The Impact of Omicron Variant in Vaccine Uptake in South Africa

Blessing Ogbuokiri (✉ blessogb@yorku.ca)
York University

Ali Ahmadi
K.N.Toosi University of Technology

Nidhi Tripathi
University of the Witwatersrand

Zahra Movahedi
York University

Bruce Melado
University of the Witwatersrand

Jianhong Wu
York University

Ali Asgary
York University

James Orbinski
York University

Jude Kong
York University

Article

Keywords: COVID−19, Emotion Analysis, Health Policymaking, Machine Learning, Natural Language Processing, Omicron, Vaccine Hesitanc

Posted Date: January 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2454619/v1

Additional Declarations: No competing interests reported.

The Impact of Omicron Variant in Vaccine Uptake in South Africa

Blessing Ogbuokiri[1,4,*], **Ali Ahmadi**[2], **Nidhi Tripathi**[3], **Zahra Movahedi**[1,4], **Bruce Mellado**[1,3], **Jiahong Wu**[1,4], **Ali Asgary**[1,5], **James Orbinski**[1,6], **and Jude Kong**[1,4]

[1]Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Department of Mathematics and Statistics, York University, Toronto, Ontario, M3J 1P3, Canada.
[2]K.N. Toosi University, Faculty of Computer Engineering, Tehran, Iran.
[3]School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, 2000, Gauteng, South Africa.
[4]Laboratory for Industrial and Applied mathematics, York University, Toronto, Ontario, M3J 1P3, Canada.
[5]Advanced Disaster, Emergency and Rapid-response Simulation (ADERSIM), York University, Toronto, Ontario, M3J 1P3, Canada.
[6]Dahdaleh Institute for Global Health Research, York University, Toronto, Ontario, M3J 1P3, Canada.
[*]Corresponding: blessogb@yorku.ca

ABSTRACT

The first report of the Omicron variant triggered a lot of emotions towards vaccination in South Africa. These emotions are mostly expressed on social media such as Twitter. The result could weaken the confidence level of users even before they are vaccinated. Identifying these emotions and how they change before and during the Omicron variant can help, in understanding the dynamics in citizens' behaviour, towards vaccination for health policy-making. In this study, 23,000 vaccine-related Twitter posts were collected in South Africa, from 1 October 2021 to 15 January 2022 using Natural Language Processing techniques. The emotional classification of Twitter posts and their associated intensities were achieved using the Text2emotion pre-trained model. The results were validated using Naive Bayes with an accuracy of 76%, Logistic Regression (91%), Support Vector Machines (84%), Decision Tree (82%), and K-Nearest Neighbours (72%). The number of tweets significantly positively correlated with the increase in vaccination across all South African provinces (Corr≤0.532, P≤0.003) except Northern Cape province. The emotional intensities for vaccine-related posts showed a strong association with the increase in vaccination during Omicron (P<0.04) in Eastern Cape, Gauteng, Limpopo, and North West provinces than other provinces. The comparison of the intensities of the emotional classes differed across provinces before and during Omicron. The result of this research showed that, social media data can be used to complement existing data, in understanding and predicting the dynamics in citizens' emotional behaviour towards vaccination, during a new COVID–19 variant or future outbreaks. The result could also inform health policy in planning, control, and management of provinces identified with vaccine hesitancy. It can also serve as a template or reference for related future academic research.

Keywords: COVID–19, Emotion Analysis, Health Policymaking, Machine Learning, Natural Language Processing, Omicron, Vaccine Hesitancy.

1 Introduction

A major characteristic of the SARS-CoV-2 virus (COVID-19) is that it changes constantly[1,2]. Since the first report of the COVID-19 pandemic in December 2019 in Wuhan, China, there have been several variations of the virus[2]. The variations include Alpha variant, Beta variant, Delta variant, and Omicron variant (whose most common lineages are BA.1, BA.1.1, and BA.2). The virus keeps mutating to even a more dangerous and transmissible disease[3–5]. The XE variant, a combination of the two highly transmissible strains of Omicron was reported in the United Kingdom on January 19, 2022[6–9]. As of January 07, 2023, the XBB.1.5, a recent subvariant of Omicron has become a variant of concern as it has shot up like a rocket in the United States. The good news is that there is no evidence the XBB.1.5 Omicron subvariant makes people sicker than earlier versions[10].

On November 24, 2021, the World Health Organisation (WHO) classified a variation of COVID–19 (B.1.1.529) as a variant of concern because of its high transmissibility and named it Omicron[10,11]. Although there were rumours of the virus in Botswana, Nigeria, Zimbabwe, and some part of Europe. The Omicron was first reported by South African scientists in Tshwane municipality in the Gauteng province and it has been tracked in more than 120 countries including the United States of America and Canada[11–13].

Omicron is said to be highly transmissible than Delta variant[14]. For example, from the earlier reports of the variant, within a period of two weeks the Omicron cases in South Africa rose from 300 a day to 3000 a day[14,15]. This may be attributed to be the reason the virus was said to be the most dominant variant in a short while in most countries[16,17]. There are over 30 mutations in Omicron's spike protein which can attack the human cells with the ability to increase the probability of infection[2]. Given the high transmissibility nature of Omicron, it can bypass the human immune responses, especially in people who are not vaccinated but were infected in the past[3,4].

Of course, these information about the nature and transmissibility of Omicron variant dominated the news and social media space at the early report in South Africa. Social Media was the common platform for interaction among users or groups to share information about Omicron while observing COVID–19 restrictions[18]. However, it became a thing of concern, given that information shared on social media do not necessarily go through a thorough editorial supervision. These could be a rumour from unverified sources[18,19]. The danger in this is that, rumours could become the reality of people who are not well informed about the subject of discuss[18,20]. Twitter, Facebook, and Whats-App were the popular social media platforms used to share news, information, opinions, and emotions about Omicron related discussions[15,21,22]. Since we used Twitter data for this research, we focus the discussion on Twitter alone.

Moreover, information sharing and discussions about the Omicron variant on Twitter triggered a lot of emotions towards vaccination in South Africa[18]. This is because, the Omicron variant became the subject of discussion at that time and it formed most of the trending topics about vaccination[20]. By trending topic, we refer to a word or phrase that is mentioned at a greater rate within a limited duration of time on a social media platform, such as Twitter[23,24]. The emotions expressed on these Twitter posts about the Omicron, do not necessarily pass through editing oversight before they are posted to the public domain[25]. The essence is to be able to manage the tone of the language of the tweet in order to appeal to the users or followers fancy[26]. The implication of not editing a tweet before it is tweeted is that, they may not necessary dunce the tension of followers at that time, rather, they may heighten or increase the tension. Uninformed users tend to believe everything they read from Twitter. Especially, if the post originated from a verified Twitter user or a prominent person in the society[26,27]. Users do not necessarily verify the sources of the tweets. This may affect some ardent follower psychologically, which could result to emotional imbalance for the user if not properly managed. This could affect the users trust and interest towards accepting vaccination.

According to the American Psychological Association (APA), emotion is defined as *"a complex reaction pattern, involving experiential, behavioural and physiological elements."* Emotion is a behaviour that reflects a person's opinion regarding the interaction between two or more persons in relation to an event[28–30]. In recent times, emotion detection in text has become very popular due to its notable relevance in application areas like marketing, artificial intelligence, psychology, etc. The availability of large amount of opinionated and self expression textual data have also played a major role. Although Armin, et al[31], argued that techniques and methodologies for text detection needs improvement. Most of them lack insufficient capacity to handle complexity in human emotions, such as the use of metaphorical language expressed on text[31,32]. According to Haji, et al[33], a hybrid based architecture was used to detect emotion from text. The system recorded a prediction accuracy of 96.43% when validated with Support Vector Machine classification algorithm.

However, modelling emotionality and neutrality of words using a generative unigram mixture model (UMM) shows significantly lower perplexity than traditional supervised Latent Dirichlet Allocation (sLDA) models[34]. Emotion from text could be detected at the sentence level, by computing an emotion vector word. The semantic relationship between words and their different emotions are calculated. The scores are normalised using the syntactic dependencies within the sentence structure[35]. Challenges in sentiment analysis and emotion detection from text includes spelling mistakes, new slang, and incorrect use of grammar[36]. These challenges make it difficult for a machine to detect if a text expresses anger or worry. Researchers are still working on ways to improve on the existing systems in order to optimise their performance. Given that it is nearly impossible to identify all the variations of text representing all possible human emotions[37].

There are different emotional classes that has been identified by Kandra in[38]. The human beings can identify and understand the emotional classes expressed on text, but machines cannot, except if they are trained to do so. The Text2emotion, a Natural Language Processing (NLP) technique has been trained to identify five basic classes of emotions from text[30]. The five emotional classes are happiness, sadness, anger, surprise, and fear. We focus on these five basic emotional classes because of the nature of this study. Given a large amount of text-based data from Twitter, by applying the appropriate tool, such as Text2emotion, the emotions on each Twitter post (tweets) could be extracted. The way users receive, process, and understand the discussions relating to Omicron from Twitter could influence or impact their emotions towards vaccination.

