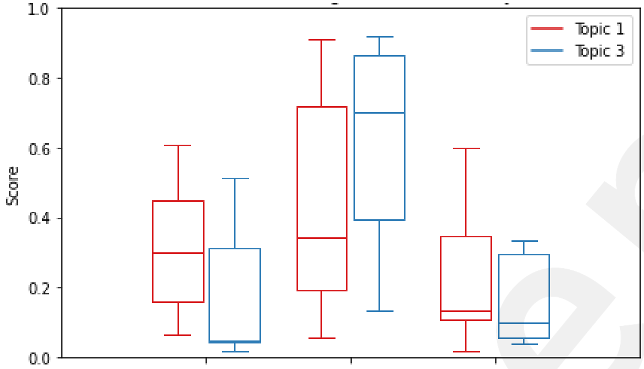
**Figure 12.** Cape Town vaccine-related discussions with the normalized COVID–19 different totals.

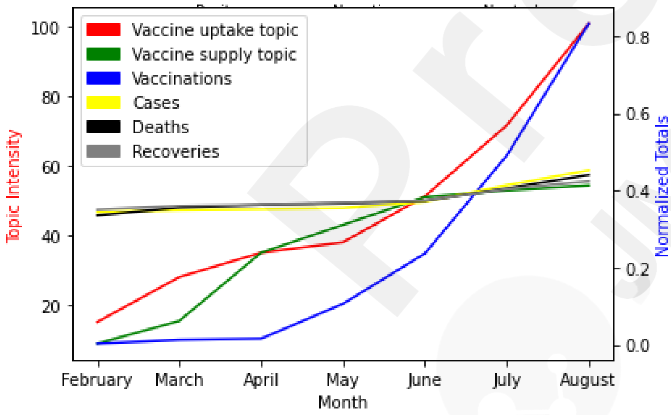
The distribution of the sentiment intensity scores is shown in Figure 13. A comparison of the differences in sentiment intensity scores between the two topics (1 and 3) in Cape Town demonstrated a higher sentiment intensity score for both topics for the neutral sentiment (P=.03) class and the positive sentiment (P=.04) class than the negative sentiment (P=.04) class respectively (see Figure 13).

**Figure 13.** Distribution and comparison of sentiment intensity scores for vaccine-related discussions in Cape Town using the Mann–Whitney *U* test.

### Durban Specific Analysis

While the intensity of topics 1 and 3 trends upward, total vaccinations increased almost at the same pace from February to August. The evaluation of the impact of vaccine-related discussions on the total vaccinations in the KwaZulu-Natal province using the Granger test for causality showed a strong statistically significant correlation P<.001 for the two topics. However, as the intensity of the two topics increased, the number of new cases, deaths, and recoveries almost flattened from February to August. The evaluation of the impact of vaccine-related discussions on the new cases, deaths, and recoveries showed a weak correlation (P =.07), see Figure 14.





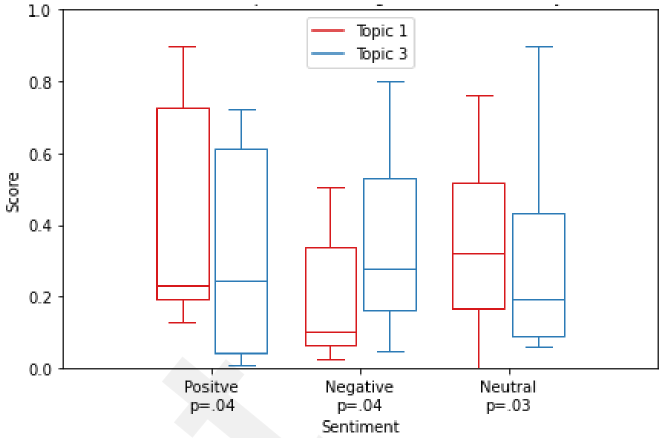


**Figure 14.** Durban vaccine-related discussions with the normalized COVID–19 different totals.

Next, we present the distribution and compare the differences in sentiment intensity scores between the two topics (1 and 3) for Durban city. We observed that the sentiment intensity score for both topics demonstrated higher scores for the negative sentiment (P=.03) than the positive sentiment (P=.02) and the neutral sentiment (P=.04) respectively (see Figure 15).

