

# Tracing Unemployment Rate of South Africa during the COVID-19 Pandemic Using Twitter Data

Zahra Movahedi Nia, Ali Asgary, Nicola Bragazzi, Bruce Melado, James Orbinski, Jianhong Wu, Jude Dzevela Kong

Submitted to: Journal of Medical Internet Research  
on: September 26, 2021

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

## ***Table of Contents***

---

<b>Original Manuscript.....</b>	<b>5</b>
---------------------------------	----------

Preprint  
JMIR Publications

# Tracing Unemployment Rate of South Africa during the COVID-19 Pandemic Using Twitter Data

Zahra Movahedi Nia<sup>1</sup>; Ali Asgary<sup>2</sup>; Nicola Bragazzi<sup>1</sup>; Bruce Melado<sup>3</sup>; James Orbinski<sup>4</sup>; Jianhong Wu<sup>1</sup>; Jude Dzevela Kong<sup>1</sup>

<sup>1</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Laboratory for Industrial and Applied mathematics, York University Toronto CA

<sup>2</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), the Advanced Disaster, Emergency and Rapid Response Program, York University Toronto CA

<sup>3</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand Johannesburg ZA

<sup>4</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), the Dahdaleh Institute for Global Health Research, York University Toronto CA

## Corresponding Author:

Jude Dzevela Kong

Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Laboratory for Industrial and Applied mathematics, York University

4700 Keele Street

Toronto

CA

## Abstract

**Background:** Global economy has been hardly hit by the COVID-19 pandemic. Many countries are experiencing a severe and destructive recession. Unemployment rate is very important to policy makers as it provide a key indicator of overall labour market and wider economic conditions. Despite its relevance, there is usually a delay in the availability of the indicator as it is traditionally based on a survey of households over several months. The speed at which the economy in most countries decline at the onset of COVID-19 highlights the importance of timely information about the labour market during the onset of a recession. In the coming year, there will be uncertainty about the timing and extent of any improvement in labour market outcomes that will also highlight the value of timely information.

**Objective:** The main goal of this study is to provide policy- and decision-makers with additional and real-time information about the labor market flow during a prolonged pandemic. The first objective of the study is to find the missing unemployment rates in cases where census measurements are incomplete. The second objective is to estimate the unemployment rate in real-time since it usually takes months for formal unemployment data to be published. In this paper, we use social media data, particularly, Twitter to trace and nowcast the unemployment rate of South Africa during the COVID-19 pandemic.

**Methods:** Unemployment rate in South Africa is estimated quarterly. We first used Google mobility index to interpolate it and find the monthly values. Next, we created a dataset of unemployment related tweets in South Africa using certain keywords such as employed, unemployed, and retrench. Principal Component Regression (PCR) was applied to estimate the unemployment rate using the tweets and their sentiment scores.

**Results:** Numerical results indicate that the number of tweets is highly correlated with the unemployment rate during and before the COVID-19 pandemic. In addition, the trend of the normalized sum of the sentiment scores of the tweets is negatively correlated with the unemployment rate of South Africa. Moreover, the estimated unemployment rate using PCR is highly correlated with the actual unemployment rate of South Africa and has a low Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

**Conclusions:** The results of this study show that social media information can be used to reasonably estimate one of the key labor market indicators, especially during disaster events such as a prolonged pandemic. This information can be used to rapidly understand and manage the impacts of the pandemic on the economy and people's life.

(JMIR Preprints 26/09/2021:33843)

DOI: <https://doi.org/10.2196/preprints.33843>

## Preprint Settings

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

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

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

Only make the preprint title and abstract visible.

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

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

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

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

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

## Original Manuscript

# Tracing Unemployment Rate of South Africa during the COVID-19 Pandemic Using Twitter Data

Zahra Movahedi Nia<sup>1</sup>, Ali Asgary<sup>2\*</sup>, Nicola Bragazzi<sup>1\*</sup>, Bruce Melado<sup>3\*</sup>, James Orbinski<sup>4\*</sup> and Jianhong Wu<sup>1\*</sup>, Jude Kong<sup>1\*†</sup>

<sup>1</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Laboratory for Industrial and Applied mathematics, York University, Canada

<sup>2</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), the Advanced Disaster, Emergency and Rapid Response Program, York University, Canada