In respect to the above, we performed NLP techniques on Twitter posts from October 1, 2021, to January 15, 2022. We identified and compared the emotional changes of users towards COVID–19 vaccine-related tweets before and during Omicron, across nine(9) provinces in South Africa. The before Omicron, refers to October 1 – November 22, 2021. While during Omicron refers to November 23, 2021 – January 15, 2022. We correlated the intensities of these emotions to vaccination, to understand the dynamics of emotional changes and its effects to vaccination, to identify provinces or locations that may likely experience COVID–19 vaccine hesitancy, based on the data available, for health policy-making.

Specifically, this study addressed the following problems:

- Identified and classified emotions in COVID–19 related discussions from Twitter posts before and during Omicron.
- Identified the relationship between emotional changes in COVID–19 vaccine related discussions on Twitter and vaccine uptake before Omicron across different provinces in South Africa.
- Identified the relationship between emotional changes in COVID-19 vaccine related discussions on Twitter and the increase in vaccine uptake during Omicron across different provinces in South Africa.
- Compared different emotional changes in COVID–19 vaccine related discussions on Twitter before and during Omicron across different provinces in South Africa.

The approach used in this research shows that social media data from Twitter can be used to complement existing vaccination data, in understanding and predicting the dynamics in citizens' emotional behaviour towards vaccination, before or during an outbreak of a new COVID–19 variant or future outbreaks. Also, to inform health policy in planning, control, and management of provinces identified with vaccine hesitancy or locations that may likely experience vaccine hesitancy for a timely response. It can also serve as a template or reference for related academic research.

The remaining parts of this manuscript is presented in five sections. Section 2 discussed different approaches used in collecting, preparing, and analysing data. Section 3 presents the outcome of the analysis of the impact of Omicron to vaccination before and during Omicron. It also compares the emotional intensities across provinces. In Section 4, the results in connection with other similar works that are relevant to this study were discussed. Finally, Section 5 we summarised the outcome of study and concludes the manuscript.

2 Methods

2.1 Data Collection

Twitter data was collected using the approved academic researcher Twitter API. The Twitter API allows retrieval of up to 10 million historical tweets per month. We used Python version 3.6 with the Twitter access token to generate scripts that was used to collect user historical tweets from the Twitter database. The historical tweets contain specific keywords concerning COVID–19 vaccines. The keywords were selected from the COVID–19 Twitter trending topics in South Africa, from October 1, 2021, to January 15, 2022. They include `omicron`, `vaccine`, `sarscov2`, `covid19`, `covid-19sa`, `coronavirus`, `coronavirus sa`, and `corona`. Others are `sars-cov-2`, `lockdown`, `covid-19`, and `pandemic`. English was used as the preferred language of the tweets.

We collected a total of 23,000 tweets using the archive search process. Each Tweet contains most of the following features described in Table 1.

Additionally, daily statistics of COVID–19 vaccinations before Omicron and increase in vaccination during Omicron, in all provinces were calculated. The South African coronavirus official website[39] provided us these data for the daily statistics together with the South Africa COVID–19 dashboard[40]. These websites are primarily updated every day.

2.2 Data Preprocessing

User tweets are usually unorganised in their raw form. They contain a lot of redundant information. In order to make sense of the data, we processed the data to retain the information relevant to our research. We extracted tweets, date created, time created, and provinces from the dataset into a dataframe using Pandas version 1.2.4[41]. We prepared the data for NLP use by removing non–English words, tweets with incomplete information, URLs, punctuations and quotations, duplicate tweets, Stopwords, and special or non–alphabetical characters. This process was achieved using tweets-preprocessor 0.6.0[42] and NLTK 3.6.2[43]. The Spacy 2 3.2 from the Spacy2 toolkit was used to tokenize the tweets[44,45]. In the end, the tweets were reduced to 20,766 after preprocessing.

2.3 Emotion Analysis

Emotion is the state of mind that is related to how human beings feel and think about a specific object. It is a behaviour that reflects a person's opinion in relation to the interaction between two or more persons about a certain event[30]. Emotion can be expressed physically or textually. The physical expression of emotions can be easily understood by a close observation on one's facial expression. Also, human beings can recognise and identify emotions from textual data and relate same to matter of the text. Machines are not able to recognise emotions from text unless they are trained to do so[30,38].

Text2emotion version 0.0.5[46] was used to extract emotions from the Twitter data (tweets) we generated and processed. Text2emotion is a python packaged developed to identify the correct emotion expressed on text data. It can recognise five

Table 1. Dataset features

SN	Attribute	Description	Type
1	TweetText	Twitter post that represents users' opinions	Text
2	TweetID	Unique identity of a tweet	Numeric
3	CreateDate	The date when the tweet was posted	Date
4	RetweetCount	Number of re-post a tweet gets	Numeric
5	ReplyCount	A processed post that represents users' opinions	Text
6	LikeCount	Number of users interested or agreed to the tweet	Numeric
7	GeoId	Unique identifier of the the region or location of the the tweet	alphanumeric
8	GeoCityProvince	The city and province the tweet originated	Text
9	GeoCountry	The country of origin of he tweet	Text
10	GeoCoordinate(bbox)	The area of the tweet defined by two longitudes and two latitudes	Numeric
11	AuthorID	Unique number that identifies a user	Text
12	UserName	An identification used by a user with access to Twitter.	Alphanumeric
13	Hashtags	Metadata tag that is prefaced by the pound symbol	Alphanumeric
14	CreatedAccountAt	Time and date user account was created	Datetime
15	FollowerCount	Total number of users' followers	numeric
16	FollowingCount	Total number of users a user is following	numeric
17	TweetCount	Total number of tweets	Numeric

different emotion categories as Happy, Angry, Sadness, Surprise and Fear from text data. Figure 1 shows how Text2emotion classifies text based on Happy, Angry, Sadness, Surprise and Fear categorises and assigns intensity to each emotion category.

The **Tokenized word** in Figure 1 is where sentences are broken down into words or pair of words. For instance, the sentence:

```
"I am bitter, agonized, and confused because Covid-19 infected and killed by
                                   cousin"
```

when tokenized becomes

```
["Bitter", "agonized", "confused", "infected", "killed"].
```

The tokenized words are then passed to the emotion list.

The **Emotion list** contains a dictionary of the five emotional categories: Happy, Angry, Sadness, Surprise, and Fear, including the words that describe them. A sample emotion list is presented in Figure 2. When a word from the tokenized word list is contained in the emotion list of words, the **Emotion found** gains a count of one. For instance, the count of words in the emotional categories of the sentence whose tokenized words are `["Bitter", "agonized", "confused", "infected", "killed"]` will be: Happy = 0, Sadness = 1, Anger = 2, Surprise= 1, Fear = 1.

The value of each emotional category determines the emotional intensity. The emotional intensity in this context is the impact of a particular emotional class in a sentence. The emotional intensity can be calculated using the formula in Equation (1). The score of the emotional intensity of any emotional class is normalised between 0 lowest intensity to 1 highest intensity.

$$\text{Emotional Intensity} = \frac{\text{Count of emotional class}}{\text{Total count of emotional class}}. \qquad (1)$$

For instance, the intensity for each of the emotional class will be calculated using Equation 1 as thus:

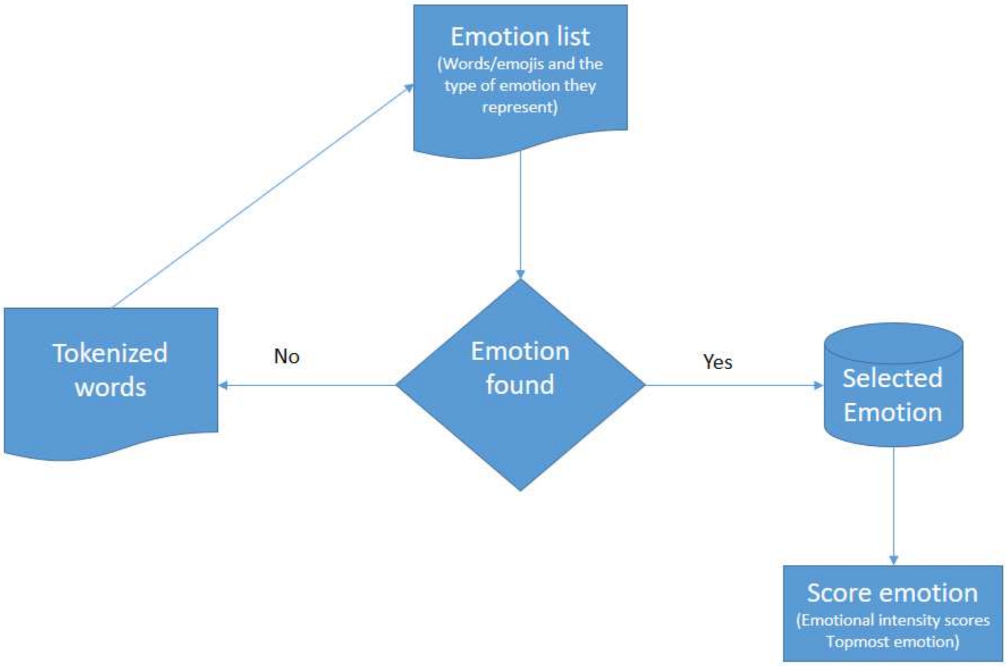

Emotion list
(Words/emojis and the type of emotion they represent)
Tokenized words
No
Emotion found
Yes
Selected Emotion
Score emotion
(Emotional intensity scores Topmost emotion)

