Machine Learning with Applications 20 (2025) 100644

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

# Emotional reactions towards vaccination during the emergence of the Omicron variant: Insights from twitter analysis in South Africa

Blessing Ogbuokiri [a,f,h,j,*], Ali Ahmadi [b], Nidhi Tripathi [d,f], Laleh Seyyed-Kalantari [f,i], Woldergebriel Assefa Woldegerima [c,f], Bruce Mellado [d,f,j], Jiahong Wu [c,f,j], James Orbinski [f,g,j], Ali Asgary [b,f,j], Jude Dzevela Kong [e,f,j,**]

[a] Department of Computer Science, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, L2S 3A1, Ontario, Canada
[b] Advanced Disaster, Emergency and Rapid-response Simulation (ADERSIM), York University, 4700 Keele St, North York, Toronto, M3J 1P3, Ontario, Canada
[c] Laboratory for Industrial and Applied mathematics, York University, 4700 Keele St, North York, Toronto, M3J 1P3, Ontario, Canada
[d] School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, 1 Jan Smuts Ave, Braamfontein, Johannesburg, 2017, Gauteng, South Africa
[e] Artificial Intelligence & Mathematical Modeling Lab (AIMMLab), Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, M5T 3M7, Ontario, Canada
[f] Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), University of Toronto, 155 College St, Toronto, M5T 3M7, Ontario, Canada
[g] Massey College, University of Toronto, 4 Devonshire Pl, Toronto, M5S 2E1, Ontario, Canada
[h] Responsible & Applied Machine Learning Laboratory (RAML LAB), Brock University, 1812 Sir Isaac Brock Way, St. Catharines, L2S 3A1, Ontario, Canada
[i] Department of Electrical Engineering and Computer Science, York University, 124 Campus Walk #116, North York, Toronto, M3J 1P3, Ontario, Canada
[j] Global South Artificial Intelligence for Pandemic and EPidemic Preparedness and Response Network (AI4PEP), University of Toronto, 155 College St, Toronto, M5T 3M7, Ontario, Canada

## ARTICLE INFO

Dataset link: https://www.kaggle.com/datasets/ogbuokiriblessing/x-omicron-dataset

*Keywords:*
COVID–19
Emotion analysis
Health policymaking
Natural language processing
Omicron

## ABSTRACT

The emergence of the Omicron variant triggered intense emotional reactions toward vaccination in South Africa, particularly evident on platforms like Twitter. These emotions have the potential to significantly influence vaccine confidence and uptake, posing a challenge for public health efforts. However, existing research lacks a detailed understanding of how emotional dynamics during variant-specific outbreaks, such as Omicron, impact vaccination rates, especially at a province level. This gap limits the ability of policymakers to design targeted interventions. Our study addresses this problem by analyzing emotional reactions to vaccination during the Omicron outbreak using geotagged Twitter data and the Text2emotion pre-trained model. We validated the model by hand-labeling a random 10% of tweets and comparing results with BERT-labeled tweets, finding no significant differences ($p < 0.001$ for hand-labeled, $p = 0.002$ for BERT). Using statistical methods such as $\chi^2$, Mann–Whitney U, Granger causality, and Jaccard similarity, we identified a strong association between emotional intensities in vaccine-related posts and vaccination rates during the Omicron period ($p < 0.04$) in specific provinces. Additionally, Latent Dirichlet Allocation (LDA) was employed for topic modeling, revealing variations in emotional reactions across topics and provinces before and during the Omicron variant. Our findings provide actionable insights for health policy-making by highlighting the role of emotional dynamics in vaccine acceptance and offering a province-level analysis of Twitter discussions. This study demonstrates the potential of social media data to understand public sentiment during disease outbreaks and serves as a valuable reference for future academic research.

## 1. Introduction

The SARS-CoV-2 virus (COVID-19) has undergone continuous mutation since its emergence in December 2019, resulting in several variants, including Alpha, Beta, Delta, and Omicron (Aden & Zaheer, 2024; Hulswit, de Haan, & Bosch, 2016; Lin, Qi, & Jiancheng, 2020; Meganck et al., 2024; Wimalawansa, 2024). Among these, the Omicron

* Corresponding author.
** Corresponding author at: Artificial Intelligence & Mathematical Modeling Lab (AIMMLab), Dalla Lana School of Public Health, University of Toronto, 155 College St, Toronto, M5T 3M7, Ontario, Canada.
*E-mail addresses:* bogbuokiri@brocku.ca (B. Ogbuokiri), ahmadia@yorku.ca (A. Ahmadi), nidhi.tripathi@wits.ac.za (N. Tripathi), lsk@yorku.ca (L. Seyyed-Kalantari), wassefaw@yorku.ca (W.A. Woldegerima), bruce.mellado@wits.ac.za (B. Mellado), wujh@yorku.ca (J. Wu), jorbinski@masseycollege.ca (J. Orbinski), asgary@yorku.ca (A. Asgary), jude.kong@utoronto.ca (J.D. Kong).

https://doi.org/10.1016/j.mlwa.2025.100644
Received 30 December 2024; Received in revised form 15 March 2025; Accepted 19 March 2025
Available online 27 March 2025
2666-8270/

variant (B.1.1.529), first identified in South Africa in November 2021, has raised significant concerns due to its heightened transmissibility and potential to evade immune responses (Abdullah et al., 2022; Dyer, 2021; Hulswit et al., 2016; Zhao, Huang, Zhang, Chen, Gao et al., 2022).

The rapid spread of Omicron, coupled with the proliferation of both accurate and inaccurate information on social media platforms like Twitter, has created substantial challenges for public health communication and vaccine acceptance (Tasnim, Hossain, & Mazumder, 2024; Wella, Mtambo, Lazaro, & Mapulanga, 2025).

Despite extensive research on COVID-19 sentiment and emotions, several critical gaps remain. Most studies have focused on general sentiment trends during the pandemic, with limited attention to emotions specific to emerging variants like Omicron (Kumar, Singh, & Singh, 2023; Ogbuokiri, Ahmadi, Nia, et al., 2023). Additionally, there is a lack of research examining province- or region-level variations in emotional responses to vaccination, which is crucial for designing targeted public health interventions (Ogbuokiri et al., 2022; Ogbuokiri, Ahmadi, Nia, et al., 2023). Another significant gap is the absence of studies directly correlating emotional intensities in social media discussions with actual vaccination rates (Ogbuokiri et al., 2022; Wang et al., 2022). Furthermore, there is a pressing need for real-time sentiment analysis to understand public reactions during critical periods, such as the emergence of highly transmissible variants like Omicron (Ogbuokiri, Ahmadi, Mellado, et al., 2023; Ogbuokiri, Ahmadi, Nia, et al., 2023).

The Omicron variant is highly transmissible, surpassing the Delta variant in its ability to spread rapidly (Lin et al., 2022). Within two weeks of its emergence, daily Omicron cases in South Africa surged from 300 to 3,000 (Cloete et al., 2022; Lin et al., 2022), leading to its dominance in many countries (Frederic, Marek, & Tomasz, 2022; Xuemei, Weiqi, Xiangyu, Guangwen, & Xiawei, 2021). The variant's spike protein contains over 30 mutations, enhancing its ability to infect human cells and potentially evade immune responses, particularly in unvaccinated individuals (Aden & Zaheer, 2024; Hulswit et al., 2016; Mukherjee & Satardekar, 2021; Wimalawansa, 2024; Zhao et al., 2022).

During this period, Twitter became a primary platform for information sharing and discussion. However, the lack of editorial oversight allowed unverified rumors and misinformation to spread widely (Bangyal et al., 2021; Tasnim et al., 2024; Wella et al., 2025). This contributed to heightened anxiety and distrust, ultimately influencing vaccine hesitancy. Since emotions play a significant role in decision-making and public health behavior (Wang et al., 2022), understanding emotional responses to Omicron-related vaccination efforts is critical for improving vaccine acceptance.

Our study addresses these gaps by analyzing emotional reactions to vaccination during the Omicron variant outbreak in South Africa. Using geotagged Twitter data and advanced emotion classification models, we provide real-time insights into public sentiment. Specifically, we investigate emotional responses to vaccination during the Omicron outbreak, offering a unique perspective on how public sentiment evolves in response to emerging variants. By analyzing emotional reactions across South Africa's nine provinces, we identify regional disparities in vaccine hesitancy and acceptance, enabling targeted public health interventions. Additionally, we establish a direct link between emotional intensities in Twitter discussions and actual vaccination rates, providing actionable insights for policymakers. Our approach also captures real-time public sentiment during the Omicron surge, offering valuable guidance for healthcare decision-making during critical periods. Furthermore, we evaluate the effectiveness of different emotion classification models, including Text2Emotion, hand-labeled, and BERT-based methods, demonstrating the reliability of Text2Emotion for analyzing vaccine-related discussions.

The primary objective of this research is to employ natural language processing (NLP) techniques on Twitter posts from October 1, 2021, to January 15, 2022, to identify and compare users' emotional reactions to COVID-19 vaccine-related tweets before and during the emergence of the Omicron variant across South Africa's nine provinces. The study aims to correlate these emotional intensities with actual vaccine uptake to understand the dynamics of emotional changes and their potential influence on vaccination acceptance rates. Additionally, it seeks to identify provinces that may be more prone to vaccine hesitancy based on the emotional sentiments expressed in tweets, providing insights to guide health policy-making decisions.

To achieve these goals, the research addresses the following specific objectives:

1. Identify and categorize emotions in Twitter discussions related to COVID-19 vaccines before and during the Omicron outbreak.
2. Analyze the correlation between emotional trends in tweets and vaccine uptake rates across South African provinces.
3. Examine the impact of emotional fluctuations on vaccine acceptance during the Omicron surge.
4. Compare emotional reactions to vaccine-related discussions before and during the Omicron era.
5. Generate and classify discussion topics on COVID-19 vaccines and Omicron using topic modeling.

By achieving these objectives, our study provides a comprehensive understanding of public emotional reactions during the Omicron outbreak and offers evidence-based recommendations to enhance vaccine acceptance during future outbreaks.

The remainder of this manuscript is organized as follows: Section 2 reviews related literature, Section 3 details the methodology, Section 4 presents the findings, Section 5 discusses the results, and Section 6 concludes the study.

## 2. Related works

The COVID-19 pandemic has spurred extensive research on sentiment and emotion analysis of social media data, particularly Twitter, to understand public reactions, misinformation, and vaccine hesitancy. This section organizes related works into three key themes: (1) Sentiment Analysis of COVID-19 Tweets, (2) Emotion Analysis of COVID-19 Tweets, and (3) Vaccine-Related Sentiment and Emotion Analysis. By synthesizing recent studies, we identify gaps in the literature and highlight the contributions of our work.

### 2.1. Sentiment analysis of COVID-19 Tweets

Sentiment analysis of COVID-19-related tweets has been widely explored to gauge public opinion and emotional responses during the pandemic. Early studies focused on developing robust models for sentiment classification. For instance, Tharu, Pokhrel, and Lamichhane (2023) introduced novel feature extraction methods and convolutional neural networks (CNNs) for sentiment analysis of Nepali tweets, achieving strong performance in classifying positive, neutral, and negative sentiments. Similarly, Gulati et al. (2022) compared machine learning models for sentiment analysis across tweets from the USA, UK, and Bangladesh, identifying CatBoost as the most effective model. Their findings highlighted the influence of factors such as lockdowns, quarantine, and vaccines on public sentiment.

Recent advancements in deep learning have further improved sentiment analysis accuracy (Bangyal, Amina, Shakir, Ubakanma, & Iqbal, 2023; Zamir et al., 2024). Kumar et al. (2023) proposed the CT-BERT-LSTM model, which combines COVID-Twitter-BERT and LSTM, achieving an F1-score of 0.88 for sentiment analysis of COVID-19 vaccination tweets. This model demonstrates the potential of hybrid approaches for enhancing transparency and decision-making in public health contexts. Additionally, Ainapure et al. (2022) employed deep learning and lexicon-based techniques to analyze sentiments expressed by Indian citizens regarding COVID-19 vaccination, providing highly accurate models for understanding public sentiment during crises.

### 2.2. Emotion analysis of COVID-19 Tweets

Emotion analysis extends beyond sentiment classification by identifying specific emotional reactions, such as fear, anger, sadness, and trust, in COVID-19-related tweets. Xie et al. (2021) analyzed over 2 million English tweets to examine public perception of COVID-19 vaccines during the vaccine development period. Their study revealed a shift from fear to trust as the dominant emotion over time, highlighting the importance of monitoring public attitudes for successful vaccine deployment.

Similarly, Braig et al. (2022) conducted a systematic literature review of machine learning-based sentiment and emotion analysis techniques applied to COVID-19 Twitter data. They found that ensemble models, particularly BERT and RoBERTa fine-tuned on Twitter data, were highly effective in assessing public sentiment and emotions. However, the review also identified limitations in handling metaphorical language and slang, which are common in social media text.

Recent studies have also explored the emotional impact of COVID-19 lockdowns. For example, a study analyzing German-language Twitter posts found that rates of anxiety and sadness decreased after the introduction of social contact restrictions, while positive emotions remained stable (Hanschmidt & Kersting, 2023). These findings underscore the value of real-time social media analysis for understanding evolving public sentiments during health crises.

### 2.3. Vaccine-related sentiment and emotion analysis

Vaccine hesitancy and acceptance have been central themes in COVID-19-related social media research. Yousefinaghani, Dara, Mubareka, Papadopoulos, and Sharif (2021) analyzed over 4.5 million tweets to understand public sentiments and opinions toward COVID-19 vaccines. Their study found a prevalence of positive sentiments and higher engagement with positive content, but also identified varying patterns of vaccine rejection and hesitancy across countries. Similarly, Cooper, van Rooyen, and Wiysonge (2021) highlighted the social nature of COVID-19 vaccine hesitancy in South Africa, emphasizing the influence of factors such as age, race, education, and geography.