<sup>3</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), School of Physics, Institute for Collider Particle Physics, University of the Witwatersrand, Johannesburg, South Africa

<sup>4</sup>Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), the Dahdaleh Institute for Global Health Research, York University, Canada

\*These authors have contributed equally to this work and share last authorship

†Corresponding Author: jdkong@york.ca

## Abstract

**Background:** Global economy has been hardly hit by the COVID-19 pandemic. Many countries are experiencing a severe and destructive recession. Unemployment rate is very important to policy makers as it provide a key indicator of overall labour market and wider economic conditions. Despite its relevance, there is usually a delay in the availability of the indicator as it is traditionally based on a survey of households over several months. The speed at which the economy in most countries decline at the onset of COVID-19 highlights the importance of timely information about the labour market during the onset of a recession. In the coming year, there will be uncertainty about the timing and extent of any improvement in labour market outcomes that will also highlight the value of timely information.

**Objectives:** The main goal of this study is to provide policy- and decision-makers with additional and real-time information about the labor market flow during a prolonged pandemic. The first objective of the study is to find the missing unemployment rates in cases where census measurements are incomplete. The second objective is to estimate the unemployment rate in real-time since it usually takes months for formal unemployment data to be published. In this paper, we use social media data, particularly, Twitter to trace and nowcast the unemployment rate of South Africa during the COVID-19 pandemic.

**Methods:** Unemployment rate in South Africa is estimated quarterly. We first used Google mobility index to interpolate it and find the monthly values. Next, we created a dataset of unemployment related tweets in South Africa using certain keywords such as employed, unemployed, and retrench. Principal Component Regression (PCR) was applied to estimate the unemployment rate using the tweets and their sentiment scores.

**Results:** Numerical results indicate that the number of tweets is highly correlated with the unemployment rate during and before the COVID-19 pandemic. In addition, the trend of the normalized sum of the sentiment scores of the tweets is negatively correlated with the unemployment rate of South Africa. Moreover, the estimated unemployment rate using PCR is highly correlated with the actual unemployment rate of South Africa and has a low Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

**Conclusion:** The results of this study show that social media information can be used to reasonably estimate one of the key labor market indicators, especially during disaster events such as a prolonged pandemic. This information can be used to rapidly understand and manage the impacts of the pandemic on the economy and people's life.

**Keywords:** sentiment analysis, social media, Twitter data, Google mobility index, unemployment rate, labor market, COVID-19, South Africa.

### 1. Introduction

The novel coronavirus known as “Severe Acute Respiratory Syndrome-related Coronavirus type 2” (SARS-CoV-2), responsible for the “Coronavirus Disease 2019” (COVID-19) pandemic, was first detected in the metropolitan city of Wuhan, Hubei Province, mainland China, in late December 2019. Since then, it quickly spread around the globe, causing more than 230 million infections and 4.7 million deaths, as of September 24, 2021. The World Health Organization (WHO) officially declared the COVID-19 outbreak, initially, as a Public Health Emergency of International Concern (PHEIC) on January 30, 2020, and, later, as a global pandemic on March 11, 2020.

Since then, countries have enforced non-pharmaceutical interventions (NPIs) to curb the diffusion of the virus and prevent its spread, including lockdowns and different levels of restrictions. Even though effective both from a clinical and epidemiological perspective, consecutive rounds of NPIs have had devastating effects on the economy and caused bankruptcy to many companies and businesses. As a result, many people and individuals have lost their jobs and countries are experiencing economic recession [18]. To better manage the economic impacts of the pandemic on the economy and people, it is highly important to have complete, reliable, and real-time information about the effects of the pandemic on unemployment rate as one of the key macroeconomic indicators.

Traditional census methods that are used by most countries to generate unemployment data are often conducted on a seasonal or annual basis. While this provides sufficient information for public policy in normal situations, they lack the details and urgency that is required for decision-making during a pandemic event. Census data often uses questionnaires on a sample of households to collect employment data. Despite using new technologies in data collection (such as online surveys) and analysis, censuses are still expensive, time- and resource-consuming, and difficult to handle. The census method faces many other challenges and limitations such as privacy concerns, low public cooperation, errors caused by response burden, cybersecurity attacks (e.g. denial of service), and missing out hard-to-reach populations. Migration, homelessness, and nomadism may result in under- or over-registration, making collected data not representative of the entire population. Low levels of literacy and language issues may cause some people to struggle with the census forms and fail to provide correct information.