Figure 1. Text emotion classification using Text2emotion.

```
{Happy: happy, amused, love, attracted, like}
{Sadness: sad, hate, unhappy, grieved, infected}
{Anger: angry, agonized, bitter, appalled}
{Surprise: surprise, confused}
{Fear: fear, fearful, death, killed}
```

Figure 2. Sample emotion list.

$$\text{Happy: } \frac{0}{0+1+2+1+1} = \frac{0}{5} = 0,$$

$$\text{Sadness: } \frac{1}{0+1+2+1+1} = \frac{1}{5} = 0.2,$$

$$\text{Anger: } \frac{2}{0+1+2+1+1} = \frac{2}{5} = 0.4,$$

$$\text{Surprise: } \frac{1}{0+1+2+1+1} = \frac{1}{5} = 0.2,$$

$$\text{Fear: } \frac{1}{0+1+2+1+1} = \frac{1}{5} = 0.2.$$

The final output will be represented in a dictionary as follows:

$$\{\text{Happy: 0.0, Sadness: 0.2, Anger: 0.4, Surprise: 0.2, Fear: 0.2}\}.$$

The emotional class with the highest emotional score becomes the dominance emotion of the sentence. Therefore, the sentence

`"I am bitter, agonized, and confused because COVID-19 infected and killed by cousin"` is labelled as `Anger` because the anger emotional class carries the highest emotional score.

2.4 Tweet Labelling

The tweets were labeled according to their corresponding emotional class using the approach in Section 2.3. A distribution of the tweets labelled using Text2emotion is shown in Figure 3.

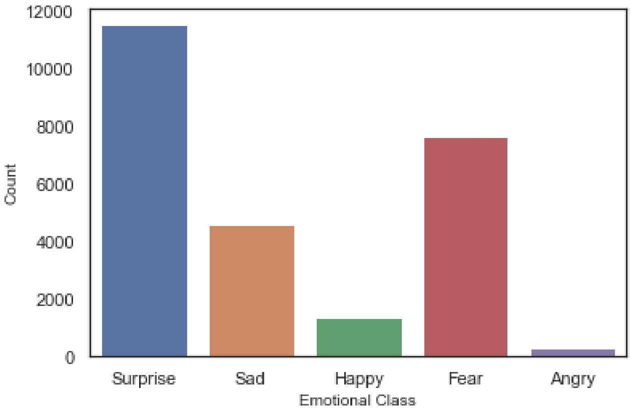


Figure 3. Distribution of emotional class.

The emotional class of every tweet was assigned a corresponding emotion score. The outputs of the labelling were validated using Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVMs), Decision Tree (DT), and K-Nearest Neighbours (KNN) machine learning classification algorithms. These classification algorithms were used because they have been successfully used in similar works, such as[47] and[48]. The outcome of this process was saved into a dataframe. Further explanation is presented in the result section of this manuscript.

2.5 Classification Algorithms

2.5.1 Naive Bayes

We used the *MultinomialNB* Naive Bayes machine learning algorithm to classify the tweet data. Here, the distribution of tweet data is parameterized into vectors $\theta_y = (\theta_{y1}, \ldots, \theta_{yn})$ for each class y (positive, negative, neutral), where n represents the number of features in tweet classification or the size of the vocabulary and θ_{yi} is the probability $P(x_i|y)$ of feature i appearing in a sample belonging to class y. Hence, the parameters θ_y is estimated by a smoothed version of maximum likelihood. The equation is adapted from Yang[49] is given as:

$$\tilde{\theta_{yi}} = \frac{N_{yi} + \alpha}{N_y + \alpha_n} \tag{2}$$

where $N_{yi} = \sum_{x \in T} x_i$ represents the number of times feature i appears in tweet dataset of class y in the training set T. While $N_y = \sum_{i=1}^{n} N_{yi}$ represents the total count of all features for class y. If $\alpha \geq 0$, then the features may be missing in the learning sample. This prevents a zero probability in the further computations. A Laplace smoothing is when $\alpha = 1$ and Lidstone smoothing is $\alpha < 1$[49].

2.5.2 Logistic regression

The Logistic regression (LR) algorithm was used to create a model that classifies tweets as positive, negative , or neutral. It predicts the probability of a tweet being positive, negative or neutral. The formula is adapted from Obaido et al[50] and it is given as thus:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_m x_m \tag{3}$$

where π is the likelihood of a particular tweet to be classified as positive, negative or neutral, β_i represents the regression coefficients, and x_i denotes the predictor variables[50].

2.5.3 Support Vector Machines

We created a classify using the SVMs to classify the tweets. The model separate the negative tweets and positive tweets in the dataset. The empirical risk minimization method was employed by computing a decision boundary. Given a training set of n linearly separable samples and a feature vector x with dimensions d, for a dual optimization problem where $\alpha \ \varepsilon \ R^n$ and $y \in \{1, -1\}$, the SVMs solution adapted from Obaido et al[50] can be minimized using:

$$\underset{a}{maximize} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j (X_i^T, X_j) \tag{4}$$

$$\text{subject to } \alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{5}$$

The SVMs has the ability to separate a linear dataset using one hyperplane in binary classification problems. However, for handling nonlinear datasets with a multiclass target variable, kernel functions are employed to project the data to a higher dimension space where it can be linearly separable by a single hyperplane[50]. The radial basis function (RBF) was used as the kernel functions.

2.5.4 Decision tree

We used the classification and regression tree (CART) algorithm to classify the tweets. The CART algorithm uses the Gini index measure in building the decision tree. The Gini index measure is an attribute selection measure used to split data, assign weight and rank to the various features[50]. Given a tweet dataset with J labels, for all $i \in 1, 2, \ldots, J$, the Gini index is calculated using:

$$Gini = 1 - \sum_{i=1}^{J} P_i^2 \tag{6}$$

where P_i is the probability that a tweet could be classified as positive, negative, or neutral.

2.5.5 K-Nearest Neighbours

We used the KNN to identify a predefined number of training samples from the Twitter data that is closest in distance to the new point, and predict the label from these. We defined the number of samples to a constant during training, this approach is called k-nearest neighbor learning. The distance can, in general, be any metric measure: standard Euclidean distance is the most common choice. We applied the neighbourhood component analysis on our data as adapted from Obaido et al[50]. To maximise the sum of the over all samples i of the probability p_i that i is correctly classified, thus:

$$argmax_L = \sum_{i=0}^{N-1} p_i \tag{7}$$

where $N = n_{samples}$ and p_1 represents the probability of sample i being correctly classified according to a stochastic nearest neighbors rule in the learned embedded space:

$$p_i = \sum_{j \in C_i}^{N-1} p_{ij} \tag{8}$$

where C_i is the set of points in the same class as sample i, and p_{ij} is the softmax over Euclidean distances in the embedded space:

$$p_{ij} = \frac{exp(-||Lx_i - Lx_j||)^2}{\sum_k exp - (||Lx_i - Lx_k||^2)}, p_{ii} = 0 \tag{9}$$

2.6 Model User Hyperparameter Selection

By tuning the user defined parameters, model performance could be optimised for better result. The user defined parameters for each classifier is presented in this part. There are other parameters that are not stated here, because their default values used from the sklearn package. Each classifier's paramater was taken and defined from the sklearn package in Python, see Table 2.

Table 2. Description of machine learning User-defined hyperparameters used in the models.

Model	Hyperparameter	Description
NB	alpha=1.0	Controls the form of the model
LR	multi_class=’rbf’	Decision function for multi classification
	C=3.0	Strength of inverse regularisation
	kernel = ’rbf’	Specifies the type of kernel used.
SVM	C = 1.0	Regularization parameter.
	random_state = 42	Control random number for data shuffling.
DT	max_depth=32	Maximum depth of tree.
KNN	n_neighbors=7	Maximum data point in a neighbourhood.

2.7 Test Statistic

A comparison of the impact of the intensities of vaccine-related tweets to vaccination was calculated using the `granger causality tests` package in python[51,52]. The focus was on people living in South Africa whose ages are between 18 years to 34 years. The correlation of the increase in tweets and increase in vaccination across all provinces in South Africa was calculated using the `scipy.stats.pearsonr` package in python[53]. Further, a comparison of the intensities of different emotional classes was conducted using the `Mann-WhitneyU` test of independent groups, from the `scipy.stats.mannwhitneyu` package in python[54].

3 Results

We present the results in two parts. The first part deals with the classification of tweets according to their corresponding emotional classes using machine learning. Secondly, an analysis of the impact of COVID–19 related discussions on Twitter to vaccination by province before and during Omicron is conducted.

3.1 Tweet Emotion Classification

The tweets data labelled with Text2Emotion pre-trained model was fitted into Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVMs), Decision Tree (DT), and K-Nearest Neighbour (KNN) classification algorithms. The summary of the performance of these classification algorithms is shown in Table 3.