Recent studies have also explored the role of healthcare professionals in shaping public opinion. Elyashar, Plochotnikov, Cohen, Puzis, and Cohen (2021) analyzed Twitter posts from healthcare professionals in the US,UK, and South Africa finding that they discussed common health, social, and political issues related to the pandemic. This suggests the potential of using Twitter discourse to monitor public health responses during emergencies.

Ogbuokiri et al. (2022) conducted a city-level analysis of public sentiments toward COVID-19 vaccines in South Africa using Twitter data. Their study identified key themes and sentiments in vaccine-related discussions but did not explore emotional reactions during the emergence of specific variants, such as Omicron. Additionally, Ogbuokiri, Ahmadi, Mellado, et al. (2023) investigated the impact of post-vaccination sentiment on booster jab acceptance, highlighting the role of public sentiment in shaping vaccine uptake. However, their study did not address the correlation between emotional intensities and vaccination rates at a granular level. Finally, Ogbuokiri, Ahmadi, Nia, et al. (2023) analyzed geotagged Twitter posts to identify vaccine hesitancy hotspots in Africa, providing valuable insights into regional variations in vaccine acceptance but lacking a focus on emotional dynamics during variant-specific outbreaks.

### 2.4. Gaps in the literature

While existing studies have made significant contributions to sentiment and emotion analysis of COVID-19-related tweets, several gaps remain. First, most studies focus on general sentiment and emotions during the pandemic, with limited attention to variant-specific outbreaks, such as Omicron. Second, few studies have explored province- or region-level variations in emotional responses to vaccination, which is essential for targeted public health interventions. Third, there is a lack of research directly correlating emotional intensities in social media discussions with actual vaccination rates. Finally, limited studies have examined real-time emotional reactions during critical periods, such as the emergence of a highly transmissible variant. Our study addresses these gaps by analyzing emotional reactions to vaccination during the Omicron variant outbreak in South Africa, conducting a province-level analysis, and correlating emotional intensities with vaccination rates. By leveraging geotagged Twitter data and advanced emotion classification models, we provide timely insights for health policy-making and interventions.

## 3. Methodology

### 3.1. Inclusion and exclusion criteria for data

Twitter data were collected using the approved academic researcher Twitter API. At the time of data collection, the Twitter API allowed the retrieval of up to 10 million historical tweets per month. Using Python 3.6 and a Twitter access token, we developed scripts to collect historical tweets from the Twitter database. These tweets contained specific keywords related to COVID-19, COVID-19 vaccines, and the Omicron variant.

#### 3.1.1. Inclusion criteria

The inclusion criteria for selecting tweets were based on their relevance to the study topic, as determined by health professionals on our team. The keywords were derived from COVID-19-related trending topics on Twitter in South Africa between October 1, 2021, and January 15, 2022. This period was chosen due to the emergence of the Omicron variant, which led to a significant increase in COVID-19 cases and related discussions. The data collection period was divided into two phases:

- Pre-Omicron (September 30, 2021, to November 23, 2021): The period before the widespread presence of the Omicron variant.
- During Omicron (November 24, 2021, to January 14, 2022): The period when the Omicron variant was prevalent and significantly impacted the population.

This division allowed us to compare emotional dynamics related to vaccination between the two periods. Only English-language tweets were included. For the search keyword code snippet, please refer to Appendix A.

We selected tweets containing geographical information, as shown in Table 1. Using the Twitter API, we retrieved the userID (authorID) for each tweet. A Python script (‘df['userID'].unique()‘) was applied to the dataset to identify unique users. We then extracted the latitude and longitude coordinates from the Geocoordinate (bbox) property of the tweets to determine their precise locations. These locations were manually verified to ensure they were within South Africa.

Additionally, we collected daily COVID-19 vaccination statistics for all provinces before and during the Omicron variant. These data were sourced from the South African coronavirus official website (coronavirus, 2022) and the South Africa COVID-19 dashboard (Covid19SA, 2022), which are updated daily.

#### 3.1.2. Exclusion criteria

To ensure data quality and relevance, we applied several exclusion criteria:

- Tweets without geographical information or originating outside South Africa were excluded, as the study focuses on location-specific data.
- Duplicate tweets were removed to avoid redundancy.
- Bot accounts and retweets were filtered out to eliminate bias and ensure the dataset reflects genuine user interactions.

**Table 1**
Dataset features.

| SN | Attribute | Description | Type |
|---|---|---|---|
| 1 | TweetText | Text content of the tweet representing users' opinions | Text |
| 2 | TweetID | Unique identifier for each tweet | Numeric |
| 3 | CreateDate | Date and time the tweet was posted | Date |
| 4 | RetweetCount | Number of times the tweet was retweeted | Numeric |
| 5 | ReplyCount | Number of replies to the tweet | Numeric |
| 6 | LikeCount | Number of likes the tweet received | Numeric |
| 7 | GeoId | Unique identifier for the tweet's geographical region | Alphanumeric |
| 8 | GeoCityProvince | City and province where the tweet originated | Text |
| 9 | GeoCountry | Country of origin for the tweet | Text |
| 10 | GeoCoordinate(bbox) | Geographical coordinates (latitude and longitude) of the tweet | Numeric |
| 11 | AuthorID | Unique identifier for the user who posted the tweet | Text |
| 12 | UserName | Username of the Twitter account | Alphanumeric |
| 13 | Hashtags | Hashtags included in the tweet | Alphanumeric |
| 14 | CreatedAccountAt | Date and time the user account was created | Datetime |
| 15 | FollowerCount | Total number of followers for the user | Numeric |
| 16 | FollowingCount | Total number of users the account is following | Numeric |
| 17 | TweetCount | Total number of tweets posted by the user | Numeric |

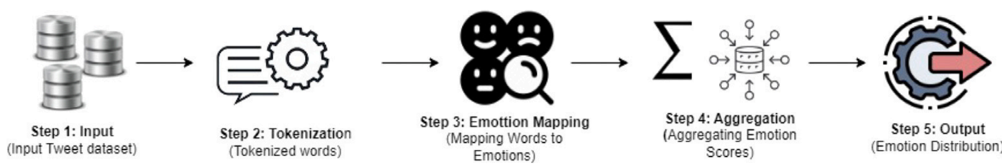


**Fig. 1.** Overview of the text emotion classification process using the Text2Emotion model. The figure highlights key steps, including text preprocessing, emotion detection through lexicon-based analysis, and the final classification of emotions.

### 3.2. Data preprocessing

Raw tweets are often unstructured and contain redundant information. To prepare the data for analysis, we performed several preprocessing steps using Pandas version 1.2.4 Li et al. (2021). These steps included extracting tweets, creation dates, times, and provinces into a dataframe. We then applied the following preprocessing techniques:

- Removed non-English words, incomplete tweets, URLs, punctuation, and quotation marks.
- Eliminated duplicate tweets and stopwords.
- Removed non-alphabetical characters.
- Identified and removed tweets from bot and organizational accounts. Bot accounts were detected by analyzing tweet frequencies, repetitive content, and specific patterns or hashtags associated with bot behavior. Retweets were also excluded to minimize bias.
- Expanded contractions (e.g., "don't" to "do not") and lemmatized words to their base forms.

These steps were implemented using the tweets-preprocessor 0.6.0 (Özcan, 2024) and NLTK 3.6.2 (NLTK Documentation, 2023) libraries. Tokenization was performed using Spacy version 2.3.2 (Aditya, 2022; Honnibal, 2024).

After preprocessing, the dataset comprised 25,293 tweets: 3,613 from the pre-Omicron period and 21,680 from the Omicron period. The tweets were distributed across South African provinces as follows: Eastern Cape (314/1.24%), Free State (389/1.54%), Gauteng (6,977/27.85%), KwaZulu-Natal (5,033/19.90%), Limpopo (597/2.36%), Mpumalanga (4,823/19.10%), North West (729/2.88%), Northern Cape (336/1.33%), and Western Cape (6,095/24.10%).

### 3.3. Emotion analysis

To analyze emotions in the tweets, we used Text2Emotion version 0.0.5 (Gupta, Band, Karan, Sharma, & Deepspacehub, 2020), a Python package that identifies five emotion categories: Happy, Angry, Sadness, Surprise, and Fear. Fig. 1 illustrates the emotion classification process.

The process involves:

- Input: Loading preprocessed tweets discussing COVID-19, vaccines, and Omicron.
- Tokenization: Breaking tweets into individual words or word pairs.
- Emotion Mapping: Comparing tokenized words against an internal dataset to identify associated emotions.
- Aggregation: Calculating the proportion of each emotion in the text, termed emotional intensity.
- Output: Generating a dictionary detailing the proportions of each emotion.

Emotional intensity is calculated using Eq. (1):

$$\text{Emotional Intensity} = \frac{\text{Count of emotional class}}{\text{Total tokenized word count}} \tag{1}$$

The intensity score ranges from 0 (lowest) to 1 (highest).

### 3.4. Tweet labeling

Tweets were automatically labeled using Text2Emotion, as described in Section 3.3. To validate the accuracy of these labels, 10% of the tweets (2,530 tweets) were randomly selected and manually labeled. This sample included an equal distribution of 2% (506 tweets) for each emotion category.

To ensure reliability, four human annotators independently labeled the tweets. Consensus was achieved through majority voting, where the final label was determined by the majority opinion. The manually labeled tweets were then compared with the Text2Emotion-labeled tweets using tweet IDs.

Additionally, we annotated the tweets using transformer-based models, specifically BERT (Wu, Jin, Shi, Liang, & Zhan, 2024) and RoBERTa (Kamath, Ghoshal, Eswaran, & Honnavalli, 2022), which were fine-tuned for our dataset. The Chi-square test was used to compare the distributions of predicted labels from Text2Emotion, manual labeling, and the best-performing transformer model. Results are presented in the Results section.

### 3.5. Topic modeling

We employed Latent Dirichlet Allocation (LDA) for topic modeling. LDA is a probabilistic model that identifies topics within a collection of documents by assuming each document is a mixture of topics and each word is attributed to a topic with a certain probability (Blei, Ng, & Jordan, 2003; Ogbuokiri, Ahmadi, Mellado, et al., 2023).

Topics were selected based on their Jaccard similarity score, calculated using the `sklearn.metrics.jaccardscore` package in Python (GeeksForGeeks, 2023; Scikit Learn, 2024). The Jaccard similarity between two topics with representative keyword sets $A$ and $B$ is given by:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{2}$$

where:

- $J(A,B)$ is the Jaccard similarity coefficient.
- $|A \cap B|$ is the number of common elements.
- $|A \cup B|$ is the total number of unique elements.

### 3.6. Statistical analysis

#### 3.6.1. Cohen's Kappa test

We used Cohen's Kappa to measure agreement between manual and Text2Emotion labels (Tal, 2025). The Kappa score ranges from −1 to 1, with 1 indicating perfect agreement. The score was calculated using the `cohen_kappa_score` function from the `sklearn.metrics` library. The formula is:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

where:

- $\kappa$ is the Cohen's Kappa score.
- $P_o$ is the observed agreement.
- $P_e$ is the expected agreement by chance.

#### 3.6.2. Granger causality test

To analyze the relationship between emotional intensities and vaccination rates, we performed Granger causality tests using the `statsmodels` library (Mighri, Ragoubi, Sarwar, & Wang, 2022; Rosol, Młyńczak, & Cybulski, 2022). This test determines whether changes in emotional intensities precede changes in vaccination rates, suggesting a causal influence.

#### 3.6.3. Pearson correlation

We explored the correlation between vaccine-related tweets and vaccination rates across South African provinces using the `scipy.stats.pearsonr` function (Li, Zhang, Zhang, Wu, & Zhan, 2022).

#### 3.6.4. Mann–Whitney U test

To compare emotional intensities between groups, we used the Mann–Whitney U test from the `scipy.stats.mannwhitneyu` package (Bedre, 2020). This non-parametric test identifies differences in emotional expressions between independent groups.

These analyses provided insights into the relationships between vaccine-related tweets, vaccination rates, and emotional expressions. The results are discussed in detail in the subsequent sections.

## 4. Results

This section presents the findings of our research, organized into three parts. The first part provides summary statistics of the collected data. The second part focuses on the classification of emotions in hand-labeled tweets and compares these results with those obtained from Text2Emotion-labeled tweets. Additionally, we conduct a topic analysis of these labeled tweets using the LDA model. The third part examines the impact of COVID-19-related discussions on Twitter regarding vaccination by province, both before and during the Omicron variant period. This analysis highlights how online conversations influenced vaccination behaviors across South African provinces. By employing advanced methodologies and extensive data analysis, our research addresses critical questions related to emotions on social media and explores the interplay between Twitter discussions and vaccination patterns during the Omicron era.

### 4.1. Data collection

We collected a total of 28,000 tweets from 14,100 unique users across South Africa using the Twitter API's archive search feature. Of these, 26,188 tweets included geographical information, as shown in Table 1. The GeoCoordinate (bbox) property of the tweets was used to extract latitude and longitude coordinates. The bounding box (bbox) is defined by two sets of latitude and longitude coordinates, representing the southwest and northeast corners. To estimate the exact location of each tweet, we calculated the average of these coordinates. This process ensured the accuracy of the geographical data for the 14,000 unique users.

The dataset was divided into two periods:

- Pre-Omicron: 4,440 tweets from 2,500 unique users.
- During Omicron: 23,560 tweets from 11,600 unique users.