Due to such difficulties, the unemployment rate in South Africa is also estimated quarterly. In contrast, social media data is readily available. Statistics and demographic information can be easily extracted and processed in real-time. Many of the problems and limitations of the classical census approach do not exist when data are extracted and estimated using social media [28, 29].

Access to socio-economic data such as unemployment rates is very critical for rapid and effective decision-making and public health policies, during devastating disasters such as the still ongoing COVID-19 pandemic. In the present study, we propose a method for estimating unemployment rates during the COVID-19 using social media, particularly Twitter data. Accessing data extracted from social media is fast, easy, and low-cost. It can be done in real-time and does not have the difficulties and limitations of census-based methods.

Social media has provided promising methods and approaches to study different matters such as opinion mining, information dissemination, healthcare, and economy in real-time [19-22]. Twitter as a pervasive social media is widely used for understanding economic behavior and measuring its metrics [23, 24]. It is also one of the most popular social media in Africa [25, 26]. With the implementation of NPIs, such as lockdowns and the closure of workplaces and public areas, people spend even more of their time on social media [27].

In this paper, we aim to examine how the number of tweets can be used to nowcast unemployment rate using South Africa as a case study. The rest of this paper is organized as follows. Section two provides a review of the related research. Section three provides information about the materials and

methods of the study. Section four presents the numerical results for the study followed by the limitations of the study in section five. Section six includes the discussions. Finally, section seven concludes the paper with some suggestions for future studies.

## 2. *Background and Literature Review*

Social media, especially Twitter, has long been used for investigating economic issues. Authors in [13] searched for tweets with hashtags for different keywords on jobs and gathered tweets sent by popular users in the USA. Sentiment analysis showed that most of the tweets had negative sentiments. In [14] a sentiment-based model was designed with 0.6787 accuracy for tweets, news articles and movie reviews and concluded that the sentiment scores were correlated with economic indexes such as the exchange rate. Although social media has long been used for studying economic issues and related concerns, very few studies have considered using social media to understand the unemployment rate. One of the first works that used Twitter to estimate unemployment rate is presented in [7]. In this paper, 19.3 billion tweets were gathered from July 2011 to November 2013 on unemployment in the USA. Principal Component Analysis (PCA) was used to reduce the dimension of the dataset. The unemployment rate of the USA was then estimated using the principal components. A similar approach was proposed in [8] for studying the correlation between number of tweets and the unemployment rate in Greece. Sentiment analysis has not been considered in these two studies to improve the results further. Ryo in [15] analyzed the sentiments of Korean tweets, blogs, and news articles, and used sentiments to predict unemployment rate with autoregression analysis (like ARIMAX and ARX). It concluded that predicting with the Twitter datasets had the lowest error. In addition, predicting with news articles had a lower error compared to end blogs. Another study [16] used keywords to extract tweets from the USA on both employment and unemployment rates. Then using sentiment analysis, they found that a peak of negative or positive sentiments was found for different users around the time they gained or lost their jobs. They also used sentiment analysis to predict the unemployment rate of the USA. Authors in [17] built a linear model to predict employment and unemployment rates using tweets from the USA. Although the papers mentioned above have presented novel methods for studying unemployment rate using social media, they have not investigated unemployment rate changes during the occurrence of a disaster such as the COVID-19 pandemic.

Authors in [9] have hydrated a Twitter dataset to study the correlation between the number of tweets and the unemployment rate and track the unemployment rate of the USA during the COVID-19 pandemic. However, because of the limitations in their dataset they were not able to properly understand how the unemployment rate changed over time. Moreover, they have not considered using techniques such as sentiment analysis to further investigate the tweets related to unemployment rate during the COVID-19.