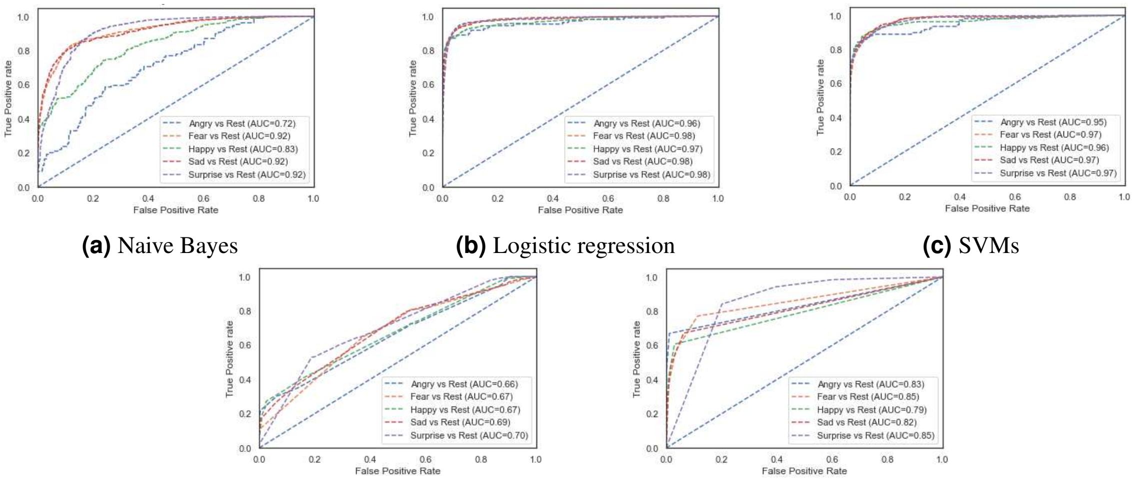
Table 3. Performance of tweet emotion classification models.

SN	Algorithm	Accuracy (%)	Average F1-Score (%)	Average AUC (%)
1	NB	0.76	0.55	0.86
2	LR	0.91	0.85	0.97
3	SVMs	0.84	0.77	0.96
4	DT	0.82	0.78	0.68
5	KNN	0.72	0.64	0.83

While there is a clear difference in the accuracy scores of the models. It was observed that the Logistic Regression model performed better than other models with accuracy score of 0.91, average F1-score of 0.85, and average AUC score of 0.97. Similarly, the SVM has an average AUC score of 0.96, slightly lower than the average AUC score for Logistic Regression model. The accuracy score of 0.84 for SVM is also slightly higher than the accuracy score of 0.82 for Decision Tree. Meanwhile, the average F1-score of 0.78 for DT is slightly higher than the mean F1-score of 0.77 for SVM. Although NB and KNN have the lowest accuracy scores and lowest average F1-scores. Both maintained a 0.86 and 0.83 average AUC scores respectively, higher than KNN with an average AUC score of 0.68.

Given the analysis above, it is clear that these models can classify tweets according to the type of emotion expressed in them. Interestingly, Logistic Regression model proved to be more suitable for this type of classification problem given all the measures. One such measure is that the large feature set generated from the the 20,766 tweets appears to be more suitable for the Logistic Regression higher performance.

Further, we visualised the Receiver Operating Characteristic (ROC) metric to evaluate the quality of the multi classification output, together with the Area Under the curve (AUC). This is to validate the performance of the models. See Figure 4.



(a) Naive Bayes

(b) Logistic regression

(c) SVMs

(d) Decision Tree

(e) KNN

Figure 4. ROC–AUC for the models used to classify tweet emotions.

The ROC curve shows the true positive rate against the true negative rate. The true positive rate is the probability of the model to accurately predict the properly labelled emotions from the tweets. While the true negative rate is the probability of the models to accurately predict the mislabelled emotions from the tweets. This is to ascertain how well the models are classifying each class.

The more the curve aligns towards the upper left corner of the plot, the better the model performs in the classification of the tweets into various emotional classes. As shown in Figure 4b, the Logistic Regression model performed well in the classification of the tweets into various emotional classes, followed by SVMs model, see Figure 4c. Unlike the NB model with an average performance in the classification of the tweet emotions, see Figure 4a, the DT and KNN models performed poorly in the classification of the tweets into various emotional classes, see Figures 4d and 4e respectively. The AUC was used to ascertain how much of the plot is located under the curve. If the AUC score is closer to 1, we assume that the model has performed well. We can conclude that LR model performed well with a large feature set and multiclass prediction.

Given that the Logistic Regression model performed better than other models, we therefore, analyse the features that may have influenced the emotion classification of the tweets. The ELI5[55] interpretable machine learning tool was used to visualise the top ten features of the Logistic Regression model in their order of importance. Table 4 shows the weight and features of the top ten features that influenced the emotional classes of the tweets as classified by the logistic regression model.

Next, we examine the discussions about COVID–19 on Twitter before and during the Omicron outbreak in South Africa.

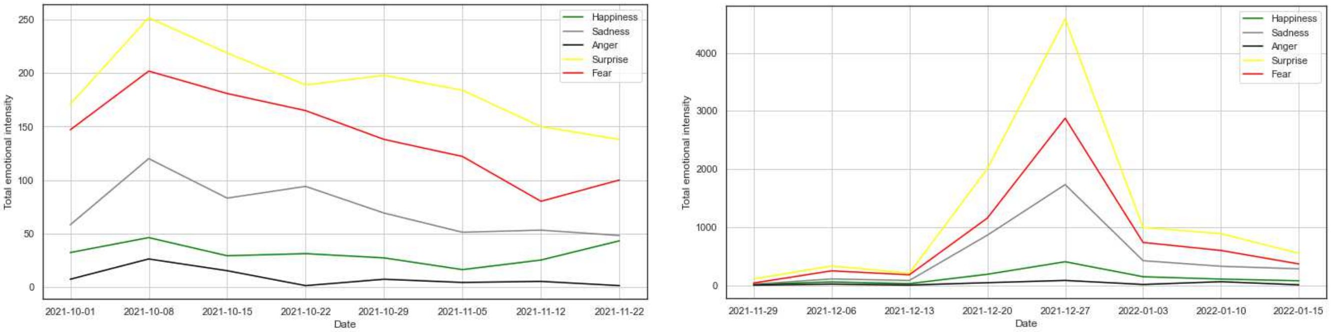
Table 4. Logistic regression model feature interpretation using ELI5.

SN	Happy		Sadness		Anger		Surprise		Fear	
	Weight	Feature	Weight	Feature	Weight	Feature	Weight	Feature	Weight	Feature
1	+3.328	identified	+2.880	deaths	+2.960	say	+2.649	impact	+3.209	bans
2	+3.061	good	+2.716	titled	+2.796	first	+2.558	stay	+3.111	ban
3	+2.920	day	+2.563	confirmed	+2.697	discovered	+2.040	might	+3.006	concerns
4	+2.872	important	+2.350	high	+2.531	really	+2.039	swear	+2.940	surge
5	+2.801	ready	+2.330	panic	+2.529	said	+1.961	red	+2.764	concern
6	+2.733	great	+2.323	wave	+2.507	coming	+1.946	sharp	+2.685	past
7	+2.701	protect	+2.321	rules	+2.413	breaking	+1.780	hard	+2.650	yes
8	+2.634	excellent	+2.318	death	+2.407	help	+1.767	hits	+2.553	please
9	+2.603	approves	+2.242	stock	+2.364	taking	+1.760	despite	+2.549	response
10	+2.472	days	+2.220	relief	+2.265	alarm	+1.749	shut	+2.468	quickly

3.2 COVID–19 vaccine related discussions on Twitter before and during Omicron

In this part, we examine the discussions about COVID–19 on Twitter before and during the Omicron outbreak in South Africa. The word cloud of most dominant words during and before Omicron is presented in Appendix A. The word Omicron was seen as one of the dominant words that was used in the COVID–19 vaccine related tweets during the Omicron outbreak. Similarly, the trend in tweets with time before and during Omicron is presented in Appendix B. As shown in Appendix B, there was an upsurge of tweets during the first week of the Omicron variant outbreak, precisely, between November 22nd, 2021, to December 08, 2021. Then, the graph remained flattened till January 15, 2022.

Further, the total emotional intensities of COVID-19 vaccine related discussions on Twitter with time before and during Omicron is shown in Figure 5. This shows how users emotions changed with time during an outbreak of a new variant in South Africa.



(a) Before Omicron

(b) During Omicron

Figure 5. Emotional changes expressed on tweets

Before the Omicron outbreak, all emotional intensities went up on the week of October 08, 2021, and started to go down till the week of November 22, 2021, see Figure 5a. Similarly, during the Omicron outbreak, all the emotional intensities were at their peak in the week of December 27, 2021, see Figure 5b. This period was possibly when the Omicron variant was the most dominated cases in the hospital, as such, it may have triggered a lot of discussions and emotions on social media and Twitter in particular. Next, we present the impact of COVID–19 vaccine related discussions on vaccination before the Omicron outbreak.