Each tweet contained features such as `TweetText`, `CreatedDate`, `GeoCityProvince`, and `GeoCoordinate`. The trends in tweets over time and their distribution across provinces are shown in Figs. B.7 and B.8 of Appendix B.

### 4.2. Tweet emotion classification

This section compares the results of hand-labeled emotions with those from Text2Emotion. Our analysis revealed minimal variation in the distribution of hand-labeled tweets compared to Text2Emotion-labeled tweets, indicating strong consistency between the two methods. Tables 2 and 3 provide examples of misclassified tweets, along with the associated keywords and their weights.

To further validate the comparison between hand-labeled and Text2Emotion-labeled datasets, we performed a Cohen's Kappa test using the contingency table (see Table 4). The Kappa score of 0.92 indicates almost perfect agreement between the two labeling methods, demonstrating the reliability of Text2Emotion in replicating human judgment.

The results from Table 4 show an average accuracy of 93.5%, with 92.88% of Angry, 91.89% of Fear, 97.03% of Happy, 93.08% of Sad, and 92.68% of Surprise tweets correctly classified. The high Cohen's Kappa score (0.92) confirms the reliability of Text2Emotion for emotion classification in this context.

We further evaluated transformer-based models, including BERT and RoBERTa, to validate our results. Fine-tuned BERT achieved an overall accuracy of 95%, while fine-tuned RoBERTa achieved 55% accuracy (see Table 5). A comparison between fine-tuned BERT and Text2Emotion revealed statistically significant similarity (p = 0.002), supporting the robustness of our findings.

**Table 2**
Comparison of hand-labeled tweets and Text2Emotion-labeled tweets: sample distribution of misclassified tweets.

| SN | Original tweet | Hand label | Text2Emotion label |
|---|---|---|---|
| 1 | The Omicron variant is spreading so fast, it is terrifying! Stay safe everyone and follow the guidelines! | Fear | Angry |
| 2 | Just heard about the first case of Omicron in my city. I am so scared right now! | Fear | Sad |
| 3 | The government's slow response to the Omicron variant is infuriating. We need action now! | Angry | Fear |
| 4 | Sad to see the impact of the Omicron variant on businesses. The government must step in to support them. | Sad | Angry |
| 5 | Feeling hopeful as the government ramps up vaccination efforts. Together, we can beat Omicron! | Happy | Surprise |

**Table 3**
Emotional keywords of misclassified tweets and their weights by Text2Emotion.

| S/N | Angry | | Fear | | Happy | | Sad | | Surprise | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Keyword | Weight | Keywords | Weights | Keywords | Weights | Keywords | Weights | Keywords | Weights |
| 1 | irate | 0.701 | terrified | 0.453 | joyful | 0.420 | sorrow | 0.412 | astonishment | 0.771 |
| 2 | furious | 0.472 | infuriating | 0.451 | delighted | 0.382 | grief | 0.410 | amazement | 0.637 |
| 3 | outraged | 0.405 | frightened | 0.441 | cheerful | 0.372 | scared | 0.409 | shock | 0.420 |
| 4 | terrifying | 0.347 | jab | 0.416 | pleased | 0.338 | heartache | 0.378 | ecstatic | 0.401 |
| 5 | infuriated | 0.305 | incensed | 0.389 | content | 0.268 | headache | 0.354 | hopeful | 0.345 |
| 6 | afraid | 0.272 | dreadful | 0.386 | startle | 0.256 | depression | 0.320 | stun | 0.349 |
| 7 | enraged | 0.205 | appalled | 0.251 | elated | 0.246 | misery | 0.262 | dismay | 0.340 |
| 8 | wrathful | 0.201 | horrified | 0.220 | overjoyed | 0.236 | suffering | 0.219 | disbelief | 0.333 |
| 9 | spreading | 0.200 | fearful | 0.201 | thrilled | 0.226 | anguish | 0.210 | wonder | 0.309 |
| 10 | omicron | 0.177 | apprehensive | 0.165 | blissful | 0.215 | disaster | 0.129 | jolt | 0.212 |

**Table 4**
Contingency table of the distribution of hand label vs. Text2Emotion label.

| | | Text2Emotion Label (Predicted) | | | | | |
|---|---|---|---|---|---|---|---|
| | | Angry | Fear | Happy | Sad | Surprise | Total |
| Hand label (Actual) | Angry | 470 | 16 | 0 | 12 | 8 | 506 |
| | Fear | 20 | 465 | 15 | 5 | 1 | 506 |
| | Happy | 0 | 0 | 491 | 0 | 15 | 506 |
| | Sad | 9 | 26 | 0 | 471 | 0 | 506 |
| | Surprise | 7 | 13 | 17 | 0 | 469 | 506 |
| | Total | 506 | 520 | 523 | 488 | 493 | **2530** |

**Table 5**
Pre-trained transformer-based model performance on our dataset.

| | BERT | | | | | RoBERTa | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Class | Angry | Fear | Happy | Sad | Surprise | Angry | Fear | Happy | Sad | Surprise |
| Precision | 0.40 | 0.53 | 0.33 | 0.44 | 0.42 | 0.34 | 0.25 | 0.22 | 0.45 | 0.22 |
| Recall | 0.41 | 0.52 | 0.40 | 0.45 | 0.43 | 0.27 | 0.27 | 0.26 | 0.30 | 0.41 |
| F1-score | 0.31 | 0.22 | 0.43 | 0.27 | 0.35 | 0.33 | 0.21 | 0.18 | 0.22 | 0.30 |
| Accuracy | | | | | 0.45 | | | | | 0.37 |
| | Fine-Tuned BERT | | | | | Fine-Tuned RoBERTa | | | | |
| Class | Angry | Fear | Happy | Sad | Surprise | Angry | Fear | Happy | Sad | Surprise |
| Precision | 0.94 | 0.97 | 0.95 | 0.91 | 0.96 | 0.45 | 0.55 | 0.47 | 0.53 | 0.48 |
| Recall | 0.92 | 0.87 | 0.91 | 0.97 | 0.91 | 0.50 | 0.52 | 0.60 | 0.54 | 0.50 |
| F1-score | 0.90 | 0.88 | 0.92 | 0.89 | 0.91 | 0.44 | 0.56 | 0.47 | 0.53 | 0.54 |
| Accuracy | | | | | 0.95 | | | | | 0.55 |

### 4.3. Feasible alternatives and advantages of Text2Emotion

In this study, we adopted the Text2Emotion lexicon-based approach for emotion classification due to its computational efficiency, simplicity, and strong agreement with human-labeled data. However, several feasible alternatives exist, each with its own advantages and limitations. Lexicon-based approaches, such as Text2Emotion, are lightweight, interpretable, and do not require labeled training data, making them ideal for large-scale text analysis (Mohammad & Turney, 2013). However, they may struggle with slang, sarcasm, or domain-specific language, as highlighted by Kiritchenko, Zhu, and Mohammad (2014). Machine learning models, such as Support Vector Machines (SVM) and Random Forest, can capture complex patterns in data but require extensive labeled data and computational resources for training (Pang, Lee, & Vaithyanathan, 2002). Transformer-based models, such as BERT (Devlin, Chang, Lee, & Toutanova, 2019) and RoBERTa (Liu et al., 2019), achieve state-of-the-art performance in text classification tasks but are computationally intensive and require expertise for fine-tuning. Hybrid approaches, which combine the strengths of lexicon-based and machine learning methods, offer a potential middle ground but are more complex to implement and may still require significant computational resources (Zhang, Zhao, & LeCun, 2018).

The choice of Text2Emotion was driven by its ability to provide reliable results with minimal computational overhead. As demonstrated by the high Cohen's Kappa score (0.92) and comparison with fine-tuned BERT ($p = 0.002$), Text2Emotion performs comparably to more resource-intensive models in this context. Its simplicity and interpretability make it an ideal choice for rapid analysis of large datasets, particularly when computational resources are limited (Cambria, Poria, Hazarika, & Kwok, 2018). However, it is important to note that the generalizability of these findings may vary across datasets and domains. Future work could explore hybrid approaches or evaluate the performance of Text2Emotion on diverse datasets to further validate its applicability (Wang, Huang, Zhu, & Zhao, 2020).

Having explored the emotion classification aspect, we now shift our focus to examining discussions about COVID-19 on Twitter before and during the Omicron outbreak in South Africa. This investigation aims to gain insights into how the online discourse surrounding the pandemic evolved during this critical period.

### 4.4. Topic generation and analysis

Using our LDA model, we generated 40 topics from the labeled tweet dataset. During the topic modeling process, we observed that several topics shared overlapping words and lacked distinct themes. To enhance the quality of our topics, we calculated the coherence score for each using the `CoherenceModel` class, employing the `c_v` metric. This metric evaluates the semantic coherence of topics based on word co-occurrence within the corpus. Upon evaluation, Topics 1 to 5 achieved high coherence scores of 0.85, 0.86, 0.87, 0.83, and 0.82, respectively, indicating that these topics were coherent and distinct in capturing meaningful themes from the dataset. In contrast, many other topics scored below 0.5, suggesting lower coherence and less clarity in topic representation. Therefore, we selected Topics 1 to 5 for further analysis and interpretation, focusing on their robust thematic coherence and interpretability. This approach ensured that the topics derived from our LDA model were not only numerous but also meaningful and representative of distinct themes present in the labeled tweet dataset.

Additionally, we employed the Jaccard similarity measure to compare the similarity between two topics. The Jaccard similarity between two topics is defined as the size of the intersection of the topics' words or tokens divided by the size of the union of the topics' tokens. The Jaccard similarity score ranges from 0 to 1, where 0 indicates no similarity (the two topics have no words in common) and 1 indicates complete similarity (the topics are identical). We selected topics with a Jaccard similarity score of $\leq 0.5$ (see Fig. 2). The yellow color gradient along the diagonal indicates that the topics are similar in terms of the words or tokens they contain, while the blue color indicates dissimilarity. The darker the blue color gradient, the more dissimilar the topics are. See Fig. 2 for reference.

For each topic, we summarized the number of tweets and the percentage of tweets from the dataset that discussed the topic over time (see Fig. 3).

Moreover, we prepared a table and figures summarizing the top five LDA-generated topics. This table includes representative tweets with corresponding emotions, keywords, and the location of these tweets. The detailed information can be found in Table 6.

### 4.5. COVID-19 vaccine-related discussions on twitter before and during omicron

In this section, we examine discussions about COVID-19 vaccines on Twitter before and during the Omicron outbreak in South Africa. The word "Omicron" was one of the most frequently used terms in vaccine-related tweets during the outbreak. The trends in tweets over time, both before and during Omicron, are presented in Figs. B.8 of Appendix B.

Additionally, the total emotional intensities of COVID-19 vaccine-related discussions on Twitter over time are shown in Fig. 4. Emotional intensity was calculated by summing the weights of the emotional classes identified in each tweet. Each emotional class has an associated weight that reflects its intensity in conveying a specific emotion. By aggregating these weights across the dataset, we obtained the total emotional intensity, illustrating how users' emotions evolved over time during the outbreak of the new variant in South Africa.

Before the Omicron outbreak, we observed that all emotional intensities peaked during the week of October 08, 2021, and then gradually declined until the week of November 22, 2021 (see Fig. 4(a)). Similarly, during the Omicron outbreak, emotional intensities reached their highest levels in the week of December 27, 2021 (see Fig. 4(b)). This period coincided with the surge in Omicron cases, which likely triggered widespread discussions and heightened emotions on social media platforms, particularly Twitter.

Next, we examine the emotional reactions to vaccine-related discussions before the Omicron outbreak.

#### 4.5.1. Emotional reactions toward vaccine-related discussions before omicron

In this section, we analyze the emotional reactions expressed in COVID-19 vaccine-related discussions on Twitter and their relationship with vaccination rates in South Africa before the emergence of the Omicron variant. The vaccination data used in this analysis pertains to individuals aged 18 to 34 years, as this age group is highly active on social media, particularly Twitter (Statista, 2024). The study period spans from October 1, 2021, to November 22, 2021, focusing on the months leading up to the Omicron outbreak. We used the Granger causality test to assess the relationship between emotional intensities in tweets and vaccination rates across South African provinces. Our goal was to determine whether emotional intensities in tweets are associated with vaccination rates.

We formulated the following hypotheses:

**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.

Granger causality suggests that emotional intensities in tweets may predict future vaccination rates. The Granger causality test produces an F-test statistic and a corresponding $p$-value ($p$). If the $p$-value is below the significance level ($\alpha = 0.05$), we reject $H_0$, indicating that emotional intensities Granger-cause vaccination rates over time. This implies an association between emotional intensity and vaccination.

The results of the Granger causality test revealed statistically significant evidence for Mpumalanga (F-test = 2.6939, $p = 0.0011$) and Western Cape (F-test = 30.4013, $p = 0.0053$), indicating an association between emotional intensities and vaccination rates before the Omicron outbreak. Despite the relatively low number of tweets in these provinces, we reject the null hypothesis ($H_0$) and conclude that emotional intensities in tweets may predict vaccination uptake in these regions. Refer to Fig. B.9 of Appendix B.

We further analyzed the variations in emotional classes before the Omicron outbreak for these two provinces. In Mpumalanga, the Angry emotional class ($p = 0.04$) showed a significant association with vaccination. In Western Cape, both Angry ($p = 0.01$) and Fear ($p = 0.005$) emotional classes were significantly associated with vaccination. These findings suggest that Angry and Fear emotions in vaccine-related tweets may predict vaccination uptake in Mpumalanga and Western Cape, particularly during an outbreak. See Fig. B.10 of Appendix B.

For the remaining seven provinces, we found no statistically significant evidence to reject $H_0$. Table 7 summarizes the results.