There are other studies that have focused on the economy during the COVID-19 pandemic. Authors in [10] have used Twitter to study the effect of different factors on reopening sentiments. They found that people with low income, low education level, high housing rent, and in the labor force are more positive about reopening. In [11] Twitter was used to study the economy of the USA during the COVID-19 pandemic. In this work, the Area Deprivation Index (ADI) of different geographical locations was used to assess the economic situation of people. It concluded that in low resource areas people were more concerned with economic hardship while in high resource areas people were more focused on public health. In [12] Twitter and newspaper articles have been used to study economic uncertainty in the UK and the USA during the COVID-19 pandemic. Numerical results show that with COVID-19 pandemic, a huge uncertainty jump was found in economic related indicators such as business growth, GDP growth, and stock market volatility.

These papers have investigated the effect of the COVID-19 pandemic on the economy. However, they do not consider studying and estimating the unemployment rate using social media. The main contribution of this study is to fill the existing gaps in using social media data to estimate unemployment rate during the pandemic using a combination of methods. This combination has



significantly improved the classical method for estimating unemployment rate.

### 3. Materials and Methods

To estimate the unemployment rate for South Africa using Twitter data we performed three steps. In the first step we try to find the missing unemployment data using Google Mobility Index. In the second step we generate unemployment related tweets from Twitter data using relevant keywords. In this step we also conduct a sentiment analysis to get further information about the labor market conditions. Finally, in step three we use PCA to estimate the unemployment rate from the number of tweets. Details of each of these steps are described in this section.

#### 3.1 Finding Missing Unemployment Data

The real unemployment data for South Africa is provided on a seasonal basis. Since we are using Twitter data to estimate unemployment rate and would like to compare them with real data, we have used Google Mobility data to interpolate the unemployment rates in between seasons [1, 2]. Google mobility index shows the movement trends over time and space in six different categories of places namely, retail and recreation, grocery and pharmacies, parks, transit stations, workplaces, and residential. Figure 1 shows the indexes of these categories over time for South Africa.

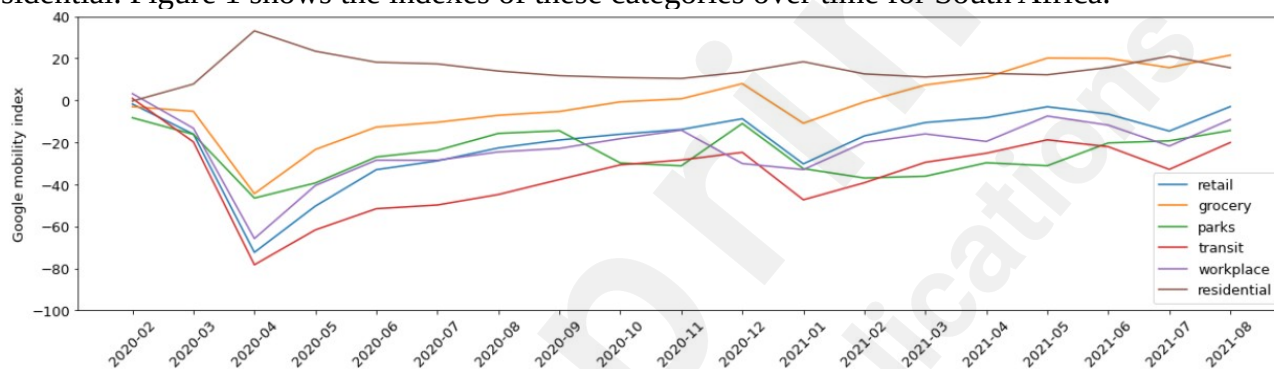


Figure 1: Google mobility index for different categories of places over time for South Africa [1]

Because residential activity is not a work related function and its index has a negative correlation with the rest of the indexes we have excluded it from our analysis. We averaged the indexes of all other categories and used linear regression to interpolate the unemployment rate of South Africa with the Google mobility index. Equation 1 shows the coefficients of the linear regression method used for interpolation.

$$unemp = -0.1214 \times \text{GMI} + 28.5283 \quad (1)$$

Where  $\text{GMI}$  is the Google mobility index averaged over all the categories except the residential places and  $unemp$  is the interpolated unemployment rate. Figure 2 shows the quarterly unemployment rate and the unemployment rate interpolated using the mobility index for South Africa.