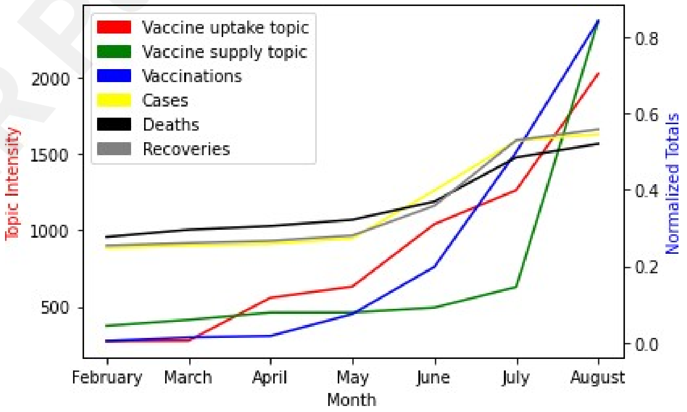


**Figure 15.** Distribution and comparison of sentiment intensity scores for vaccine-related discussions in Durban using the Mann–Whitney *U* test.

### Johannesburg Specific Analysis

Unlike Cape Town and Durban, Johannesburg showed a strong correlation on the impact of the vaccine-related discussions on new cases, deaths, and recoveries in the Gauteng province (P=.03) with time. Similarly, as the intensity score for topics 1 and 3 trends upward, total vaccinations also increased from February to August. The evaluation of the impact of vaccine- related discussions on the total vaccinations showed a strong statistically significant correlation P<.001 for the two topics (see Figure 16).

Further, we show the distribution of the sentiments for each topic and compare the differences in sentiment intensity scores between the two topics as well.

**Figure 16.** Johannesburg vaccine-related discussions with the normalized COVID-19 different values.

We observed that the sentiment intensity scores for both topics demonstrated higher scores for the neutral sentiment (P=.04) than the positive sentiment (P=.04) and the negative sentiment (P=.02) respectively (see Figure 17).

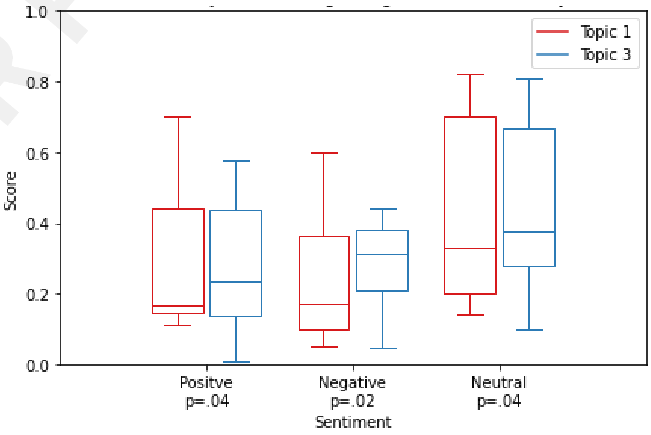
**Figure 17.** Distribution and comparison of sentiment intensity scores for vaccine-related discussions in Johannesburg using the Mann–Whitney *U* test.

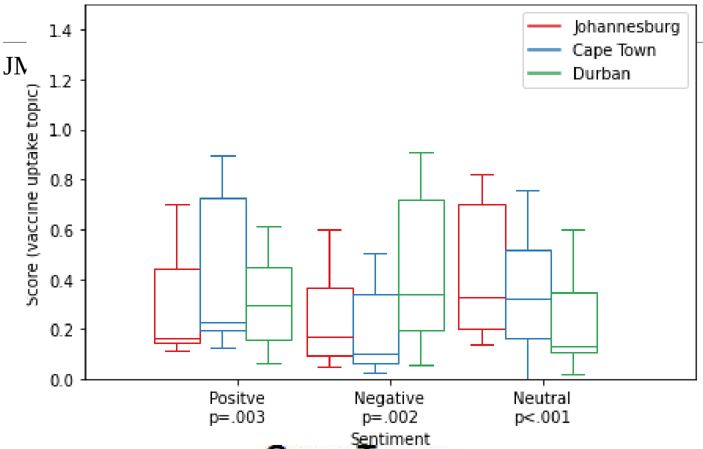
### Comparison Across Cities

Sentiment towards vaccine uptake deferred across cities (see Figure 18). Cape Town demonstrated a higher intensity score for positive sentiment than Durban and Johannesburg. Durban demonstrated a higher negative sentiment intensity score than Cape Town and Johannesburg. Similarly, Johannesburg demonstrated a higher neutral sentiment than Durban and Cape Town. There is a statistically significant neutral sentiment (p<.001) for the vaccine uptake topic across the cities than the negative sentiment (p=.002), and positive sentiment (p=.003) respectively.