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

#### 4.5.2. Emotional reactions toward vaccine-related discussions during the omicron outbreak

In this section, we analyze the emotional reactions expressed in COVID-19 vaccine-related discussions during the Omicron outbreak and their relationship with vaccination rates. The vaccination data pertains to individuals aged 18 to 34 years. To measure the increase in vaccination uptake, we calculated the difference between daily vaccination totals before and during the Omicron outbreak for all provinces.

We first examined the correlation between daily tweets and the increase in daily vaccination rates across provinces. The results revealed diverse patterns of association, as shown in Fig. 5(a)–(i).

In Eastern Cape (Fig. 5(a)), a moderate positive correlation (corr = 0.504, $p < 0.001$) suggests that an increase in tweets corresponds to a moderate increase in vaccinations. In Free State (Fig. 5(b)), a strong positive correlation (corr = 0.630, $p < 0.001$) indicates a significant increase in vaccinations with more tweets. Gauteng (Fig. 5(c)) shows

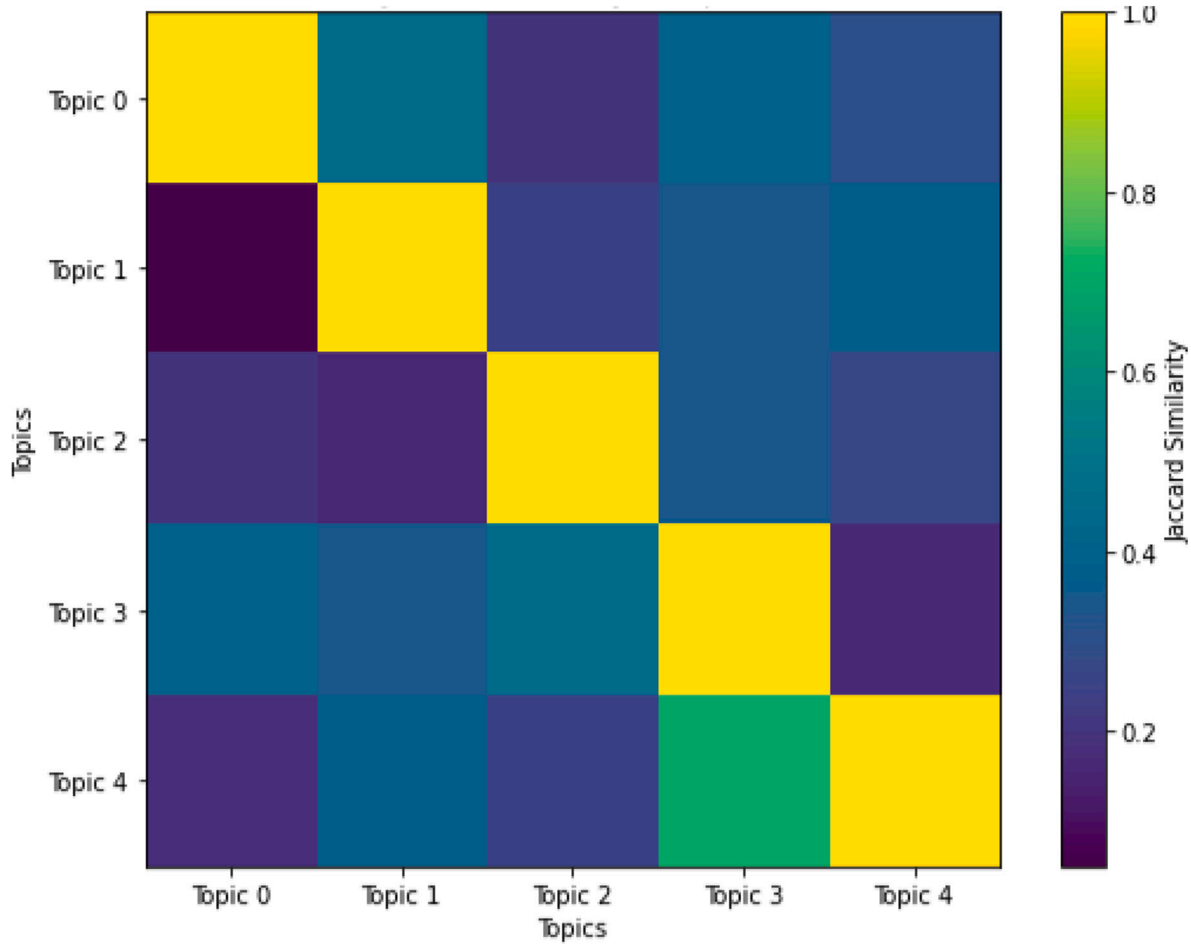


**Fig. 2.** Jaccard similarity of topics.

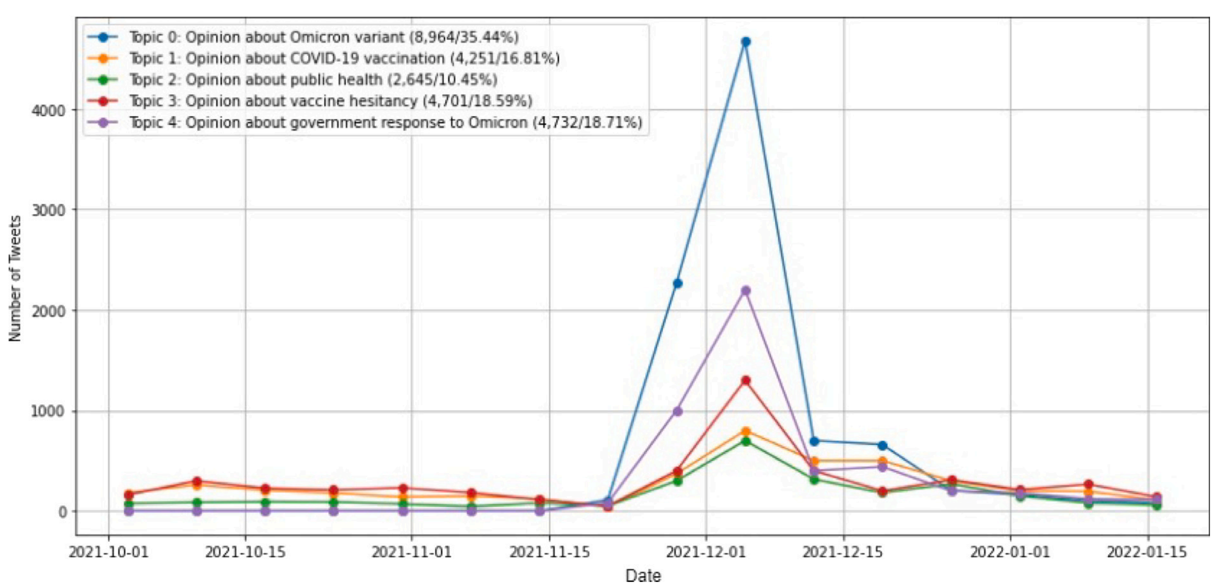


**Fig. 3.** Weekly tweets of the top 5 topics with respect to time (N=25,293).

a strong positive correlation (corr = 0.723, $p$ < 0.001), highlighting a robust association between tweet volume and vaccination rates.

In KwaZulu-Natal and Mpumalanga, the correlations are moderate but significant. KwaZulu-Natal shows a moderate positive correlation (corr = 0.382, $p$ = 0.005), while Mpumalanga has a similar correlation (corr = 0.378, $p$ = 0.006) (Figs. 5(d) and (e)). In contrast, Northern Cape shows almost no correlation (corr = −0.006, $p$ = 0.968), indicating that tweet volume does not significantly impact vaccination rates (Fig. 5(f)).

Limpopo, North West, and Western Cape exhibit strong or moderate positive correlations. Limpopo (corr = 0.707, $p$ < 0.001) and North West

**Table 6**
LDA-generated topics and representative tweets with emotions and keywords.

| Topics | Keywords | Representative tweets | Emotions | Tweet location |
|---|---|---|---|---|
| 0: Opinion about Omicron variant | omicron, variant, spreading, terrified, safe, guidelines, first case, city, scared, believe, quickly, nightmare, anxious, pandemic, worried | "The Omicron variant is spreading so fast, it's terrifying! Stay safe everyone and follow the guidelines!" | Fear | Gauteng |
| | | "Just heard about the first case of Omicron in my city. I'm so scared right now!" | Fear | Gauteng |
| | | "I can't believe how quickly the Omicron variant is spreading. It's like a nightmare come true!" | Fear | Kwazulu Natal |
| | | "Feeling anxious about the Omicron variant taking over. Will we ever be safe from this pandemic?" | Sad | Mpumalanga |
| | | "The news about Omicron is giving me serious anxiety. I'm worried about what's to come." | Sad | Western Cape |
| 1: Opinion about COVID-19 vaccination | booster shot, relieved, protect, omicron threat, vaccination, angry, refuse, vaccinated, vulnerable, healthcare workers, vaccines, safe | "Just got my booster shot and feeling relieved! It's essential to protect ourselves and others from the Omicron threat." | Happy | Western Cape |
| | | "I'm angry at those who refuse to get vaccinated. It's not just about us, but the vulnerable ones too!" | Anger | Gauteng |
| | | "Feeling hopeful and grateful for the healthcare workers administering vaccines. Thank you for keeping us safe." | Happy | Western Cape |
| | | "My heart breaks seeing the rising cases. We need more people to get vaccinated to fight the Omicron wave." | Sad | Limpopo |
| | | "Got my second dose today! Feeling happy and more protected against the Omicron variant." | Happy | Northwest |
| 2: Opinion about public health | social media, flooded, updates, essential, verify information, fear, panic, twitter, health authorities, communication game, misinformation | "Social media is flooded with updates about the Omicron variant. It's essential to verify information from the SA command council before sharing!" | Fear | Western Cape |
| | | "Seeing the fear and panic on Twitter due to Omicron. Health authorities must step up their communication game!" | Fear | Gauteng |
| | | "Misinformation about COVID-19 vaccines is rampant on social media. We need better regulation to combat this." | Anger | Mpumalanga |
| | | "The power of social media in spreading awareness is evident, but we must also tackle the anxiety it generates." | Fear | Kwazulu Natal |
| | | "Social media can be both a blessing and a curse during a pandemic. Let's use it responsibly to promote accurate information." | Surprise | Gauteng |
| 3: Opinion about vaccine hesitancy | effects, vaccine, hesitancy, changing, covid19vaccine, doubt, effectiveness, fear, hesitant, transparency, awareness | "I'm scared of side effects from the vaccine. Can't decide whether to get it or not." | Fear | Free State |
| | | "The constant changing information about vaccines makes me doubt their effectiveness." | Fear | Northern Cape |
| | | "Angry at the government for not providing enough information about the vaccine. Transparency is crucial!" | Anger | Gauteng |
| | | "Feeling relieved after getting vaccinated. The fear of Omicron pushed me to finally take the step." | Happy | Kwazulu Natal |
| | | "Surprised to see people still hesitant about vaccines despite the rising cases. Let's protect ourselves and others!" | Surprise | Western Cape |
| 4: Opinion about government response to Omicron | government, slow, response, infuriating, action, measures, spread, cooperate, public safety, covid hoax, support, economic, ramps, omicron | "The government's slow response to the Omicron variant is infuriating. We need action now!" | Anger | Kwazulu Natal |
| | | "Happy to see the government implementing strict measures to curb the spread of Omicron. Let's cooperate!" | Happy | Gauteng |
| | | "Surprised that some people still believe COVID is a hoax. We need better public education!" | Surprise | Western Cape |
| | | "Sad to see the impact of the Omicron variant on businesses. The government must step in to support them." | Sad | Mpumalanga |
| | | "Feeling hopeful as the government ramps up vaccination efforts. Together, we can beat Omicron!" | Happy | Northwest |

(corr = 0.641, $p < 0.001$) show strong correlations, while Western Cape shows a moderate correlation (corr = 0.375, $p = 0.006$). These findings suggest that social media activity, as measured by tweet volume, generally correlates with vaccination rates, with varying degrees of strength across provinces.