3.2.1 Impact of COVID-19 vaccine related discussions on vaccination before Omicron

In this section, we present the impact of COVID–19 vaccine related discussions to South African vaccination data before Omicron. The vaccination data we used was for people within the ages of 18 years to 34 years[56]. We choose the data for the above age range because they are regarded as the most active on social media, especially Twitter. The duration was from October 1, 2021, to November 22, 2022. We choose this duration because we are interested in the immediate months before the outbreak of the Omicron variant, to enable us to compare the result with the period during the Omicron outbreak. We performed Granger causality test to evaluate the impact of the COVID–19 related tweets intensities against vaccination in each South African province within the selected period of time. We want to know if there is an association between the emotional intensities expressed on these tweets and vaccination within a period. Thus, we test the following hypothesis as follows:

Hypothesis One

- Null hypothesis (H_0): There is no association between emotional intensity and vaccination.
- Alternative hypothesis (H_1): There is an association between emotional intensity and vaccination.

By the term Granger-cause, we mean that there is an association between the emotional intensities and vaccination. That is, the emotional intensities expressed on COVID-19 related Twitter posts within the period of discussion could be useful in predicting vaccination uptake in the future. The Ganger causality test produces an F-test statistic with a corresponding p-value. If the p-value is below $\alpha = 0.05$ significant level, then, we reject H_0. We concluded that we have sufficient evidence to say that emotional intensities Granger-causes vaccination with time. That means, there is an association between emotional intensity and vaccination.

The result of the Granger causality test showed that there is a statistically significant evidence for Mpumalanga province (F-test = 2.6939, p-value = 0.0011) and Western Cape province (F-test = 30.4013, p-value = 0.0053) of the association between emotional intensities and vaccination before the Omicron outbreak. We, therefore, reject the null hypothesis, H_0, and conclude that emotional intensities expressed on Twitter posts could be used to predict vaccination uptake in the future, especially in these two provinces given our data, since we have an association between emotional intensity and vaccination.

The variations in the emotional classes before Omicron outbreak for these two provinces were analysed. The analysis showed Anger (p=0.04) emotional class as being evidently significant in association with vaccination in Mpumalanga Province. Meanwhile, Anger (p=0.01) and Fear (p=0.005) emotional classes showed statistically significant evidence in the association with vaccination in Western Cape Province. These go further to suggest that Anger and Fear emotional classes expressed on COVID–19 related tweets could be a contributing factor in predicting vaccination uptake in Mpumalanga and Western Cape provinces especially when there is an outbreak.

However, our result of the analysis of the seven remaining Provinces in South Africa showed that there is no statistically significant evidence enough to reject H_0. Table 5 summarises the outcome of the analysis.

Table 5. Outcome of the association of emotional intensities from discussions on vaccine related post to vaccination before Omicron for the remaining seven provinces.

SN	Province	F-test	P-values
1	Eastern Cape	1.2838	0.3205
2	Free State	1.83591	0.2463
3	Gauteng	0.1820	0.6916
4	Kwazulu Natal	4.4556	0.1024
5	Limpopo	0.0046	0.9491
6	North West	0.0126	0.9161
7	Norther Cape	1.1014	0.3532

Next, we discuss the impact of COVID–19 related discussions on vaccination during Omicron outbreak.

3.2.2 Impact of COVID-19 vaccine related discussions on vaccination during Omicron Outbreak

The vaccination data was taken from people within the ages of 18 years to 34 years as usual. We conducted the Granger causality test of emotional intensities expressed on COVID-19 vaccine related discussions to increase in vaccination uptake during Omicron. The increase in vaccination uptake was achieved from the difference of daily total vaccinations before Omicron and daily total vaccinations during Omicron in all the South African provinces. We started by checking if there is a correlation between daily tweets with increase in daily vaccinations in all the provinces. The results showed that the increase in daily tweets significantly positively correlated with increase in daily vaccinations in all the provinces (see Figures 6a–6g and 6i) except in Northern Cape province, see Figure 6h.

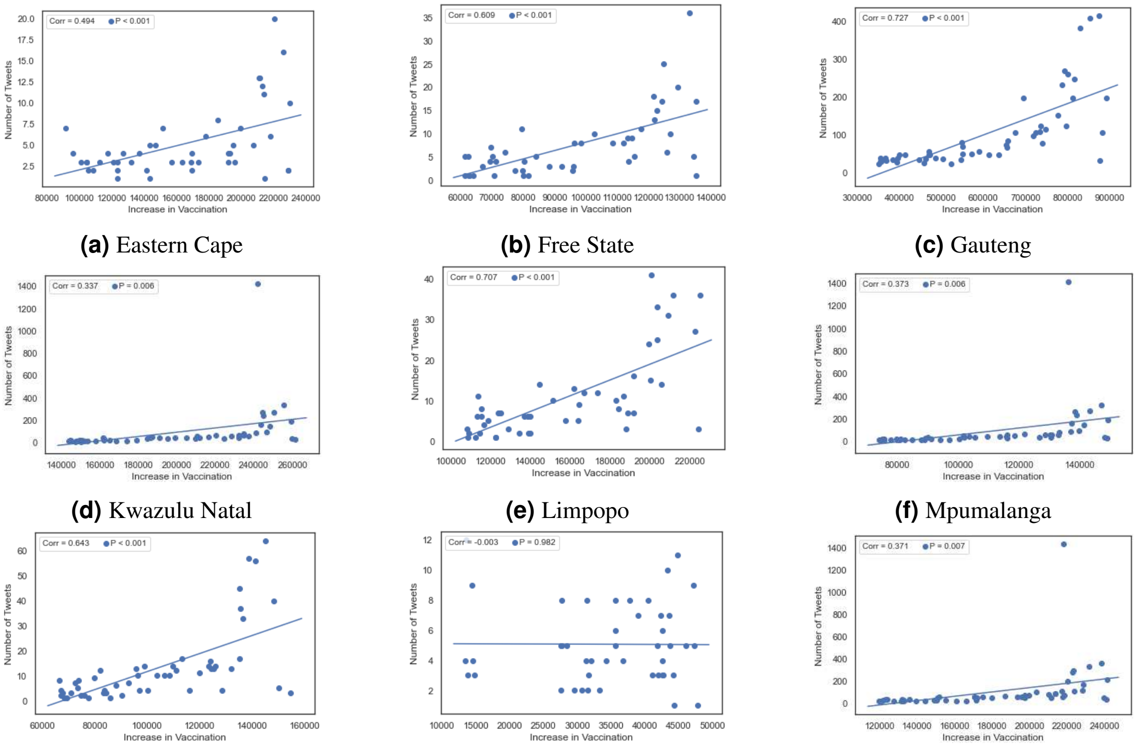
The result of the Granger causality test on our data during Omicron, showed a statistically significant evidence for Eastern Cape province (F-test = 0.8125, P-value = 0.0115), Gauteng province (F-test = 0.1625, P-value = 0.0001), Limpopo province (F-test = 1.2898, P-value = 0.0078), and North West province (F-test = 1.0572, P-value = 0.3623) of the association between emotional intensities expressed on COVID–19 related tweets and increase in vaccination during the Omicron outbreak. Given the above result, we therefore reject the null hypothesis, H_0, and conclude that increase in emotional intensities expressed on COVID–19 related Twitter posts could impact the increase in vaccination uptake in these provinces. The variations in the emotional classes for these four provinces were further analysed.

The analysis showed Anger (p=0.001) and Surprise (p=0.001) emotional classes as being evidently significant in association with increase in vaccination in Eastern Cape province. Happy, Sadness, Anger, Surprise, and Fear emotional classes with p<0.001 also showed a statistically significant evidence in association with increase in vaccination in Gauteng province. Similarly, Sadness (p=0.0009), Surprise (p<0.001) and Fear (p=0.001) were statistically significant in association with vaccination in Limpopo province, while Anger (p=0.02), Surprise (p<0.001), and Fear (p=0.0001) emotional classes were also statistically significant in association with increase in vaccination in North West province. The above results suggest that the emotional classes expressed on COVID–19 related tweets could contribute in predicting vaccination uptake of a certain age range (18 - 34 years) in Eastern Cape, Gauteng, Limpopo, and North West provinces, especially during an outbreak of a new variant.

Further, our result for the remaining five provinces in South Africa during Omicron showed that there is no statistically significant evidence enough to reject H_0. Table 6 summarises the outcome of the analysis.

3.2.3 Comparison of emotional classes by provinces during Omicron Outbreak

In this section, we present the comparison of the emotional classes identified on the COVID–19 vaccine related discussions before and during Omicron outbreak. The essence of this analysis is to identify if there is an association between the emotional classes before and during Omicron. Should there be one, then, it goes to suggest that the cause of the emotional class before Omicron is likely to be the same as the cause of the emotional class during Omicron. This is expected to be useful in health policy making, especially in managing and predicting vaccine hesitancy. The `Mann-Whitney U` test was used to perform the



(a) Eastern Cape **(b)** Free State **(c)** Gauteng

(d) Kwazulu Natal **(e)** Limpopo **(f)** Mpumalanga

(g) North West **(h)** Northern Cape **(i)** Western Cape

Figure 6. Association of daily tweets with increase in daily vaccination by provinces in South Africa.

Table 6. Outcome of the association of emotional intensities from discussions on COVID–19 vaccine related post to increase in vaccination during Omicron for the remaining five provinces.