**Figure 18.** Distribution and comparison of sentiment intensity scores for vaccine uptake across cities using the Kruskal-Wallis *H* test.

The Word clouds for vaccine uptake topic across the three cities are shown in Figure 19.





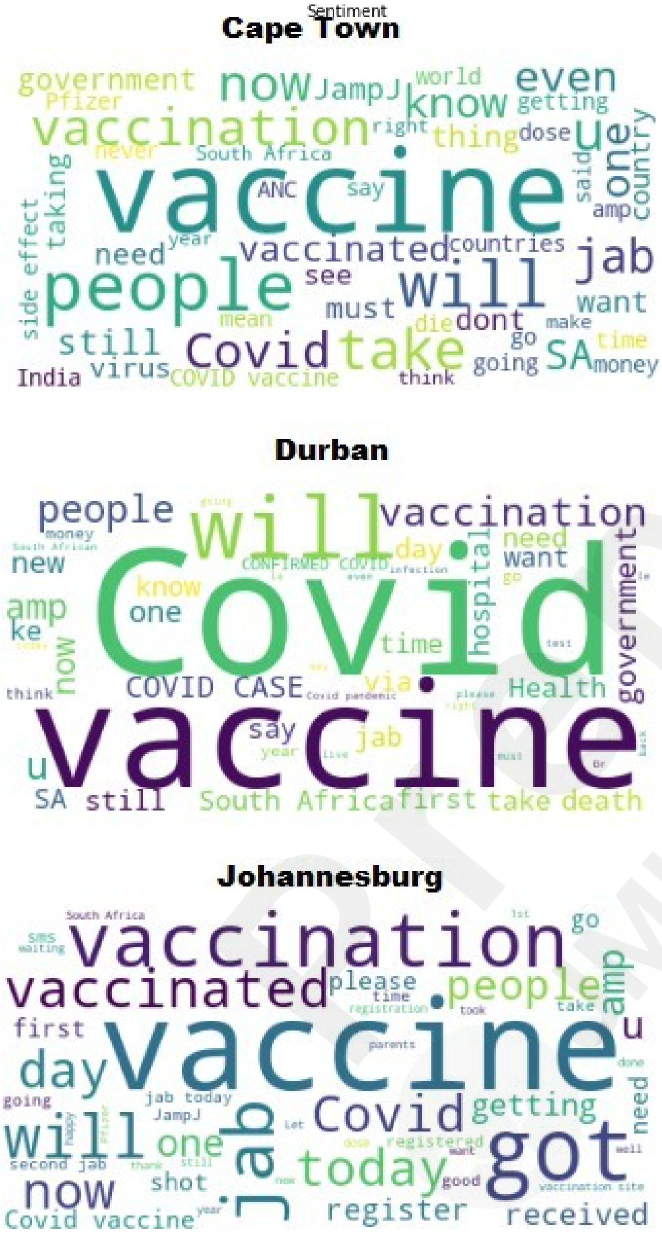


**Figure 19.** City-Level word cloud for vaccine uptake topic.

Next, we present the uncommon challenges encountered during this research.

### Limitations

The dataset used for this research only reflects the opinion of Twitter users whose geolocation was South Africa from January 2021 to August 2021. South Africa, a population of about 60 million people have only 15% online adults who use Twitter, and of the age 24 – 35 [42]. Therefore, this research does not, at large, represent the opinion of people of South Africa towards COVID-19 vaccines. However, this research only provided an insightful prediction from the dataset to support policymaking.

It is also relevant to state here that most NLTK for sentiment analysis techniques does not have the capacity to properly label figurative language, such as sarcasm. However, since the approach we used was able to label and score a large amount of the tweets in our dataset and was verified with the manual labeling of randomly selected (10%) of the tweets, in addition to the 79% accuracy achieved with the SVM classification algorithm, we assume it was able to deal with the noise generated by this obvious challenge. Finally, since our data generation ended in August 2021, the suggested area of further studies could be the generation and use of a dataset up to a recent date.

### Ethical considerations

The study was approved by Twitter and granted access to the Twitter academic researcher API which was used to retrieve the tweets. All retrieved tweets are in the public domain and are publicly available. However, the authors strictly followed the highest ethical principles in handling the personal information of Twitter users, as such, all personal information was removed.