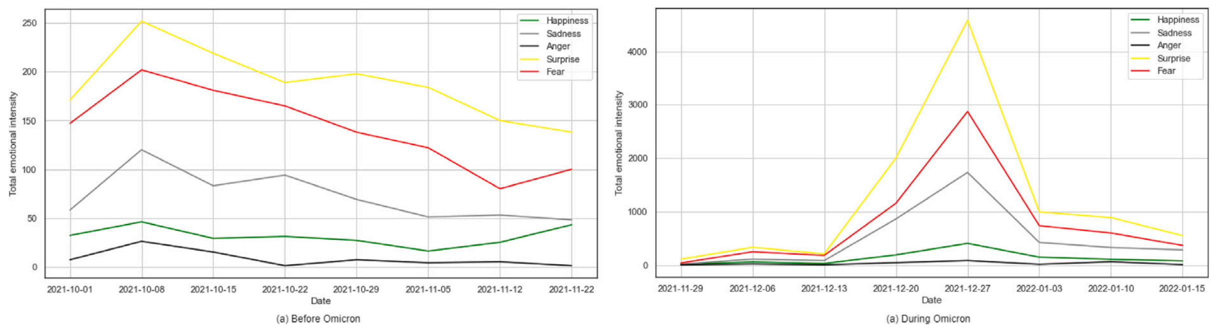


**Fig. 4.** Emotional changes expressed in tweets.

**Table 7**
Outcome of the association between emotional intensities and vaccination before omicron for the remaining seven provinces.

| SN | Province | F-test | $p$ |
|---|---|---|---|
| 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 | Northern Cape | 1.1014 | 0.3532 |

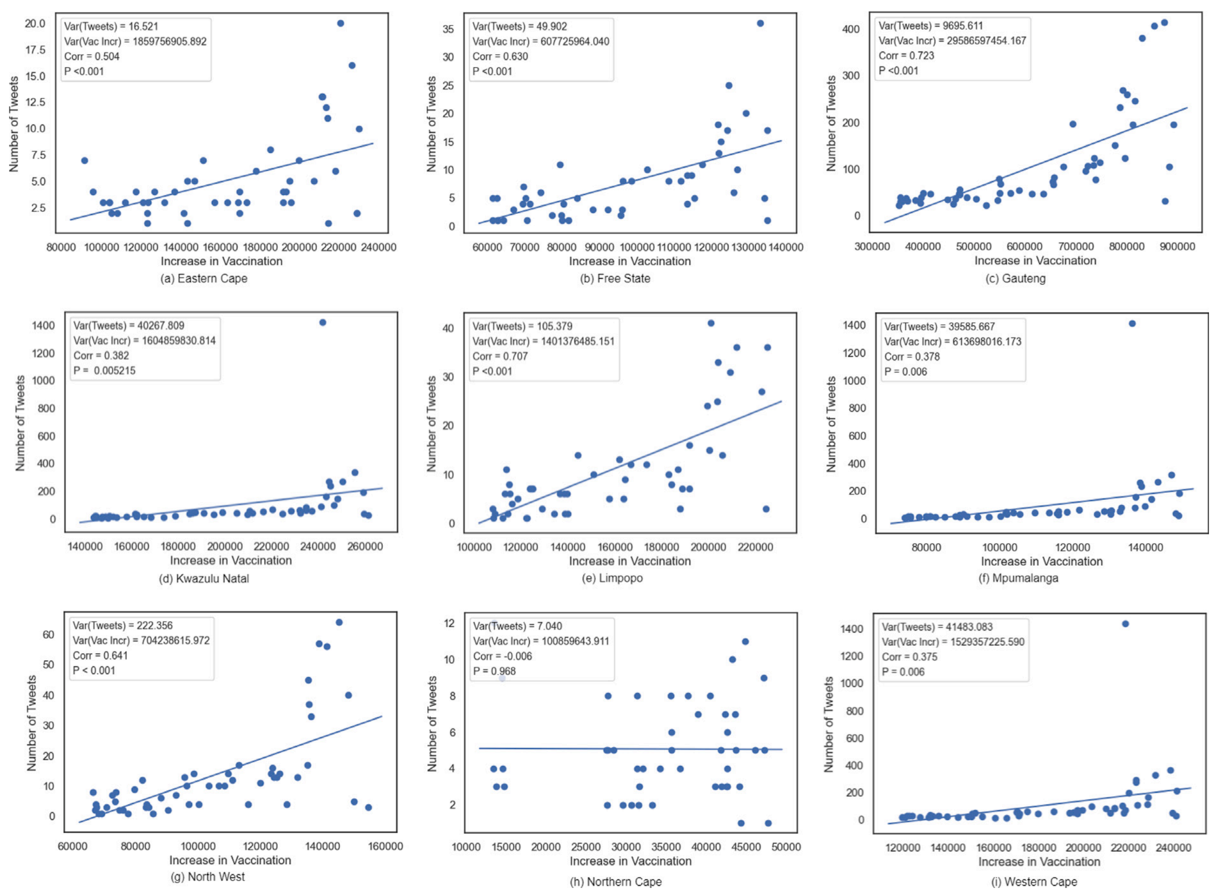


**Fig. 5.** Association of daily tweets with the increase in daily vaccination rates by province in South Africa.

The Granger causality test during the Omicron outbreak revealed statistically significant evidence for Eastern Cape (F-test = 0.8125, $p$ = 0.0115), Gauteng (F-test = 0.1625, $p$ = 0.0001), Limpopo (F-test = 1.2898, $p$ = 0.0078), and North West (F-test = 1.0572, $p$ = 0.3623). Based on these results, we reject the null hypothesis ($H_0$) and conclude that emotional intensities in tweets may influence vaccination uptake in these provinces. See Fig. B.11 of Appendix B.

We further analyzed the variations in emotional classes for these provinces. In Eastern Cape, Angry ($p$ = 0.001) and Surprise ($p$ = 0.001) emotions were significantly associated with vaccination. In Gauteng, all

**Table 8**
Outcome of the association between emotional intensities and vaccination during omicron for the remaining five provinces.

| SN | Province | F-test | $p$ |
|---|---|---|---|
| 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 |

emotional classes (Happy, Sadness, Angry, Surprise, and Fear) showed significant associations ($p < 0.001$). In Limpopo, Sadness ($p = 0.0009$), Surprise ($p < 0.001$), and Fear ($p = 0.001$) were significant. In North West, Angry ($p = 0.02$), Surprise ($p < 0.001$), and Fear ($p = 0.0001$) were significant. These results suggest that emotional classes in tweets may predict vaccination uptake for individuals aged 18 to 34 years in these provinces during an outbreak. See Fig. B.12 of Appendix B.

For the remaining five provinces, we found no statistically significant evidence to reject $H_0$. Table 8 summarizes the results.

#### *4.5.3. Comparison of emotional classes by provinces during the omicron outbreak*

In this section, we compare the emotional classes identified in COVID-19 vaccine-related discussions before and during the Omicron outbreak. The goal is to determine whether the emotional classes before and during Omicron are associated, suggesting that the same external factors may have influenced them. This information is valuable for health policy-making, particularly in managing vaccine hesitancy.

We used the Mann–Whitney U test to compare emotional classes for each province. Since provinces are treated independently, the Mann–Whitney U test is suitable for comparing independent variables. We formulated the following hypotheses:

**Hypothesis Two**

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

The Mann–Whitney U test produces a $p$-value for each emotional class in each province. To account for multiple testing, we applied the Bonferroni correction ($\alpha_{\text{adjusted}} = 0.01$). If any $p$-value falls below this threshold, we reject $H_0$, indicating an association between emotional intensities before and during Omicron. Fig. 6 summarizes the results.

The results show statistically significant evidence after applying the Bonferroni correction: - In Eastern Cape, only Surprise ($p = 0.04$) was significant but did not meet the adjusted threshold ($\alpha_{\text{adjusted}} = 0.01$). Thus, we do not reject $H_0$. - In Free State and North West, Happy ($p \leq 0.01$), Sadness ($p \leq 0.03$), Surprise ($p \leq 0.03$), and Fear ($p \leq 0.02$) were significant. After correction, Happy ($p \leq 0.01$) in Free State and North West, and Sadness ($p \leq 0.001$) in Northern Cape, were significant. Thus, we reject $H_0$ for these emotional classes. - In Limpopo, Happy ($p = 0.03$) and Surprise ($p = 0.04$) were significant but did not meet the adjusted threshold. Thus, we do not reject $H_0$. - In Northern Cape, Sadness ($p \leq 0.01$) was significant, so we reject $H_0$.

For Gauteng, KwaZulu-Natal, Mpumalanga, and Western Cape, no emotional classes met the adjusted threshold ($\alpha_{\text{adjusted}} = 0.01$). Thus, we do not reject $H_0$.

Finally, our results address vaccine hesitancy by examining its impact on emotional classes before and during the Omicron outbreak. The analysis reveals statistically significant evidence of emotional classes such as Fear, Sadness, and Surprise correlating with vaccination rates in provinces like Eastern Cape, Free State, North West, Limpopo, and Northern Cape during Omicron. These findings suggest that vaccine hesitancy may have contributed to the observed emotional responses, with the increase in Fear in Eastern Cape indicating potential hesitancy or concern. The presence of significant emotional classes like Sadness and Surprise in other provinces also suggests underlying vaccine hesitancy issues.

### *4.6. Assumptions and justifications*

In this section, we discuss the key assumptions underlying our analysis, provide justifications for their use, and evaluate their potential impact on the results.

1. Independence Between Provinces:
   - The `Mann–Whitney U` test assumes that emotional intensities across provinces are independent. This assumption is justified by the distinct demographic and socio-economic characteristics of each province, which influence public sentiment independently.
   - While violations of this assumption could affect the significance of the results, the diversity of South African provinces supports its validity.
2. Linear Granger Causality:
   - The Granger causality test assumes a linear relationship between emotional intensities and vaccination rates. This assumption is supported by prior studies demonstrating linear relationships between public sentiment and health behaviors (Statista, 2024).
   - Although nonlinear relationships could reduce the test's sensitivity, the strong statistical significance observed in key provinces suggests that the assumption holds.
3. Homogeneity in Emotional Expressions:
   - We assume that emotional expressions in tweets are homogeneous within each province. This assumption is supported by consistent language and sentiment patterns observed in regional social media data (Cambria et al., 2018).
   - While variations within provinces could affect generalizability, our province-level analysis minimizes this impact.
4. No Confounding Factors:
   - We assume that no external factors significantly influenced emotional intensities and vaccination rates during the study period. This assumption is justified by the controlled time frame of the analysis, which focuses on specific periods before and during the Omicron outbreak.
   - Although confounding factors could bias the results, the short study duration reduces this risk.
5. Bonferroni Correction for Multiple Testing:
   - The Bonferroni correction is applied to control for Type I errors in multiple hypothesis tests. While this conservative approach reduces the risk of false positives, it may also decrease the likelihood of detecting true associations.
   - Despite this limitation, our analysis identified significant emotional classes in several provinces, demonstrating the robustness of the findings.

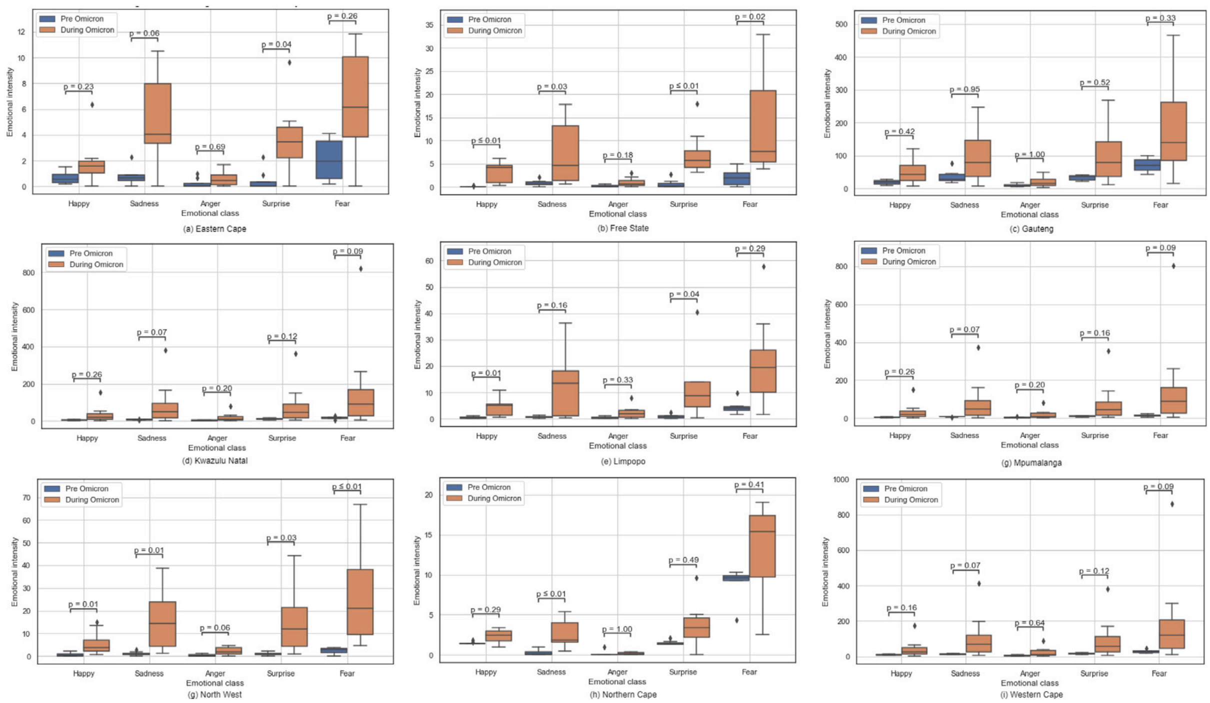

**Fig. 6.** Comparison of emotional intensities before and during omicron by province using the Mann–Whitney U Test.

## 5. Discussion

### 5.1. Findings

In this study, we collected and analyzed COVID-19 vaccine-related tweets from South Africa between October 1, 2021, and January 15, 2022, using the Twitter API. The data were divided into two periods: before the Omicron variant (October 1–November 22, 2021) and during the Omicron variant (November 23, 2021–January 15, 2022). As expected, during the Omicron period, there was a notable increase in daily tweets from the week of November 23, 2021, to the week of December 1, 2021 (see Fig. B.7 in Appendix B). However, after this initial surge, there was a sharp decline in daily tweets, with the curve remaining relatively flat until January 15, 2022. This decline may be attributed to the fact that the Omicron variant, while causing more infections, resulted in fewer hospital admissions compared to the Delta variant (Dyer, 2021; Lulan & Genhong, 2022). Consequently, many South Africans may have shown less interest in vaccination (coronavirus, 2022; Covid19SA, 2022; Lulan & Genhong, 2022), potentially leading to a decline in COVID-19 vaccine-related discussions on Twitter (Abdullah et al., 2022; Dyer, 2021).