SN	Province	F-test	P-values
1	Free State	0.5701	0.4923
2	Kwazulu Natal	0.9825	0.3778
3	Mpumalanga	1.0557	0.3623
4	Northern Cape	0.7514	0.4349
5	Western Cape	7.9087	0.2438

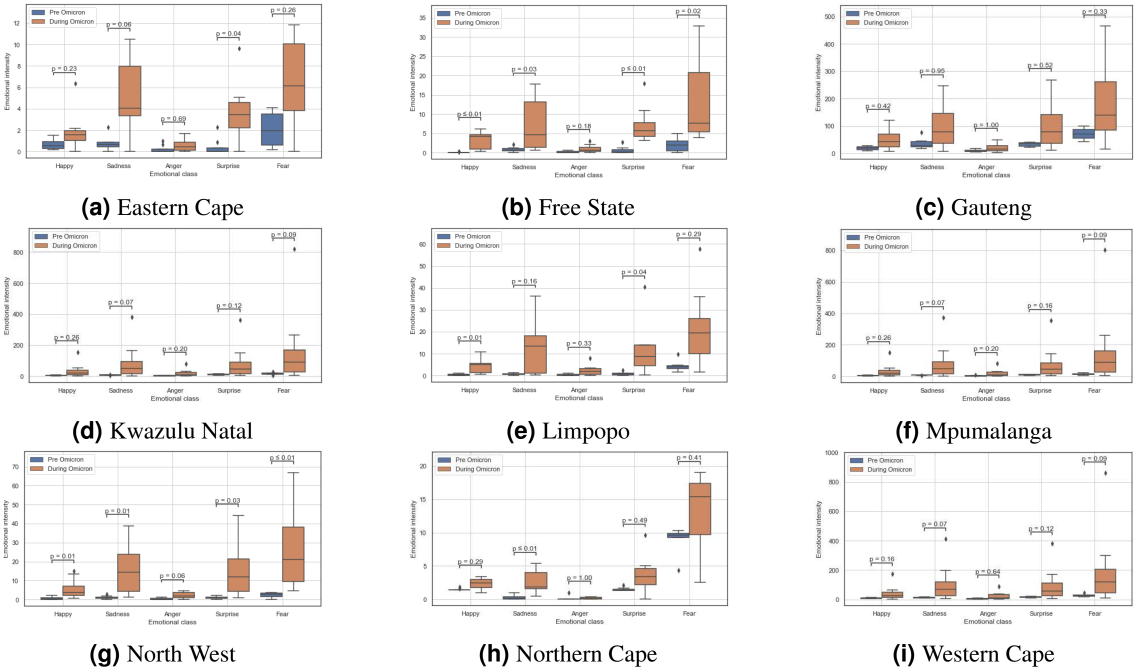
comparison of the emotional classes for each province. Given that the provinces are treated independently, the `Mann-Whitney U` was used for this study because it is good for comparison of independent variables. Thus we test the following hypothesis as follows:

Hypothesis Two

- H_0: There is no association between emotional intensities before and during Omicron outbreak.
- H_1: There is an association between emotional intensities before and during Omicron outbreak.

The `Mann-Whitney U` test produces a corresponding p-value. If the p-value is below $\alpha = 0.05$ significant level, then, we reject H_0. We concluded that we have sufficient evidence to say that there is an association between emotional intensities before and during Omicron. This goes further to suggest that the said emotional classes identified before and during Omicron could be triggered by the same external factor. Figure 7 summarises the outcome of the analysis by different provinces in South Africa.

The result of the `Mann-Whitney U` test on the emotional intensities before and during Omicron by each provinces in South Africa showed a statistically significant evidence for Surprise (p=0.04) emotional class in Eastern Cape province, see Figure 7a. Free State and North West provinces demonstrated statistically significant evidence before and during Omicron for Happy ($p \leq 0.01$), Sadness ($p \leq 0.03$), Surprise ($p \leq 0.03$), and Fear ($p \leq 0.02$) emotional classes, see Figures 7b and 7g



(a) Eastern Cape **(b)** Free State **(c)** Gauteng

(d) Kwazulu Natal **(e)** Limpopo **(f)** Mpumalanga

(g) North West **(h)** Northern Cape **(i)** Western Cape

Figure 7. Comparison of emotional intensities before and during Omicron by each provinces in South Africa using the `Mann-Whitney U` test

respectively. Similarly, there is also statistically significant evidence for Happy (p=0.03) and Surprise (p=0.04) emotional classes in Limpopo province (see Figure 7e). While Sadness (p$\leq$0.01) emotional class is statistically significant in Northern Cape province (see Figure 7h). Given the above significant levels, we reject the null hypothesis, H_0, and conclude that there is an association between emotional intensities before and during Omicron in these provinces. This means that similar external factor could be responsible for the increase in emotional intensities of these emotional class expressed on Twitter posts from different provinces.

Finally, our result of the emotional intensities before and during Omicron for the remaining four provinces, Gauteng, Kwazulu Natal, Mpumalanga, and Western Cape showed that there is no statistically significant evidence enough to reject H_0, see Figure 7.

3.3 Limitations

The Twitter data used for this research only reflects the opinion of Twitter users whose location was South Africa from Octorer 1, 2021 to January 15 2022. South Africa, is a population of about 60 million people and have only an estimated 15% online adults who use Twitter, and of the age 18–34 years[57]. Therefore, this research does not, at large, represent the opinion of people of South Africa towards COVID–19 vaccine related discussions. However, this research only provided an insightful analysis from the Twitter data to complement existing data in supporting policy making.

It is also important to state here that the Text2Emotion pre trained model used for emotion classification does not have the capacity to properly label figurative language, such as sarcasm, pidgin English, and vernacular. 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 0.86, 0.97, 0.96, 0.68, and 0.83 ROC-AUC scores achieved with the NB, LR, SVMs, DT, and KNN classification algorithms, we assume it was able to deal with the noise generated by this obvious challenge in multiclass classification and prediction.

3.4 Ethical considerations

The study was approved by Twitter and we were 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, hence, all personal information was removed.

4 Discussion

We used the Twitter API to generate and process data of COVID–19 vaccine related posts from South Africa between October 1, 2021 and January 15 2022. The tweets were divided into two, namely: before Omicron (October 1 – November 22, 2021) and during Omicron (November 23, 2021 – January 15, 2022). As expected, during Omicron, we observed an upsurge of daily tweets from the week of November 23, 2021 to the week of December 01, 2021, see Appendix **??**. However, there was a sharp decrease in daily tweets and further flattening of the curve after the week of December 01, 2021 till the week of January 15, 2022. This could be attributed to the fact that after the first week of the Omicron outbreak, the variant appeared to be causing more infections but fewer hospital admissions than delta according the South African data[12,58]. As such, a lot of people in South Africa seem to show less interest in vaccination[39,40,58]. It is believed at this time that people cared less or became used to the variant and it possibly reflected on the rate of COVID–19 vaccine related discussions on Twitter[11,12].

The Text2Emotion pretrained model was used to classify and label tweets. The tweets were labelled as Happy, Sadness, Anger, Surprise, and Fear with their corresponding emotional intensities. This approach could be likened to be similar as the study in[39,59,60]. For example, in Naila et al[59], the Text2Emotion model was used to detect emotions on cryptocurrency related tweets. Here, Text2Emotion model was used to identify the embedded emotion on cryptocurrency related tweets and presented the output in the form of a dictionary, classifying tweets as Happy, Sadness, Anger, Surprise, or Fear. Similarly, in[61], Text2Emotion models was used to classify emotions in Twitter communication and stock prices of firms during the COVID–19 pandemic. In this study, emotions in organizational tweets were classified as positive (happiness) and negative (anger, fear, and sadness) and their association with stock prices of firms. The above discussions shows that the choice of Text2Emotion is good for our type of analysis as it has been successfully used is similar works.

Further, the tweets labelled with the Text2Emotion model was fit into five different machine learning classification models. The result of the machine learning models, trained to validate the labelled tweets, showed that LR with 0.91 accuracy has higher performance than NB, SVMs, DT, and KNN. Of course, this higher performance of the LR model is as result of large feature set. The 20,766 processed tweets generated a large feature set that was suitable for the LR higher performance. While SVMs and DT achieved 0.84 and 0.82 accuracy scores respectively, NB and KNN performed poorly. In all, LR demonstrated to have performed better with a large feature set and multiclass prediction. A similar experiment was conducted in[59] where SVMs, Logistic Regression, GNB, ETC, DT, and KNN classification models where used to classify emotions expressed on cryptocurrency related tweets. The LR and SVMs performed better than the other models with an accuracy of 0.90 each. Meanwhile, KNN and GNB had the lowest performance. In this study, SVMs and LR demonstrated to perform better on a large feature set, as such, the 40,000 tweets used for the study generated a large feature set which was suitable to the models performance in multiclass prediction.

As demonstrated in our study and in Naila et al[59], large feature set from textual data is a function of performance for LR in multiclass prediction, while, for KNN the story is not the same, it performs otherwise. Therefore, the choice of Logistic Regression model for classification of tweets is suitable for this type of study.

The result of the Granger causality test performed on our data before the Omicron outbreak showed a statistically significant evidence for Mpumalanga and Western Cape provinces of the association between emotional intensities and vaccination. This suggests that the intensity of COVID–19 vaccine related discussions on Twitter could be used to predict vaccination uptake especially in these two provinces with the age range 18–34 years. However, during the Omicron outbreak, the result of the analysis showed no statistically significant evidence of the association between emotional intensities expressed on COVID–19 related discussions and increase in vaccination uptake. This means that emotional intensity is not a factor in predicting increase in vaccination in these provinces. This corroborated with the observation that people were no longer interested or scared by the variant. It possibly reduced the rate vaccination and such discussions on Twitter in these provinces[11,12].