## Discussions

### Principal Findings

We have used the Twitter API to generate and process a dataset of vaccine-related posts in South Africa from January 2021 to August 2021. We observed a decline in daily tweets and new cases between March 2021 and April 2021. This could be attributed to the effect of the xenophobic attacks [43] and the Zuma unrest [44] that took place in South Africa during this period. Our result showed that the number of new COVID-19 cases in South Africa significantly positively correlated with the number of Tweets.

The LDA topic modeling approach was used to generate topics on South Africa Twitter dataset we processed. We identified 10 topics, namely, vaccine uptake, social distancing, xenophobic attack, travel restrictions, alcohol ban, religion, sports, border closer, politics, and vaccine supply. The vaccine uptake and vaccine supply were the most dominant topics. This approach could be likened to be similar but not the same as the study in [12]. In [12], the LDA topic modeling and aspect-based sentiment analysis (ABSA) were used on Twitter data at a continental level (North America). The LDA was used to identify different topics relating to COVID–19 in the USA and Canada respectively. According to the study in [12], travel and border restrictions were the most discussed topics in February 2020 and were later overtaken by discussions about physical distancing with time. Contrary to the VADER and SVM we used in our study for sentiment analysis, ABSA was used to identify various sentiments related to the overall outbreak, anti-Asian racism and misinformation, and positive occurrences related to physical distancing.

Further, our study showed that an increase in vaccine-related discussions correlated with the number of people vaccinated for the two dominant topics we identified. We observed that the sentiment intensity score for both topics had significantly higher scores for the neutral class than the other classes. This could be attributed to the fact that a lot of people at this time may be indecisive about taking the vaccine. As a result, a lot of vaccines expired without being administered to people. This is similar to what is seen in other African countries like Nigeria, Mozambique, Zimbabwe, Botswana, Estonia, Angola, Demo- cratic Republic of Congo, etc. [45, 46] , where a lot of vaccines are said to have expired without being administered to people, despite the fact that a low percentage of their populations are vaccinated. This type of information could be helpful to public health agencies to understand public concerns of Twitter users towards vaccine hesitancy especially in communities where the acceptance rate is low.

The increase in intensity score of the negative sentiment class for the vaccine uptake topic from February 2021 to April 2021 seems to have a slight effect on the number of vaccinations, especially in April 2021. This could be the result of the rumours concerning the side effect of the vaccines [11] and the presidential address that vaccination is not compulsory at that time, as such, many people seem to be hesitant in taking the vaccine because of one reason or another [21]. In July 2021, there was a lot of continuous media and physical sensitization campaigns and awareness of the need to be vaccinated by the health agencies in South Africa [21, 22] , hence, as the number of vaccinations continued to increase, the sentiment intensity scores for the vaccine uptake topic also increased for the neutral and positive sentiment classes. In August 2021, our result appeared to behave differently. While the intensity scores for the neutral and positive sentiment classes for the vaccine supply topic decreased the number of vaccinations also decreased. Furthermore, analysis on the three selected cities, Cape Town, Durban, and Johannesburg, showed different sentiment intensity scores on the two topics within the period of discussion. This suggests that city-specific policy can be helpful in addressing the sentiment towards vaccine hesitancy.

For example, Cape Town showed a strong significant correlation of the impact of the upward increase in the intensity scores for both topics to the total vaccinations from February 2021 to August

2021. There was a weak correlation of the impact of the vaccine uptake and supply topics to new cases, deaths, and recoveries. Cape Town demonstrated a higher sentiment intensity score for both topics for neutral sentiment class than other sentiment classes.

In Durban, the impact of vaccine-related discussions on the total vaccinations showed a strong correlation for the two topics. However, the impact of vaccine-related discussions on the new cases, deaths, and recoveries demonstrated a weak correlation from February 2021 to August 2021. Addition- ally, the sentiment intensity score for both topics demonstrated higher scores for the negative sentiment than the other sentiment classes.

From February to August, Johannesburg demonstrated a strong correlation on the impact of the vaccine-related discussions to new cases, deaths, and recoveries for both topics. The sentiment intensity scores for both topics demonstrated higher scores for the neutral sentiment than the other classes.