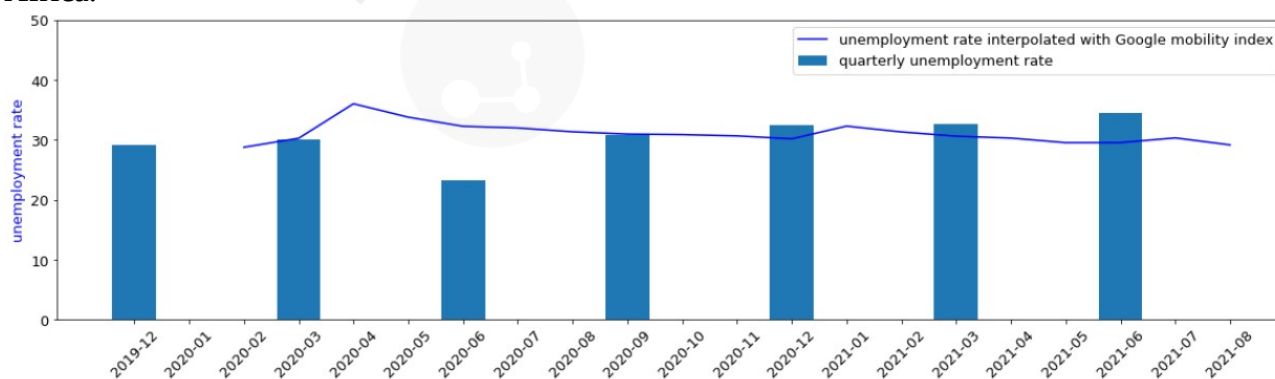


Figure 2: Unemployment rate of South Africa interpolated with mobility index

#### 3.2. Twitter Dataset and Sentiment Analysis

All the geotagged tweets posted from South Africa, except for retweets, since Jan 1<sup>st</sup>, 2016 until

August 31<sup>st</sup>, 2021 for certain keywords, namely, employed, unemployed, retrench, retrenched, retrenches, retrenching, retrenchment, and retrenchments were retrieved using full archive search and the Twitter Academic Researcher account. About 10 million tweets per month, from 2006 until today are retrievable from this archive. In order to do sentiment analysis, we prepared the dataset for Natural Language Processing (NLP) by cleaning the tweets and removing the duplicates, punctuations, URLs, and stop words. The dataset was divided into two parts. The first part contained tweets from Jan 1<sup>st</sup>, 2016 until Dec 31<sup>st</sup>, 2019, and the second part contained tweets from Feb 1<sup>st</sup>, 2020 until August 31<sup>st</sup>, 2021. The first part was used to analyze the tweets and their sentiments before the COVID-19 pandemic and the second part was used for the COVID-19 pandemic period.

Next, the trends of the number of tweets for both before and during the pandemic were compared with the unemployment rate trend of South Africa. Sentiment analysis was done using a pretrained model of BERT [3, 4]. The model was trained using a large Twitter dataset [5, 6]. We randomly chose 200 tweets from our dataset and manually labeled them as negative, neutral, or positive. We found that the model had 0.69 accuracy on our dataset. Next, the normalized sum of the sentiment scores over time was calculated for the two parts of the dataset and compared with unemployment rate, before and during the COVID-19 pandemic. Moreover, the sentiment classes and scores for different provinces were calculated and compared. The two datasets were concatenated to train the PCR model and estimate the unemployment rate. From the 34738 different tweets that were gathered, 13274 tweets belonged to the second part of the dataset, during COVID-19 pandemic, and the rest belonged to the first part, before COVID-19 pandemic. Figure 3 shows the word-cloud generated for our dataset. As can be seen in Figure 3, the most frequent words of our dataset are unemployed, employed, people, job, retrenchment, retrenched, need, and self-employed.



Figure 3: The most frequent words of our dataset

### 3.3. Nowcasting the Unemployment Rate

After concatenating the before and during the pandemic datasets, the number of tweets over time for the whole dataset and different keywords (i.e. employed, unemployed, and different forms of retrench) were found and stored in a vector. Next, since the normalized sum of the sentiment scores over time had a negative correlation with the unemployment rate, it was inverted and stored in a separate vector. These vectors made up the training set of the PCR. The unemployment rate was also stored in a different vector and used as labels for the PCR. The PCR method is essentially a linear regression model on the principal components of the training dataset [30]. Therefore, PCA was applied to all of the vectors of the training dataset, and five different principal components were found. According to Figure 4, the first component accounted for more than 80% of the variance. However, according to Figure 5, the cross-validation Root Mean Square Error (RMSE) indicated that

the least error is obtained when two of the principal components are used for linear regression. Therefore, we used linear regression with the first two principal components in our model.