However, our data analysis for Eastern Cape, Gauteng, Limpopo, and North West provinces showed a statistically significant evidence of an association between emotional intensities expressed on COVID–19 related tweets, and increase in vaccination uptake during the Omicron outbreak. This suggests that emotions may have caused more people to vaccinate as this was not the case in these provinces before Omicron as identified in our data. This obvious emotional change which was demonstrated as Happy ($p \leq 0.001$), Sadness ($p \leq 0.001$), Anger ($p \leq 0.001$), Surprise ($p \leq 0.001$), and Fear ($p \leq 0.001$) emotional classes could also be attributed to a large media hype and awareness created by health authorities with the South Africa's National Coronavirus Command Council (NCCC), especially, when they announced that the country will retain its lockdown to the lowest five-tier system of restrictions[62] due to the increase in the number of COVID–19 cases at that time. This announcement was made in the couple two weeks of the Omicron outbreak, precisely, December 16, 2021, because of its high infectious rate as suggested by the health authorities[12,14,62].

Finally, we conducted an analysis of the association between the emotional classes before and during Omicron variant using the `Mann-WhitneyU` test of independent variable. Surprise emotional class showed a strong statistically significant evidence in Eastern Cape province. There is also a statistically significant evidence for Happy, Sadness, Surprise, and Fear emotional classes in Free State and North West provinces. Similarly, Happy and Surprise emotional classes showed strong statistically significant evidence in Limpopo province. Meanwhile, Sadness emotional class was statistically significant in Northern Cape province. This suggest that the same external factors could be responsible for the dynamics of these emotional classes, as such, can affect their emotional intensities. An understanding of these dynamics in emotions expressed on Twitter posts during an outbreak could be useful in health policy, planing, and management of a new COVID–19 variant or future pandemics. The study in[63] provided a comprehensive finding of the spatiotemporal dynamics of Omicron to control the spread of the Omicron virus in the nine South African provinces. This corroborated our claim in understanding the emotional dynamics during an outbreak for control, planning and management of health policy.

5 Conclusion

In this research, Twitter posts containing daily updates of location-based COVID–19 vaccine–related tweets before and during Omicron were collected from South Africa. Our focus was during the Omicron era, that is, October 1, 2021 to January 15 2022. We observed that number of tweets significantly positively correlated with the increase in vaccination across all South African provinces (Corr$\leq$0.532, P$\leq$0.003) except Northern Cape province.

Additionally, emotional intensities for vaccine-related tweets were validated using machine learning classification algorithms, such as, NB, LR, SVMs, DT, and KNN. These emotional intensities showed a strong association with the increase in vaccination during Omicron (P<0.04) in Eastern Cape, Gauteng, Limpopo, and North West provinces than other provinces. Similarly, comparison of the intensities of the emotional classes differed across provinces before and during Omicron.

These goes further to suggest that understanding the emotional dynamics expressed on social media texts and their relationship or association with vaccination uptake can be relevant for health policy, planning, and control. Our study, therefore, promises to be interesting, as we harnessed the power of social media data in complementing existing data, especially in the areas where there is little or no available data.

Ethics Statements

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 whose tweet were collected, as such all personal information was removed.

Funding

This research is funded by Canada's International Development Research Centre (IDRC) and the Swedish International Development Cooperation Agency (SIDA) (Grant No. 109559-001). JDK acknowledges support from IDRC (Grant No. 109981), New Frontier in Research Fund- Exploratory (Grant No. NFRFE-2021-00879) and NSERC Discovery Grant (Grant No. RGPIN-2022-04559).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability of Data and Materials

The dataset used and/or analysed during the current study is available from the corresponding author on reasonable request.

References

1. Lin, L., Qi, Z. & Jiancheng, X. Changes of laboratory cardiac markers and mechanisms of cardiac injury in coronavirus disease 2019. *BioMed Res. Int.* **7**, 171–174, DOI: https://doi.org/10.1155/2020/7413673 (2020).
2. Hulswit, R., de Haan, C. & Bosch, B.-J. Chapter two - coronavirus spike protein and tropism changes. In Ziebuhr, J. (ed.) *Coronaviruses*, vol. 96 of *Advances in Virus Research*, 29–57, DOI: https://doi.org/10.1016/bs.aivir.2016.08.004 (Academic Press, 2016).
3. Zhao, Y. *et al.* The global transmission of new coronavirus variants. *Environ. Res.* **206**, 112240, DOI: https://doi.org/10.1016/j.envres.2021.112240 (2022).
4. Mukherjee, R. & Satardekar, R. Why are some coronavirus variants more infectious? *J. Biosci.* **46**, DOI: https://doi.org/10.1007/s12038-021-00221-y (2021).

5. Ewen, C. Multitude of coronavirus variants found in the us — but the threat is unclear. Available on https://bit.ly/3a2dfht. [Accessed: 2022-06-01].

6. CMAJ. Xe, xd & xf: what to know about the omicron hybrid variants. *Cmajnews.com* **194**, DOI: doi:10.1503/cmaj.1095998 (2022).

7. Kewei, M. & Jieliang, C. Omicron xe emerges as sars-cov-2 keeps evolving. *The Innov.* **3**, 100248 (2022).

8. Jyoti, P. S. & Kailash, C. S. New variant xe more transmissible than omicron: Alarming towards covid 4th wave in india. *Biotica Res. Today* **4**, 100248 (2022).

9. Rahimi, F. & Talebi Bezmin Abadi, A. Hybrid sars-cov-2 variants. *Int. journal surgery (London, England)* **102**, 106656, DOI: 10.1016/j.ijsu.2022.106656 (2022).

10. Organisation, W. H. Update on omicron. Available on https://www.who.int/news/item/28-11-2021-update-on-omicron. [Accessed: 2022-06-01.

11. Abdullah, F. *et al.* Decreased severity of disease during the first global omicron variant covid-19 outbreak in a large hospital in tshwane, south africa. *Int. J. Infect. Dis.* **116**, 38–42, DOI: https://doi.org/10.1016/j.ijid.2021.12.357 (2022).

12. Dyer, O. Covid-19: Omicron is causing more infections but fewer hospital admissions than delta, south african data show. *BMJ* **375**, DOI: 10.1136/bmj.n3104 (2021). https://www.bmj.com/content/375/bmj.n3104.full.pdf.

13. for Disease Control, C. & Prevenstion. Omicron variant: What you need to know. Available on https://www.cdc.gov/coronavirus/2019-ncov/variants/omicron-variant.html. [Accessed: 2022-06-01.

14. Lin, X.-Q. *et al.* The impact of the omicron epidemic on the health behavior in cape town, south africa. *One Heal.* **14**, 100395, DOI: https://doi.org/10.1016/j.onehlt.2022.100395 (2022).

15. Cloete, J. *et al.* Paediatric hospitalisations due to covid-19 during the first sars-cov-2 omicron (b.1.1.529) variant wave in south africa: a multicentre observational study. *The Lancet Child & Adolesc. Heal.* **6**, 294–302, DOI: https://doi.org/10.1016/S2352-4642(22)00027-X (2022).

16. Xuemei, H., Weiqi, H., Xiangyu, P., Guangwen, L. & Xiawei, W. Sars-cov-2 omicron variant: Characteristics and prevention. *Med Comm* **2**, 838–845, DOI: https://doi.org/10.1002/mco2.110 (2021).

17. Frederic, G., Marek, K. & Tomasz, L. The spread of sars-cov-2 variant omicron with the doubling time of 2.0–3.3 days can be explained by immune evasion. *Viruses* **14**, 294, DOI: 10.3390/v14020294 (2022).

18. Tasnim, S., Hossain, M. & Mazumder, H. Impact of rumors and misinformation on covid–19 in social media. *J. Prev. Medicine & Public Heal.* **202**, 171–174, DOI: https://doi.org/10.3390/journalmedia2010007 (2021).

19. Shu-Feng, T. *et al.* What social media told us in the time of covid-19: a scoping review. *The Lancet Digit. Heal.* **3**, e175–e194 (2013, Available from: https://doi.org/10.1016/S2589-7500(20)30315-0 [Accessed 2022-02-28]).

20. Marcec, R. & Likic, R. Using twitter for sentiment analysis towards astrazeneca/oxford, pfizer/biontech and moderna covid–19 vaccines. *Postgrad. Med. J.* **10**, 1–7, DOI: Availablefrom:https://pmj.bmj.com/ (2021).

21. Al-Zaman, M. Covid–19–related social media fake news in india. *J. Media* **2**, 100–114, DOI: https://doi.org/10.3390/journalmedia2010007 (2021).

22. Yan, C., Law, M., Nguyen, S., Cheung, J. & Kong, J. Comparing public sentiment toward covid-19 vaccines across canadian cities: Analysis of comments on reddit. *J Med Internet Res* **23**, e32685, DOI: 10.2196/32685 (2021).

23. Lee, K. *et al.* Twitter trending topic classification. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 251–258, DOI: 10.1109/ICDMW.2011.171 (2011).