Finally, a comparison across cities for the most trending topic, vaccine uptake, showed that Cape Town demonstrated a higher intensity score for positive sentiment, while Durban and Johannesburg demonstrated higher negative and neutral sentiments, respectively. There is a statistically significant neutral sentiment for the vaccine uptake topic across the cities. Our analysis showed that Twitters posts can be used to better understand the city–specific sentiment on vaccines related topics. Given that this approach is fast and less expensive, health policymakers could adopt this approach in monitoring citizens' responses to related policies. For example, the study in [47] showed how sentiment analysis could be used to understand public perceptions of policies in Italy. This was very helpful in the accountability and responsiveness of policymakers. Similarly, the study in [15] showed engagement on Reddit correlated with COVID-19 cases and vaccination rates in Canadian cities. This showed that discussion on social media can serve as predictors for real-world statistics.

## Conclusion

In this research, Twitter posts containing daily updates of location-based COVID–19 vaccine-related tweets were used to generate topics and understand the sentiments around the topics. Trending topics regarding the vaccine discussions were identified at local levels. The impact of the sentiment of these discussions was identified and related to the vaccinations, new cases, deaths, and recoveries at the local levels.

These go further to show that geolocation clustering of Twitter posts can be used to better analyze the sentiments towards COVID–19 vaccines at the local level. Our results, therefore, suggest that city–level analysis of sentiments about COVID–19 vaccine-related Twitter posts can be used to provide additional information to health policy and decision making regarding COVID–19 vaccine hesitancy.

## Conflicts of Interest

There is no conflict of interest.

## Acknowledgment

To The Africa-Canada AI & Data Innovation Consortium (ACADIC) team in Canada and South Africa.

## Appendix A: Keywords

1

## References

1. Food US, Administration D. Coronavirus (COVID-19) Update: FDA Authorizes First Oral Antiviral for Treatment of COVID-19; 2021. Available from: https://www.fda.gov/news-events/press-announcements/ coronavirus-covid-19-update-fda-authorizes-first-oral- antiviral-treatment-covid-19.

2. Silva J, Bratberg J, Lemay V. COVID-19 and influenza vaccine hesitancy among college students. Journal of the American Pharmacists Association. 2019;1544-3191:1–6. Available from: https://doi.org/10.1088/1757-899x/879/1/ 012116. [Accessed 2021-11-26].

3. MokhlesurRahmana M, NawazAli GGM, JunLi X, Jim- Samuel, Paul KC, Chong PHJ, et al. Socioeconomic fac- tors analysis for COVID-19 US reopening sentiment with Twitter and census data. Heliyon. 2021;7. Available from: http://www.cell.com/heliyon. [Accessed 2021-12-10].

4. World Health Organization website;. Available from: https://www.who.int/. [Accessed 2021-11-26].

5. Neha P, Eric AC, Hourmazd H, Keith G, et al. So- cial media and vaccine hesitancy: new updates for the era of COVID-19 and globalized infectious dis- eases. HUMAN VACCINES & IMMUNOTHERAPEU- TICS. 2020;16(11):2586–2593.

6. Marcec R, Likic R. Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Mod- erna COVID-19 vaccines. Postgraduate Medical Journal. 2021;p. 1–7. Available from: https://pmj.bmj.com/.

7. Cooper S, Betsch C, Sambala EZ, Mchiza N, Wiysonge CS. Vaccine hesitancy - a potential threat to the achieve- ments of vaccination programmes in Africa. Hum Vaccin Immunother. 2018;14(10):6284499–6284499.

8. Yin F, Shao X, Ji M, Wu J. Quantifying the Influence of Delay in Opinion Transmission of COVID-19 Information Propagation: Modeling Study. Journal of Medical Internet Research. 2021;23(2). Available from: https://www.jmir.org/2021/2/e25734. [Accessed 2021-12-10].

9. SAMRC. Towards understanding the complexities of vaccine hesitancy in South Africa; 2021. Available from: https://www.samrc.ac.za/news/towards-understanding- complexities-vaccine-hesitancy-south-africa. [Accessed 2021-12-10].

10. Chutel L, Fisher M. The Next Challenge to Vaccinat- ing Africa: Overcoming Skepticism; 2021. Available from: https://www.nytimes.com/2021/12/01/world/africa/ coranavirus-vaccine-hesitancy-africa.html. [Accessed 2021-12-10].

11. Tasnim S, Hossain MM, Mazumder H. Impact of Rumors and Misinformation on COVID–19 in Social Media. Jour- nal of Preventive Medicine & Public Health. 202;53:171– 174.