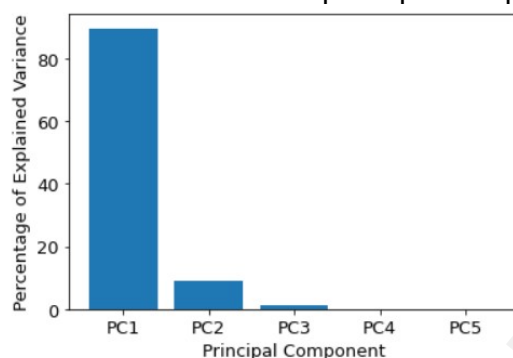


Figure 4: The percentage of explained variances for the five different principal components

No. of principal components used	RMSE_CV
1	1.245738
2	1.182032
3	1.237377
4	1.267425
5	1.411182

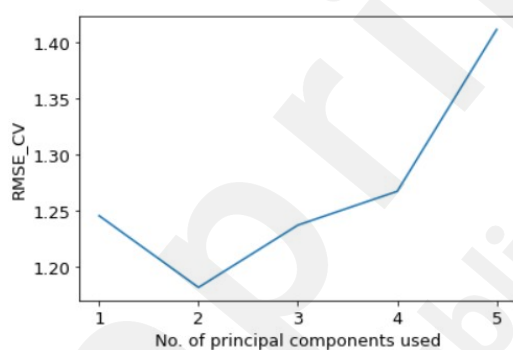


Figure 5: cross-validation RMSE for linear regression with different number of components

#### 4. Results

Tweets gathered with keywords employed, unemployed, and different forms of retrench, during COVID-19 pandemic, had the highest correlation with the interpolated unemployment rate created using the Google mobility index for South Africa. Figure 6 shows the trend of the number of tweets for the pandemic period. Tables 1 and 2 show the correlation and the p-values of the tweets with the interpolated unemployment rate and with each other, respectively. The results indicate that the keywords are related to each other and to the interpolated unemployment rate.

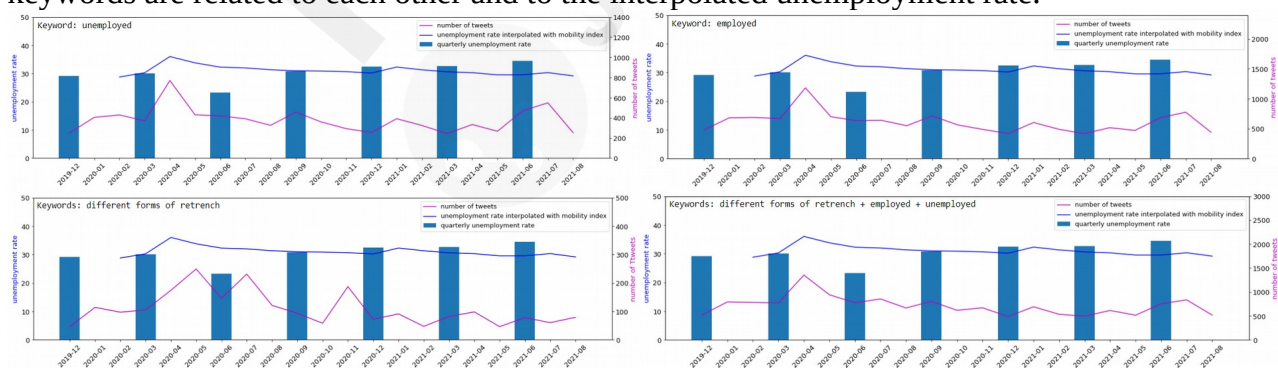


Figure 6: The second part of the dataset compared with unemployment rate during COVID-19 pandemic

Table 1: Correlation and p-values of different keywords with the interpolated unemployment rate

Keyword	employed	unemployed	different forms of retrench	Total Dataset
Correlation with				