For emotion classification, we employed the Text2Emotion pre-trained model, which categorized tweets into five emotional classes: Happy, Sadness, Angry, Surprise, and Fear, along with their corresponding emotional intensities. This approach aligns with prior studies (Aslam, Rustam, Lee, Washington, & Ashraf, 2022; coronavirus, 2022; Díaz, Shaik, Santofimio, & Quintero M., 2018), where Text2Emotion proved effective in detecting emotions across various types of tweets. Our study further validates the suitability of Text2Emotion for this type of analysis. To ensure the accuracy of the Text2Emotion labels, we compared them with a hand-labeled 10% sample of the dataset and transformer-based models (BERT and RoBERTa). The results demonstrated statistically significant agreement, with minimal variation in the distribution of labels, confirming the reliability of Text2Emotion.

The effectiveness of Text2Emotion in this study can be attributed to several factors. First, its lexicon-based approach allows for efficient and interpretable emotion classification without requiring extensive labeled training data. This makes it particularly suitable for large-scale text analysis, such as the 28,000 tweets in our dataset. Second, Text2Emotion's ability to classify emotions into distinct categories (Happy, Sadness, Angry, Surprise, and Fear) aligns well with the emotional dynamics observed in vaccine-related discussions. While more advanced models like BERT and RoBERTa achieved higher accuracy after fine-tuning, Text2Emotion provided comparable results with significantly lower computational overhead. This balance between accuracy and efficiency makes Text2Emotion a practical choice for real-time sentiment analysis during rapidly evolving situations, such as the Omicron outbreak.

Prior to the emergence of the Omicron variant, topic modeling using LDA revealed that in Gauteng, Mpumalanga, and Western Cape provinces, tweets predominantly discussed three topics: (1) Opinions on COVID-19 vaccination, (2) Views on public health, and (3) Sentiments regarding vaccine hesitancy. Within these discussions, the emotions of Happiness, Sadness, and Surprise were notably present. However, during the Omicron outbreak, all provinces exhibited an increase in tweets related to the Omicron variant itself. Specifically, tweets gravitated toward Topic 0: Opinions on the Omicron variant, Topic 4: Perceptions of the government's response to Omicron, and Topic 3. In provinces such as Gauteng, Mpumalanga, KwaZulu-Natal, and Western Cape, there was a significant surge in tweets associated with Topic 0, marked by the emotion of Fear. Similarly, discussions on Topics 1, 2, and 4 in these provinces were characterized by emotions of Angry and Surprise, more so than in other regions. The spike in tweet volume in these provinces may reflect their denser populations and status as commercial hubs within South Africa.

Regarding the correlation between COVID-19 vaccine-related discussions and vaccination rates, the Granger causality test revealed statistically significant results for Mpumalanga and Western Cape provinces before the Omicron emergence. This suggests that heightened discussions about the COVID-19 vaccine on Twitter may have positively influenced vaccination rates, particularly among individuals aged 18–34 in these provinces. However, during the Omicron outbreak, our analysis did not find a statistically significant relationship between emotional intensities and vaccination rates in the Free State, KwaZulu-Natal, Mpumalanga, Northern Cape, and Western Cape provinces. This suggests that emotional intensity in these discussions was not a pivotal factor driving vaccination rates during this period. This observation aligns with the hypothesis that residents of these provinces may have become less engaged with discussions about the variant on Twitter, potentially contributing to lower vaccination rates and increased vaccine hesitancy (Abdullah et al., 2022; Dyer, 2021).

In contrast, for the Eastern Cape, Gauteng, Limpopo, and North West provinces, there was statistically significant evidence of a positive correlation between emotional intensities in COVID-19 vaccine-related tweets and the rise in vaccination uptake during the Omicron outbreak. This suggests that emotional responses may have spurred more individuals to get vaccinated in these provinces during the Omicron wave, an influence that was not apparent in the pre-Omicron data. The detected shifts in emotions—including Happiness, Sadness, Angry, Surprise, and Fear—may be linked to intensified media coverage and heightened awareness efforts by health authorities, particularly South Africa's National Coronavirus Command Council (NCCC). This heightened response became especially pronounced when the NCCC announced the continuation of lockdown measures at the lowest level of their five-tier system on December 16, 2021, shortly after the emergence of the Omicron variant (Reuters, 2021).

Finally, the analysis of the association between emotional classes before and during the Omicron variant, using the `Mann--Whitney U` test, showed significant evidence for the Surprise emotional class in Eastern Cape province. Additionally, Happy, Sadness, Surprise, and Fear emotional classes demonstrated significant evidence in Free State and North West provinces, while Happy and Surprise emotional classes showed strong significance in Limpopo province. Furthermore, the Sadness emotional class was statistically significant in the Northern Cape province. These results indicate that the dynamics of these emotional classes and their intensities may be influenced by similar external factors. Understanding these emotional dynamics during an outbreak could be valuable for health policy planning and management to effectively control the spread of a new COVID-19 variant or future pandemics. Previous studies (Tong, Shi, Zhang, & Shi, 2022) have provided comprehensive findings on the spatiotemporal dynamics of the Omicron variant to aid in controlling its spread in the nine South African provinces, further corroborating the importance of understanding emotional dynamics during outbreaks for effective health policy implementation.

### 5.2. Policy implications

Our study explores vaccine hesitancy by examining its presence and impact on emotional classes identified in COVID-19 vaccine-related discussions before and during the Omicron outbreak in South Africa. We found that several provinces, including Eastern Cape, Free State, North West, Limpopo, and Northern Cape, showed statistically significant correlations between emotional classes such as Fear, Sadness, and Surprise and vaccination rates during Omicron. These findings suggest that vaccine hesitancy may be reflected in the observed emotional responses in these provinces.

Understanding vaccine hesitancy is crucial for health policy-making, as it can help identify areas where targeted interventions are needed to improve vaccine acceptance and uptake. Based on our findings, we recommend that public health authorities and policymakers consider implementing tailored communication strategies and community engagement initiatives to address vaccine hesitancy in these regions, especially during an outbreak. These efforts can play a crucial role in enhancing public trust in vaccination and improving overall vaccine acceptance rates.

Additionally, our study highlights factors that influenced emotional responses toward vaccines. Factors that increased Happiness included the perception of vaccines as a solution to the pandemic, the hope for a return to normalcy, and the belief in the effectiveness and safety of vaccines. Conversely, factors that increased Sadness included concerns about vaccine safety and side effects, misinformation and distrust in vaccines, and the perceived impact of vaccines on personal health and lifestyle.

Policymakers can address these factors by implementing targeted communication strategies that emphasize the safety and effectiveness of vaccines, provide clear and transparent information about vaccination, and address misinformation through educational campaigns. They can also work to improve access to vaccines, ensure equitable distribution, and provide support and resources to address vaccine-related concerns. Furthermore, policymakers can engage with communities and stakeholders to build trust and address specific concerns, tailor interventions to address the needs of different population groups, and promote vaccination as a collective effort to protect public health and prevent the spread of the virus. By addressing these factors, policymakers can help create a more supportive environment for vaccination and improve overall vaccine acceptance and uptake rates.

### 5.3. Limitations and future work

The Twitter data used for this research reflects the opinions of Twitter users located in South Africa from October 1, 2021, to January 15, 2022. It is important to acknowledge that South Africa has a population of about 60 million people, and Twitter users may not accurately represent the broader population. With only an estimated 15% of online adults aged 18–34 using Twitter (Statista, 2022), this research may not fully capture the opinions of all South Africans regarding COVID-19 vaccine-related discussions. Our study acknowledges potential biases in the demographics of Twitter users, who are typically younger, more urban, and have higher levels of education and income compared to the general population. Additionally, the use of English as the primary language may not fully represent the diverse linguistic and demographic landscape of South Africa. While English is one of the country's most widely spoken official languages, future research should aim to include other official languages to ensure a more comprehensive analysis. The reliance on Twitter data also introduces inherent biases related to internet access, digital literacy, and the platform's user demographics, which should be considered when interpreting the findings.

Furthermore, the vaccination data used in this study specifically focused on individuals aged 18–34, narrowing the scope of the analysis. Future studies could explore other age groups to gain a more comprehensive understanding of vaccination sentiments across the population. We also recognize the exclusion of the Northern Cape from certain analyses due to insufficient correlation between Twitter data and vaccination rates. As a result, we could not establish a relationship between Twitter data and vaccination rates in this province. We recommend including all provinces in future research to provide a more comprehensive analysis of the sentiment landscape across South Africa.

Another limitation is the Text2Emotion pre-trained model used for emotion classification, which cannot accurately label figurative language, such as sarcasm, pidgin English, and vernacular. Despite this limitation, the approach was validated through a comparison with a randomly selected 10% of hand-labeled tweets from the dataset. Future research should explore more advanced emotion classification models capable of handling complex emotions and figurative language to improve accuracy. While the Text2Emotion model focuses on five emotional classes—Angry, Happy, Fear, Sad, and Surprise—future studies could explore additional emotional nuances and classifications for a more detailed analysis.

To address these limitations, future research should consider incorporating additional social media platforms (e.g., Facebook, Instagram, and WhatsApp) and traditional media sources (e.g., newspapers, radio, and television) to capture a broader range of public sentiments. Including data from these sources would help provide a more comprehensive view of public opinions across different demographic groups and regions. Additionally, future studies should explore hybrid approaches that combine lexicon-based methods with advanced machine learning or transformer-based models (e.g., BERT, RoBERTa) for more accurate

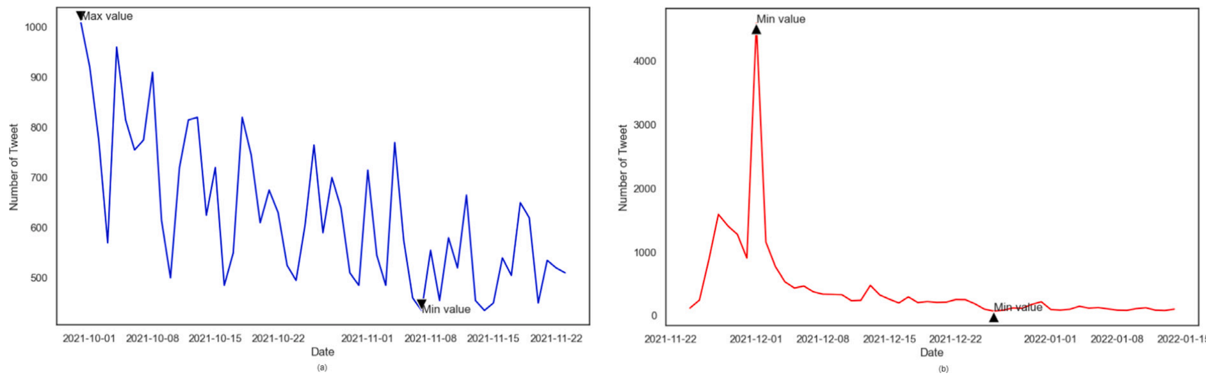


**Fig. B.7.** Tweet trends (a) before and (b) during Omicron.

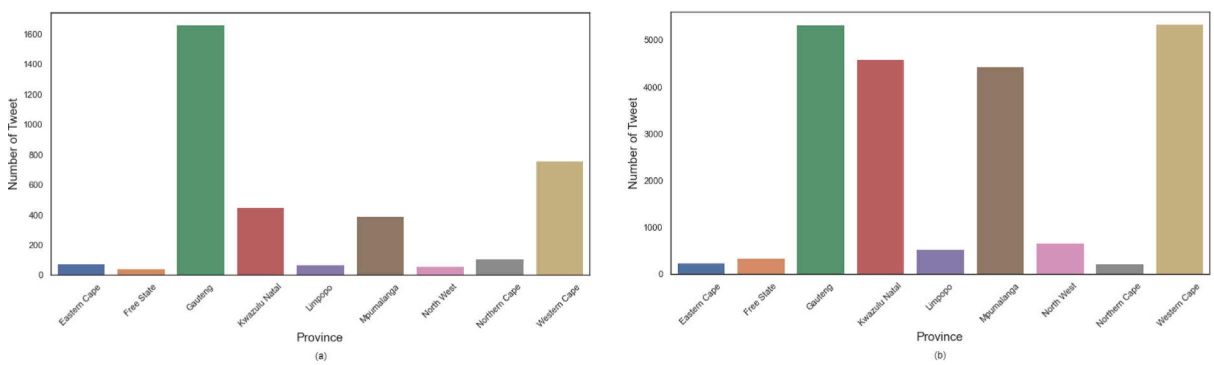


**Fig. B.8.** Tweets distribution by province (a) before and (b) during Omicron.

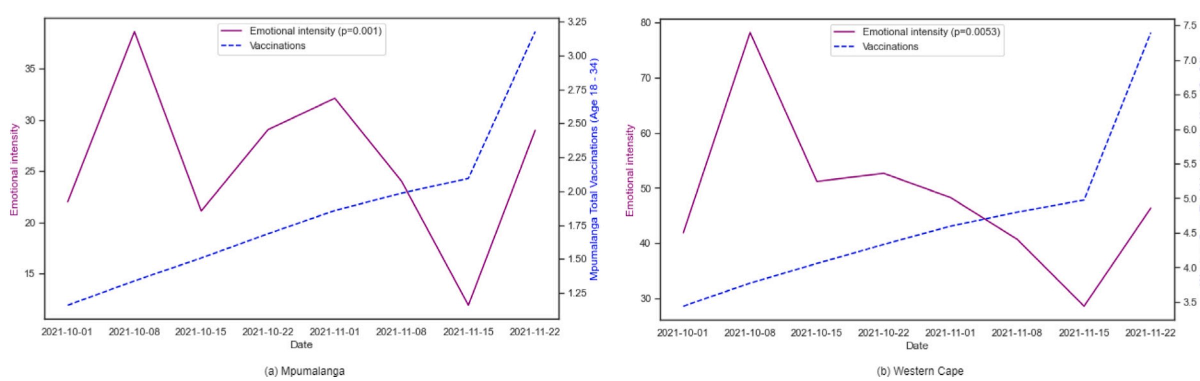


**Fig. B.9.** Impact of COVID-19 discussion on vaccination before Omicron.

and nuanced emotion classification. Longitudinal studies could also be conducted to track changes in emotional dynamics and vaccination rates over time, offering deeper insights into the factors driving vaccine hesitancy and acceptance.