12. Jang H, Rempel E, Roth D, Carenini G, Janjua NZ. Track- ing COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect- Based Sentiment Analysis. Naveed Zafar Janjua2,3,4. 2021;23(2). Available from: http://www.jmir.org/2021/2/ e25431/.

13. Menezes NP, Simuzingili M, Debebe ZY, Pivodic F, Massiah E. What is driving COVID-19 vaccine hesitancy in Sub-Saharan Africa?; 2021. Avail- able from: https://blogs.worldbank.org/africacan/what- driving-covid-19-vaccine-hesitancy-sub-saharan-africa.

14. Yin F, Xia X, Song N, Zhu L, Wu J. Quantify the role of su- perspreaders -opinion leaders- on COVID-19 information propagation in the Chinese Sina-microblog. PLoS ONE. 2020;16(6):1–20. Available from: https://doi.org/10.1371/ journal.pone.0234023.[Accessed 2021-12-10].

15. Yan C, Law M, Nguyen S, Cheung J, Kong J. Com- paring Public Sentiment Toward COVID-19 Vaccines Across Canadian Cities: Analysis of Comments on Red- dit. JOURNAL OF MEDICAL INTERNET RESEARCH. 2021;23(9). Available from: https://www.jmir.org/2021/9/ e32685. [Accessed 2021-12-10].

16. Nia ZM, Asgary A, Bragazzi N, Melado B, Orbinski J, Wu J, et al. Tracing Unemployment Rate of South Africa dur- ing the COVID-19 Pandemic Using Twitter Data. Jour- nal of Medical Internet Research;Available from: https: //preprints.jmir.org/preprint/33843.

17. Yin F, Pang H, Xia X, Shao X, Wu J. COVID-19 information contact and participation analysis and dynamic prediction in the Chinese Sina-microblog. Physica A. 2021;570:7845521–7845521.

18. Su Y, Venkat A, Yadav Y, Puglisi LB, Fodeh SJ. Twitter- based analysis reveals differential COVID-19 concerns across areas with socioeconomic disparities. Com- puters in Biology and Medicine. 2021;132(104336). Available from: http://www.elsevier.com/locate/ compbiomed. [Accessed 2021-12-10].

19. Yin F, Shao X, Tang B, Xia X, Wu J. Modeling and analyzing cross-transmission dynamics of re- lated information co-propagationModeling and analyzing cross-transmission dynamics of related information co- propagation. Scientific Reports. 2021;11(268). Avail- able from: https://www.nature.com/articles/s41598-020-79503-8#citeas. [Accessed 2021-12-10].

20. Piedrahita-Valdés H, Piedrahita-Castillo D, Bermejo- Higuera J, Guillem-Saiz P, Bermejo-Higuera JR, Guillem- Saiz J, et al. Vaccine Hesitancy on Social Media: Senti- ment Analysis from June 2011 to April 2019. Vaccines. 2021;9(28):1–12. Available from: https://www.mdpi.com/ 2076-393X/9/1/28#cite. [Accessed 2021-12-10].

21. SAcoronavirus. COVID-19 Online Resources & New Por- tal; 2021. Available from: https://sacoronavirus.co.za/ . [Accessed 2021-11-05].

22. COVID-19 South Africa Dashboard; 2020. Avail- able from: https://www.covid19sa.org/.

[Accessed 2021- 11-09].

23. Bounding Box; 2020. Accessed [2021-11-09]. Available from: https://wiki.openstreetmap.org/wiki/Bounding_ Box. [Accessed 2021-12-10].

24. Li F, Van den Bossche J, Zeitlin M, MeeseeksMachine, Team PD, Hawkins S, et al.. What's new in 1.2.4 (April 12, 2021); 2021. Available from: https://pandas.pydata.org/pandas-docs/stable/whatsnew/ v1.2.4.html, [Accessed 2021-11-08].

25. tweet-preprocessor 0.6.0; 2020. Available from: https://pypi.org/project/tweet-preprocessor/. [Accessed 2021-12-10].

26. Natural Language Toolkit; 2021. Available from: https: //www.nltk.org/. [Accessed 2021-12-10].

27. Honnibal M. spaCy 2: Natural language understand- ing with Bloom embeddings, convolutional neural net- works and incremental parsing. Sentometrics Research. 2017;1(1). Available from: https://sentometrics-research. com/ publication/72/. [Accessed 2021-12-10].