24. Liu, Y., Han, W., Tian, Y., Que, X. & Wang, W. Trending topic prediction on social network. In *2013 5th IEEE International Conference on Broadband Network & Multimedia Technology*, 149–154, DOI: 10.1109/ICBNMT.2013.6823933 (2013).

25. Shahi, G. K., Dirkson, A. & Majchrzak, T. A. An exploratory study of covid-19 misinformation on twitter. *Online Soc. Networks Media* **22**, 100104, DOI: https://doi.org/10.1016/j.osnem.2020.100104 (2021).

26. Pendyala, V. S. & Figueira, S. Towards a truthful world wide web from a humanitarian perspective. In *2015 IEEE Global Humanitarian Technology Conference (GHTC)*, 137–143, DOI: 10.1109/GHTC.2015.7343966 (2015).

27. Liu, X. Deep learning techniques for sarcasm detection. In *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*, 1–5 (2021).

28. Wolf, K. Measuring facial expression of emotion. *Dialogues Clin. Neurosci.* **17**, 457–462, DOI: 10.31887/DCNS.2015.17.4/kwolf (2015). PMID: 26869846, https://doi.org/10.31887/DCNS.2015.17.4/kwolf.

29. Gaurav, S. & James, J. G. What is an emotion? a connectionist perspective. *Emot. Rev.* **14**, 457–462, DOI: https://doi.org/10.1177/17540739221082203 (2022). https://doi.org/10.1177/17540739221082203.

30. Cherry, K. Emotions and types of emotional responses: The three key elements that make up emotion. Available on https://www.verywellmind.com/what-are-emotions-2795178. [Accessed: 2022-06-01].

31. Armin, S., Narges, T. & Wlodek, Z. Emotion detection in text: a review. *arXiv:1806.00674* DOI: https://doi.org/10.48550/arXiv.1806.00674 (2018).

32. Shivhare, S. N. & Khethawat, S. Emotion detection from text. *arXiv:1205.4944* DOI: https://doi.org/10.48550/arXiv.1205.4944 (2012).

33. Binali, H., Wu, C. & Potdar, V. Computational approaches for emotion detection in text. In *4th IEEE International Conference on Digital Ecosystems and Technologies*, 172–177, DOI: 10.1109/DEST.2010.5610650 (2010).

34. Bandhakavi, A., Wiratunga, N., Massie, S. & Padmanabhan, D. Lexicon generation for emotion detection from text. *IEEE Intell. Syst.* **32**, 102–108, DOI: 10.1109/MIS.2017.22 (2017).

35. Agrawal, A. & An, A. Unsupervised emotion detection from text using semantic and syntactic relations. In *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1, 346–353, DOI: 10.1109/WI-IAT.2012.170 (2012).

36. Nandwani, P. & Verma, R. Review on sentiment analysis and emotion detection from text. *Soc. Netw. Analysis Min.* **11**, DOI: https://doi.org/10.1007/s13278-021-00776-6 (2021).

37. Sailunaz, K., Dhaliwal, M., Rokne, J. & Alhajj, R. Emotion detection from text and speech: a survey. *Soc. Netw. Analysis Min.* **8**, DOI: https://doi.org/10.1007/s13278-021-00776-6 (2018).

38. Cherry, K. The 6 types of basic emotions and their effect on human behavior. Available on https://www.verywellmind.com/an-overview-of-the-types-of-emotions-4163976. [Accessed: 2022-06-01].

39. coronavirus, S. COVID–19 Online resources & new portal (2022). Online: https://sacoronavirus.co.za/,[Accessed: 27 May 2022].

40. Covid19SA. COVID-19 South Africa Dashboard (2022). Online: https://www.nltk.org/, [Accessed: 27 May 2022].

41. Li, F. *et al.* What's new in pandas 1.2.4. Available on https://pandas.pydata.org/pandas-docs/stable/whatsnew/v1.2.4.html. [Accessed: 2022-06-01].

42. Tweet-preprocessor 0.6.0. Available on https://pypi.org/project/tweet-preprocessor/. [Accessed: 2022-01-06].

43. Natural language toolkit. Available on https://www.nltk.org/. [Accessed: 2022-06-01].

44. Honnibal, M. spacy 2: Natural language understand- ing with bloom embeddings, convolutional neural net- works and incremental parsing. *Sentometrics Res.* **1**, 2586–2593 (2017, Available from: https://sentometrics-research.com/publication/72/. [Accessed 2022-02-28]).

45. Aditya, B. Industrial strenght natural language processing in python. Available on https://spacy.io/. [Accessed: 2022-06-01].

46. text2emotion 0.0.5. Available on https://pypi.org/project/text2emotion/. [Accessed: 2022-06-01].

47. Ogbuokiri, B. *et al.* Public sentiments toward covid-19 vaccines in south african cities: An analysis of twitter posts. *Front. public health* **10**, 987376, DOI: 10.3389/fpubh.2022.987376 (2022).

48. Ogbuokiri, B. *et al.* Vaccine hesitancy hotspots in africa: An insight from geotagged twitter posts. *TechRxiv. Prepr.* DOI: 10.36227/techrxiv.20720740.v1 (2022).

49. Yang, F.-J. An implementation of naive bayes classifier. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 301–306, DOI: 10.1109/CSCI46756.2018.00065 (2018).

50. Obaido, G. *et al.* An interpretable machine learning approach for hepatitis b diagnosis. *Appl. Sci.* **12** (2022).

51. Rosol, M., Młyńczak, M. & Cybulski, G. Granger causality test with nonlinear neural-network-based methods: Python package and simulation study. *Comput. Methods Programs Biomed.* **216**, 106669, DOI: https://doi.org/10.1016/j.cmpb.2022.106669 (2022).

52. Mighri, Z., Ragoubi, H., Sarwar, S. & Wang, Y. Quantile granger causality between us stock market indices and precious metal prices. *Resour. Policy* **76**, 102595, DOI: https://doi.org/10.1016/j.resourpol.2022.102595 (2022).

53. Li, G., Zhang, A., Zhang, Q., Wu, D. & Zhan, C. Pearson correlation coefficient-based performance enhancement of broad learning system for stock price prediction. *IEEE Transactions on Circuits Syst. II: Express Briefs* **69**, 2413–2417, DOI: 10.1109/TCSII.2022.3160266 (2022).

54. Bedre, R. Mann-whitney u test (wilcoxon rank sum test) in python [pandas and scipy]. Available on https://www.reneshbedre.com/blog/mann-whitney- u-test.html. [Accessed: 2022-06-01].

55. Angela, F. *et al.* Eli5: Long form question answering. Available on https://arxiv.org/abs/1907.09190. [Accessed: 2022-06-01.

56. Writer, S. The biggest and most popular social media platforms in south africa, including tiktok. Available on shorturl.at/cjoBU (2021). [Accessed on June 11, 2022].

57. Statista. Largest cities in south africa in 2021 by number of inhabitants. Available on https://www.statista.com/statistics/1127496/largest-cities-in-south-africa/. [Accessed: 2022-06-01].

58. Lulan, W. & Genhong, C. Sequence analysis of the emerging sars-cov-2 variant omicron in south africa. *J. medical virology* **94**, 1728–1733, DOI: https://doi.org/10.1002/jmv.27516 (2022).

59. Aslam, N., Rustam, F., Lee, E., Washington, P. B. & Ashraf, I. Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model. *IEEE Access* **10**, 39313–39324, DOI: 10.1109/ACCESS.2022.3165621 (2022).

60. Díaz, S. S., Shaik, J. M. M., Santofimio, J. C. G. & Quintero M., C. G. Intelligent execution of behaviors in a nao robot exposed to audiovisual stimulus. In *2018 IEEE 2nd Colombian Conference on Robotics and Automation (CCRA)*, 1–6, DOI: 10.1109/CCRA.2018.8588149 (2018).

61. Suparna, D. & Indranil, B. Emotions in twitter communication and stock prices of firms: the impact of covid–19 pandemic. *Decision* **47**, 385–399, DOI: https://doi.org/10.1007/s40622-020-00264-4 (2020).

62. Reuters. South africa to retain 'level 1' curbs in omicron fight (2021). Online: https://www.reuters.com/world/africa/south-africa-retain-level-1-curbs-omicron-fight-2021-12-16/, [Accessed: 31 May 2022].

63. Tong, C., Shi, W., Zhang, A. & Shi, Z. Tracking and controlling the spatiotemporal spread of sars-cov-2 omicron variant in south africa. *Travel. Medicine Infect. Dis.* **46**, 102252, DOI: https://doi.org/10.1016/j.tmaid.2021.102252 (2022).

Acknowledgement

The authors wish acknowledge the Africa-Canada AI & Data Innovation Consortium (ACADIC) team at York University, Toronto, Canada and University of the Witwatersrand Johannesburg, South Africa.

Author contributions statement

Blessing Ogbuokiri: Conceptualization of this study, methodology, and Writing - Original draft preparation. Ali Ahmadi: Supervision. Nidhi Tripathi: Data curation. Zahra Movahedi: Data curation. Bruce Mellado: Supervision. Jiahong Wu: Supervision. Ali Asgary: Supervision. James Orbinski: Supervision. Jude Kong: Editorial and project supervision.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ListofAppendices.docx