Despite these limitations, this research provides valuable complementary insights to existing data and offers important implications for policymaking. Future studies should aim to address these limitations to obtain a more comprehensive understanding of public sentiments toward COVID-19 vaccination in South Africa.

## 6. Conclusion

This study highlights a significant correlation between Twitter discussions and vaccination rates across South African provinces during the Omicron era, underscoring the influence of social media on vaccination behavior. Emotional intensities were found to be strongly associated with vaccination increases in several provinces, emphasizing the role of emotions in shaping vaccination decisions. However, it is important to note that the study's findings are limited by the reliance on social media data, which may not fully represent the broader population. Additionally, while the study provides valuable insights for health policy and communication strategies, it does not establish causation between social media discussions and vaccination rates. Nevertheless, the research demonstrates the potential of social media as a tool for monitoring public sentiments and vaccination trends, offering opportunities for more informed and adaptive public health responses in future health crises.

### 6.1. Ethical considerations

This study was conducted in adherence to ethical guidelines and principles. Approval for the research was obtained from Twitter, and access to the Twitter academic researcher API was granted, allowing us to retrieve the necessary tweets. It is important to note that, at the

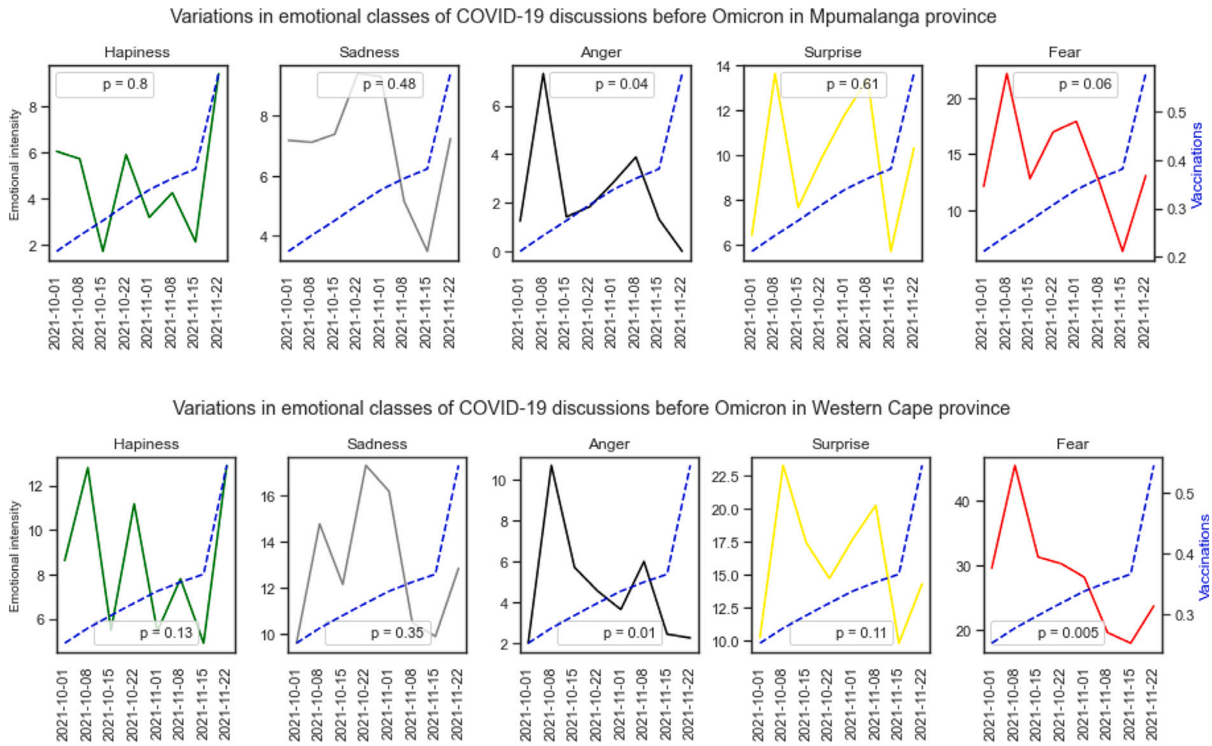


**Fig. B.10.** Emotional dynamics on vaccination before Omicron.

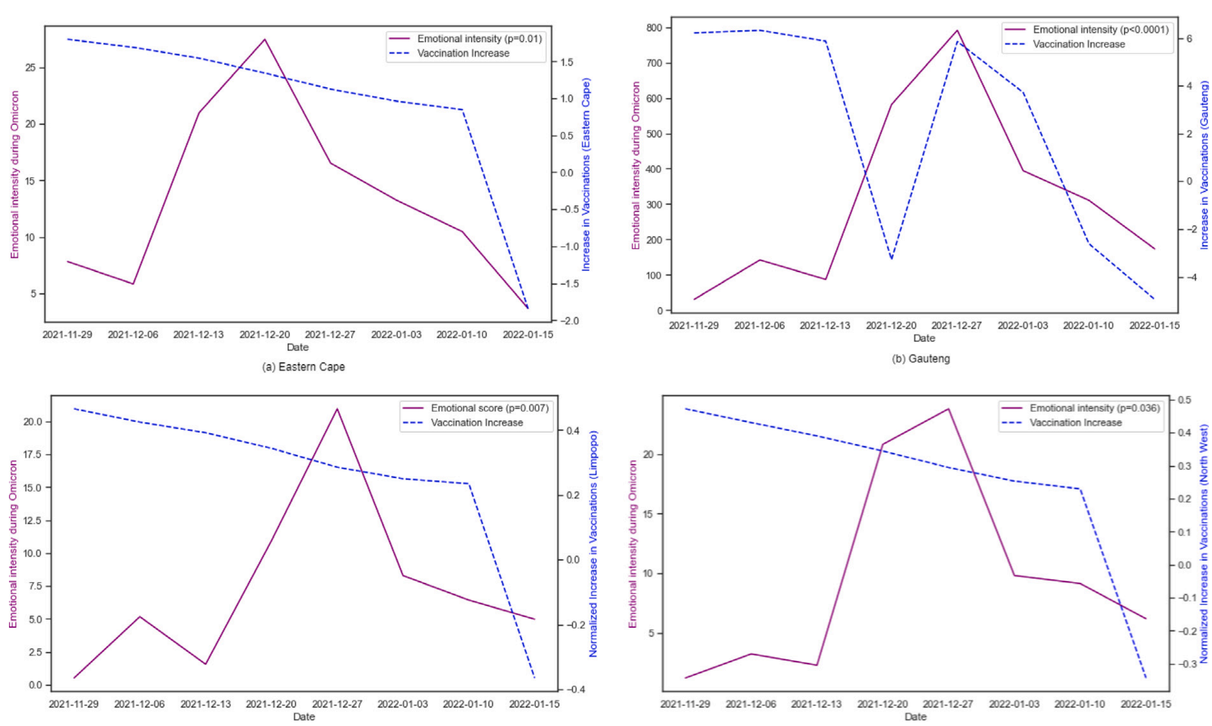


**Fig. B.11.** Impact of COVID-19 discussion on vaccination during Omicron.

time of data collection, all the tweets used in this study were publicly available and within the public domain. To uphold the highest ethical standards and respect the privacy of Twitter users, the authors took great care in handling any personal information associated with the retrieved tweets. All personal information was meticulously removed and anonymized to ensure the confidentiality and anonymity of the

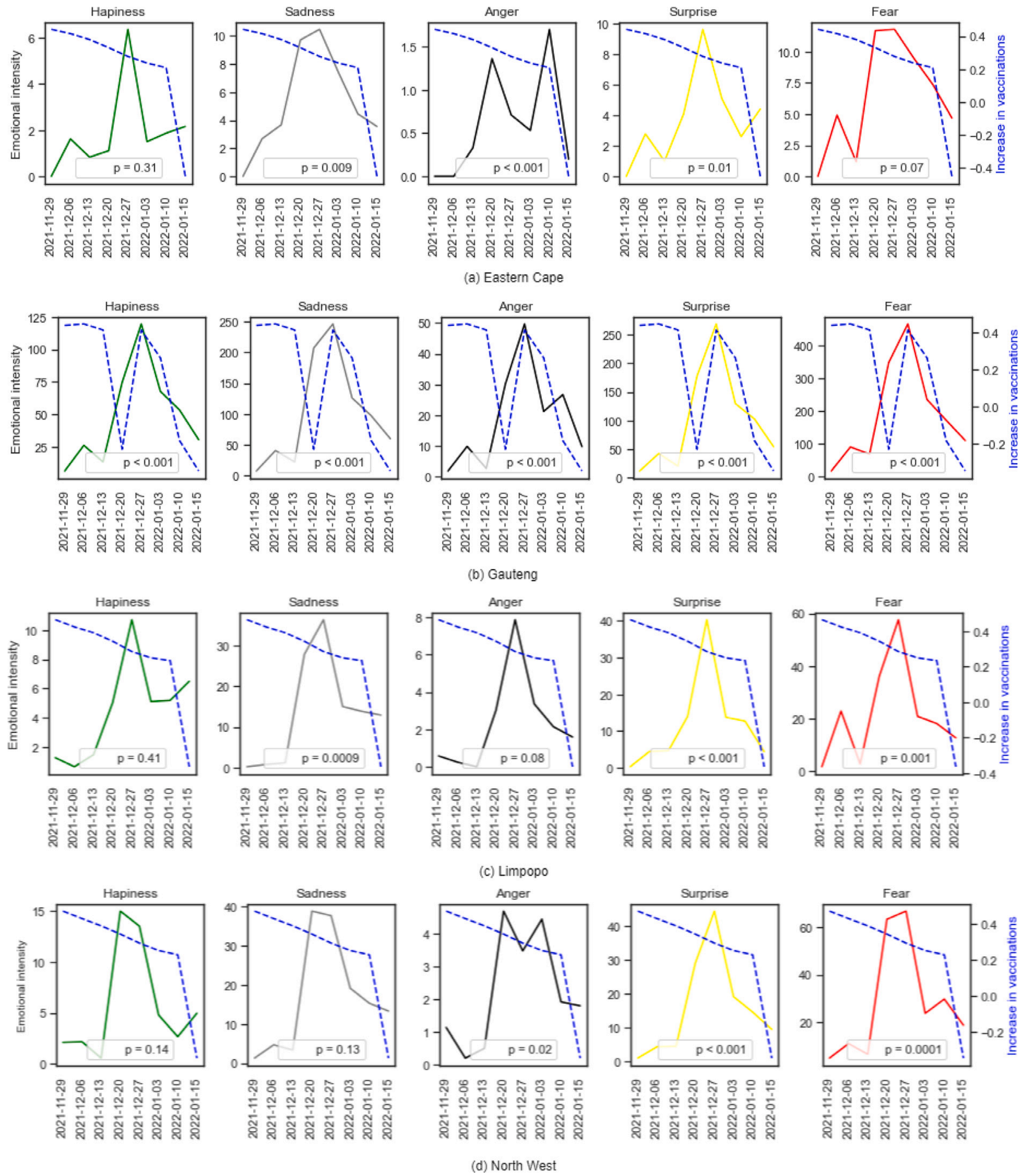


**Fig. B.12.** Emotional dynamics on vaccination during Omicron.

users. By following these ethical practices, the research team ensured that the study was conducted responsibly and with due consideration for the privacy and rights of Twitter users, as well as in compliance with the terms and conditions set forth by Twitter for academic research.

## CRediT authorship contribution statement

**Blessing Ogbuokiri:** Conception, Design, Data collection, Analysis, Draft, Revision of the manuscript. **Ali Ahmadi:** Supervision. **Nidhi Tripathi:** Vaccination Data collection. **Laleh Seyyed-Kalantari:** Manuscript editorial revision. **Woldergebriel Assefa Woldegerima:** Manuscript editorial revision. **Bruce Mellado:** Supervision. **Jiahong Wu:** Supervision. **James Orbinski:** Supervision. **Ali Asgary:** Supervision. **Jude Dzevela Kong:** Supervision, Conception, Design, Data collection, Analysis, Draft, Manuscript editorial revision.

## Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to to improve language and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

## Funding

This research is funded by Canada's International Development Research Centre (IDRC) (Grant No. 109981). JDK acknowledges support from NSERC Discovery Grant (Grant No. RGPIN-2022-04559), NSERC Discovery Launch Supplement (Grant No: DGECR-2022-00454), New Frontier in Research Fund- Exploratory (Grant No. NFRFE-2021-00879), IDRC and the Foreign, Commonwealth and Development Office (FCDO) (Grant No. 110554-001) and Canada Research Chair in Community-Oriented Artificial Intelligence and Mathematical Modelling of Infectious Diseases (Grant No. CRC-2023-00234)

## Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Acknowledgments

The authors wish to acknowledge the Responsible and Applied Machine Learning Laboratory (RAML LAB) at Brock University, Canada, the Africa-Canada AI & Data Innovation Consortium (ACADIC) team at the University of Toronto, Canada, and the University of the Witwatersrand, Johannesburg, South Africa.