28. Industrial Strenght Natural Language Processing in Python; 2021. Available from: https://spacy.io/. [Accessed 2021-12-10].

29. Aditya B. Sentimental Analysis Using; 2020. Avail- able from: https://towardsdatascience.com/sentimental- analysis-using-vader-a3415fef7664. [Accessed 2021-12-10].

30. Langkilde D. Linear SVM classification of senti- ment in tweets; 2016. from: https://www.kaggle.com/ langkilde/linear-svm-classification- of-sentiment-in-tweets. [Accessed 2021-12-10].

31. Vasista R. Sentiment Analysis using SVM; 2018. Avail- able from: https://medium.com/@vasista/sentiment- analysis-using-svm-338d418e3ff1. [Accessed 2021-12- 10].

32. Glen S. "RMSE: Root Mean Square Error" From Statis- ticsHowTo.com: Elementary Statistics for the rest of us! ; 2021. Available from: https://www.statisticshowto.com/ probability-and-statistics/regression-analysis/rmse-root- mean-square-error/. [Accessed 2021-12-10].

33.Blei DM, Ng AY, Jordan MI. Latent Dirichlet Alloca- tion. Journal of Machine Learning Research. 2003;3:993– 1022. Available from: https://www.jmlr.org/papers/ volume3/blei03a/blei03a.pdf. [Accessed 2021-12-10].

34.Rˇ ehu˚ˇrek R. Gensim Topic Modelling for Humans; 2009. Available from: https://radimrehurek.com/gensim/intro. html. [Accessed 2021-12-10].

35.Kanani B. Jaccard Similarity – Text Similar- ity Metric in NLP; 2020. Available from: https://studymachinelearning.com/jaccard-similarity-

36.text-similarity-metric-in-nlp/. [Accessed 2021-12-10]. Aletras N, Stevenson M. In: Proceedings of the 10th International Conference on Computational Semantics, 01/03/2013; 2013. p. 13–22. Available from: https:// aclanthology.org/W13-0102.pdf. [Accessed 2021-12-10].

37.Kapadia S. Evaluate Topic Models: Latent Dirich- let Allocation (LDA); 2019. Available from: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0. [Accessed 2021-12-15].

38.Li S. A Quick Introduction On Granger Causality Available Test- ing For Time Series Analysis; 2020. Available from: https://towardsdatascience.com/a-quick-introduction on-granger-causality-testing-for-time-series-analysis- 7113dc9420d2. [Accessed 2021-12-10].

39.Bedre R. Mann-Whitney U test (Wilcoxon rank sum test) in Python [pandas and SciPy]; 2021. Available from: https://www.reneshbedre.com/blog/mann-whitney- u-test.html. [Accessed 2021-12-15].

40. Bedre R. Kruskal-Wallis test in R [with exam- ple and code]; 2021. Available from: https://www.reneshbedre.com/blog/kruskal-wallis-test.html. [Accessed 2021-12-15].

41. Statista. Largest cities in South Africa in 2021, by number of inhabitants; 2021. Available from: https://www.statista.com/statistics/1127496/largest-cities- in-south-africa/. [Accessed 2021-12-10].

42. Writer S. The biggest and most popular social media plat- forms in South Africa, including TikTok.; 2021. Available from: https://businesstech.co.za/news/internet/502583/ the-biggest-and-most-popular-social-media-platforms-in-south-africa-including-tiktok/. [Accessed 2021-12- 10].

43. Isilow H. Refugees in South Africa still live in fear of xenophobic attacks; 2021. Available from: https://www.aa.com.tr/en/life/refugees-in-south-africa-still-live-in-fear-of-xenophobic-attacks/2280537. [Accessed 2021-01-11]

44. 2021 South African unrest; 2021. Available from: https://en.wikipedia.org/wiki/2021_South_African_unrest. [Accessed 2021-01-11].

45. Times G. Destruction of expired COVID-19 vaccines in Africa a shame for the West: Global Times editorial; 2021. Available from: https://www.globaltimes.cn/page/202112/1243364.shtml. [Accessed 2021-01-11].

46. Mlaba K. Why Are African Countries Throwing Away COVID-19 Vaccines?; 2021. Available from: https://www.globalcitizen.org/en/content/african-countries-throwing-away-covid-19-vaccines/. [Accessed 2021-01-11]

47. A C, F N. Public policy and social media: How sentiment analysis can support policy-makers across the policy cycle. Riv Ital Politiche Pubbliche. 2015;10(3):309–338.