## Appendix A. Code snippet: search keywords

```
(omicron OR sarscov2 OR covid19 OR covid−19sa OR
    coronavirus OR coronavirus sa OR corona OR
    sars−cov−2 OR lockdown OR covid−19 OR pandemic)
    AND (vaccine OR vaccines OR Oxford−AstraZeneca
    OR AstraZeneca OR JohnsonJohnson OR BioNTech OR
    jab OR Vaccination OR Covax OR Vaccine Rollout
    OR Sputnik OR #VaccineToSaveSouthAfrica OR
    #IChooseVaccination OR #TeachersVaccine OR
    AstraZeneca vaccine OR Pfizer OR J & J OR
    Johonson & Johnson OR Moderna OR #VaccinesWork
    OR #VacciNation OR Steriod OR #COVIDvaccine OR
    #covax OR #VaccineEquity OR #VaccineReady OR
    #Jab OR #PfizerGang)
```

## Appendix B. More figures

See Figs. B.7–B.12.

## Data availability

Our dataset is publicly available on Kaggle. Please visit: https://www.kaggle.com/datasets/ogbuokiriblessing/x-omicron-dataset to allow researchers to hydrate and reuse. However, due to Twitter's developers policy Agreement and Policy (2023), we can only share Tweet IDs publicly.

## References

Abdullah, F., Myers, J., Basu, D., Tintinger, G., Ueckermann, V., Mathebula, M., Ramlall, R., Spoor, S., de Villiers, T., der Walt, Z. V., Cloete, J., Soma-Pillay, P., Rheeder, P., Paruk, F., Engelbrecht, A., Lalloo, V., Myburg, M., Kistan, J., van Hougenhouck-Tulleken, W., ...., Boswell, M. et al. (2022). Decreased severity of disease during the first global omicron variant covid-19 outbreak in a large hospital in tshwane, South Africa. *International Journal of Infectious Diseases*, *116*, 38–42.

Aden, D., & Zaheer, S. (2024). Investigating the FLiRT variants of COVID-19: Is it an emerging concern? *Pathology-Research and Practice*, *262*, Article 155542.

Aditya, B. (2022). Industrial strenght natural language processing in Python. Available on https://spacy.io/. [Accessed 01 June 2022].

Agreement, D., & Policy (2023). X. Online: https://developer.twitter.com/en/developer-terms/agreement-and-policy. [Accessed 03 May 2024].

Ainapure, B. S., et al. (2022). Sentiment analysis of COVID-19 vaccination tweets using deep learning and lexicon-based techniques. *Journal of Public Health Informatics*, *14*(3), 123–135.

Aslam, N., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2022). Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble LSTM-GRU model. *IEEE Access*, *10*, 39313–39324.

Bangyal, W. H., Amina, S., Shakir, R., Ubakanma, G., & Iqbal, M. (2023). Using deep learning models for COVID-19 related sentiment analysis on Twitter data. In *2023 international conference on human-centered cognitive systems* (pp. 1–6). IEEE.

Bangyal, W. H., Qasim, R., Rehman, N. U., Ahmad, Z., Dar, H., Rukhsar, L., Aman, Z., & Ahmad, J. (2021). Detection of fake news text classification on COVID-19 using deep learning approaches. *Computational and Mathematical Methods in Medicine*, *2021*(1), Article 5514220.

Bedre, R. (2020). 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 01 June 2022].

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Braig, S., et al. (2022). Machine learning-based sentiment analysis of COVID-19 Twitter data: A systematic review. *Computers in Biology and Medicine*, *145*, 105–118.

Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2018). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *32*, In *Proceedings of AAAI 2018* (1), (pp. 1795–1802).

Cloete, J., Kruger, A., Masha, M., du Plessis, N. M., Mawela, D., Tshukudu, M., Manyane, T., Komane, L., Venter, M., Jassat, W., Goga, A., & Feucht, U. (2022). 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 & Adolescent Health*, *6*(5), 294–302.

Cooper, S., van Rooyen, H., & Wiysonge, C. S. (2021). COVID-19 vaccine hesitancy in South Africa: how can we maximize uptake of COVID-19 vaccines? *Expert Review of Vaccines*, *20*(8), 921–933.

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

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186).

Díaz, S. S., Shaik, J. M. M., Santofimio, J. C. G., & Quintero M., C. G. (2018). Intelligent execution of behaviors in a NAO robot exposed to audiovisual stimulus. In *2018 IEEE 2nd Colombian conference on robotics and automation* (pp. 1–6).

Dyer, O. (2021). Covid-19: Omicron is causing more infections but fewer hospital admissions than delta, South African data show. Online: https://www.bmj.com/content/375/bmj.n3104. [Accessed 16 March 2025].

Elyashar, A., Plochotnikov, I., Cohen, I.-C., Puzis, R., & Cohen, O. (2021). The state of mind of health care professionals in light of the COVID-19 pandemic: text analysis study of Twitter discourses. *Journal of Medical Internet Research*, *23*(10), Article e30217.

Frederic, G., Marek, K., & Tomasz, L. (2022). 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*(2), 294.

GeeksForGeeks (2023). Jaccard similarity. Online: https://www.geeksforgeeks.org/how-to-calculate-jaccard-similarity-in-python/. [Accessed 29 December 2024].

Gulati, K., Kumar, S. S., Boddu, R. S. K., Sarvakar, K., Sharma, D. K., & Nomani, M. (2022). Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to COVID-19 pandemic. *Materials Today: Proceedings*, *51*, 38–41.

Gupta, A., Band, A., Karan, Sharma, S. R., & Deepspacehub (2020). text2emotion 0.0.5. Available on https://pypi.org/project/text2emotion/. [Accessed 01 June 2022].

Hanschmidt, F., & Kersting, A. (2023). Emotions in Covid-19 Twitter discourse following the introduction of social contact restrictions in Central Europe. *Journal of Public Health*, *31*(6), 933–946.

Honnibal, M. (2024). spaCy 2: Natural language understand- ing with Bloom embeddings, convolutional neural net- works and incremental parsing. *Sentometrics Research*, *1*(1), 2586–2593.

Hulswit, R., de Haan, C., & Bosch, B.-J. (2016). Chapter two - coronavirus spike protein and tropism changes. In J. Ziebuhr (Ed.), *Advances in virus research: vol. 96*, *Coronaviruses* (pp. 29–57). Academic Press.

Kamath, R., Ghoshal, A., Eswaran, S., & Honnavalli, P. (2022). An enhanced context-based emotion detection model using roberta. In *2022 IEEE international conference on electronics, computing and communication technologies* (pp. 1–6). IEEE.

Kiritchenko, S., Zhu, X., & Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, *50*, 723–762.

Kumar, A., Singh, J. P., & Singh, A. K. (2023). Explainable sentiment analysis of COVID-19 vaccination tweets using CT-BERT-LSTM. *IEEE Transactions on Computational Social Systems*, *10*(2), 456–467.

Li, F., Van den Bossche, J., Zeitlin, M., Meeseeks, M., PD, T., Hawkins, S., et al. (2021). What's new in Pandas 1.2.4. Available on https://pandas.pydata.org/pandas-docs/stable/whatsnew/v1.2.4.html. [Accessed 01 June 2022].

Li, G., Zhang, A., Zhang, Q., Wu, D., & Zhan, C. (2022). Pearson correlation coefficient-based performance enhancement of broad learning system for stock price prediction. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *69*(5), 2413–2417.

Lin, X.-Q., Lv, L., Chen, Y., Chen, H.-D., Zhang, M.-X., Tung, T.-H., & Zhu, J.-S. (2022). The impact of the Omicron epidemic on the health behavior in Cape Town, South Africa. *One Health*, *14*, Article 100395.

Lin, L., Qi, Z., & Jiancheng, X. (2020). Changes of laboratory cardiac markers and mechanisms of cardiac injury in coronavirus disease 2019. *BioMed Research International*, *7*, 171–174.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Lulan, W., & Genhong, C. (2022). Sequence analysis of the emerging SARS-CoV-2 variant Omicron in South Africa. *Journal of Medical Virology*, *94*(4), 1728–1733.

Meganck, R. M., Edwards, C. E., Mallory, M. L., Lee, R. E., Dang, H., Bailey, A. B., Wykoff, J. A., Gallant, S. C., Zhu, D. R., Yount, B. L., et al. (2024). SARS-CoV-2 variant of concern fitness and adaptation in primary human airway epithelia. *Cell Reports*, *43*(4).

Mighri, Z., Ragoubi, H., Sarwar, S., & Wang, Y. (2022). Quantile granger causality between US stock market indices and precious metal prices. *Resources Policy*, *76*, Article 102595.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, *29*(3), 436–465.

Mukherjee, R., & Satardekar, R. (2021). Why are some coronavirus variants more infectious? *Journal of Biosciences*, *46*(101).

NLTK Documentation (2023). Natural language toolkit. Available on https://www.nltk.org/. [Accessed 01 June 2022].

Ogbuokiri, B., Ahmadi, A., Bragazzi, N. L., Movahedi Nia, Z., Mellado, B., Wu, J., Orbinski, J., Asgary, A., & Kong, J. (2022). Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts. *Frontiers in Public Health*, *10*, Article 987376.

Ogbuokiri, B., Ahmadi, A., Mellado, B., Wu, J., Orbinski, J., Asgary, A., & Kong, J. (2023). Can post-vaccination sentiment affect the acceptance of booster jab? In A. Abraham, S. Pllana, G. Casalino, K. Ma, & A. Bajaj (Eds.), *Intelligent systems design and applications* (716). (pp. 200–211). Switzerland, Cham: Springer Nature.

Ogbuokiri, B., Ahmadi, A., Nia, Z. M., Mellado, B., Wu, J., Orbinski, J., Asgary, A., & Kong, J. (2023). Vaccine hesitancy hotspots in Africa: An insight from geotagged Twitter posts. *IEEE Transactions on Computational Social Systems*, 1–14.

Özcan, S. (2024). Tweet-preprocessor 0.6.0. Available on https://pypi.org/project/tweet-preprocessor/. [Accessed 06 January 2022].

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on empirical methods in natural language processing* (pp. 79–86).

Reuters (2021). South Africa to retain 'Level 1' curbs in Omicron fight. Online: https://www.reuters.com/world/africa/south-africa-retain-level-1-curbs-omicron-fight-2021-12-16/. [Accessed 29 December 2024].

Rosol, M., Młyńczak, M., & Cybulski, G. (2022). Granger causality test with nonlinear neural-network-based methods: Python package and simulation study. *Computer Methods and Programs in Biomedicine*, *216*, Article 106669.

Scikit Learn (2024). sklearn.metrics.jaccard_score. Online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.jaccard_score.html. [Accessed 04 May 2024].

Statista (2022). 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 01 June 2022].

Statista (2024). Distribution of X (formerly Twitter) users worldwide as of April 2024, by age group. Available on https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/. [Accessed 04 May 2024].

Tal, E. (2025). Models and measurement. In *The Routledge Handbook of Philosophy of Scientific Modeling* (pp. 256–269). Routledge.

Tasnim, S., Hossain, M., & Mazumder, H. (2024). Impact of rumors and misinformation on COVID–19 in social media. *Journal of Preventive Medicine & Public Health*, *202*(53), 171–174.

Tharu, M., Pokhrel, S., & Lamichhane, B. R. (2023). Sentiment analysis of Nepali COVID-19 tweets using BERT-LSTM. *Journal of Engineering and Sciences*, *2*(1), 49–56.

Tong, C., Shi, W., Zhang, A., & Shi, Z. (2022). Tracking and controlling the spatiotemporal spread of SARS-CoV-2 Omicron variant in South Africa. *Travel Medicine and Infectious Disease*, *46*, Article 102252.

Wang, Q., Hu, S., Du, F., Zang, S., Xing, Y., Qu, Z., Zhang, X., Lin, L., & Hou, Z. (2022). Mapping global acceptance and uptake of COVID-19 vaccination: A systematic review and meta-analysis. *Communications Medicine*, *2*(1), 113.

Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2020). Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of EMNLP 2020* (pp. 606–615).

Wella, K., Mtambo, J., Lazaro, M., & Mapulanga, P. (2025). The role of social media for health communication during COVID-19: A scoping review. *Journal of Consumer Health on the Internet*, 1–27.

Wimalawansa, S. J. (2024). Unlocking insights: Navigating COVID-19 challenges and emulating future pandemic resilience strategies with strengthening natural immunity. *Heliyon*, *10*(15).

Wu, Y., Jin, Z., Shi, C., Liang, P., & Zhan, T. (2024). Research on the application of deep learning-based BERT model in sentiment analysis. arXiv preprint arXiv:2403.08217.

Xie, Z., Wang, X., Jiang, Y., Chen, Y., Huang, S., Ma, H., Anand, A., & Li, D. (2021). Public perception of COVID-19 vaccines on Twitter in the united states. Medrxiv.

Xuemei, H., Weiqi, H., Xiangyu, P., Guangwen, L., & Xiawei, W. (2021). SARS-CoV-2 Omicron variant: Characteristics and prevention. *Med Comm*, *2*, 838–845.

Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., & Sharif, S. (2021). An analysis of COVID-19 vaccine sentiments and opinions on Twitter. *International Journal of Infectious Diseases*, *108*, 256–262.

Zamir, M. T., Ullah, F., Tariq, R., Bangyal, W. H., Arif, M., & Gelbukh, A. (2024). Machine and deep learning algorithms for sentiment analysis during COVID-19: A vision to create fake news resistant society. *PloS One*, *19*(12), Article e0315407.

Zhang, X., Zhao, J., & LeCun, Y. (2018). Character-level convolutional networks for text classification. In *Advances in neural information processing systems: vol. 28*, (pp. 649–657).

Zhao, Y., Huang, J., Zhang, L., Chen, S., Gao, J., & Jiao, H. (2022). The global transmission of new coronavirus variants. *Environmental Research*, *206*, Article 112240.