Interpolated Unemployment Rate	0.66	0.63	0.62	0.74
P-value of the correlation	.001	.001	.001	.001

Table 2: Correlation and p-values of different keywords with each other

Keyword	employed		unemployed		different forms of retrench		total dataset	
	Correlation	P-value	Correlation	P-value	Correlation	P-value	Correlation	P-value
employed	1	0	0.98	.001	0.51	.01	0.98	.001
unemployed			1	0	0.47	.03	0.95	.001
different forms of retrench					1	0	0.68	.001
total dataset							1	0

Figure 7 shows the trend of the number of tweets compared with the quarterly unemployment rate of South Africa before COVID-19 pandemic. Tables 3 and 4 show the correlation of these keywords with the quarterly unemployment rate and with each other, respectively.

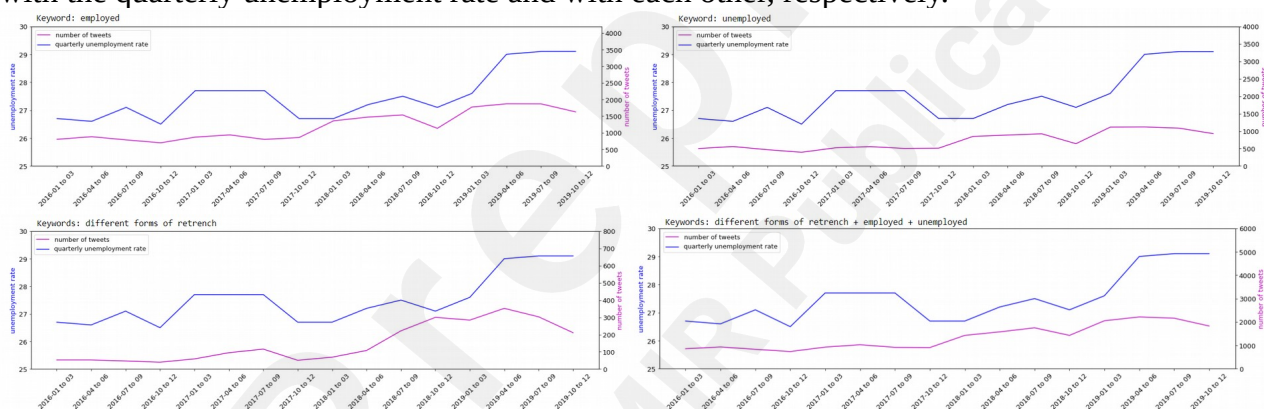


Figure 7: The first part of the dataset compared with unemployment rate before COVID-19 pandemic

Table 3: Correlation and p-values of different keywords with the unemployment rate before the COVID-19 pandemic

Keyword	employed	unemployed	different forms of retrench	total dataset
Correlation with Unemployment Rate over 2016-2019	0.7	0.66	0.71	0.72
P-value of the correlation	.001	.001	.001	.001

Table 4: Correlation and p-values of different keywords with each other before the COVID-19 pandemic

Keywords	employed		unemployed		different forms of retrench		total dataset	
	Correlation	P-value	Correlation	P-value	Correlation	P-value	Correlation	P-value
employed	1	0	0.99	.001	0.82	.001	0.99	.001
unemployed			1	0	0.8	.001	0.98	.001



different forms of retrench					1	0	0.89	.001
total dataset							1	0

According to these results, the tweets gathered using our selected keywords are significantly correlated with the unemployment rate of South Africa, during and before COVID-19, and therefore can be used to estimate the unemployment rate in real-time or find the missing data.

Next, sentiment classes and scores of the tweets were found using a pretrained model of BERT on the Twitter dataset [5]. As shown in Figure 8, more than 50% of the tweets were negative, during and before the COVID-19 pandemic. This outcome was expected as the dataset is on unemployment. Moreover, it can be seen in Figure 8 that sentiment classes are more negative and less positive and neutral during the COVID-19 pandemic compared to before. Figure 9 shows the distribution of the sentiment classes in different provinces of South Africa. As shown in Figure 9, Northern Cape has the least negative, the least positive, and the most neutral sentiments and Western Cape has the most positive and the least neutral sentiments. Moreover, Mpumalanga has the most negative sentiments.

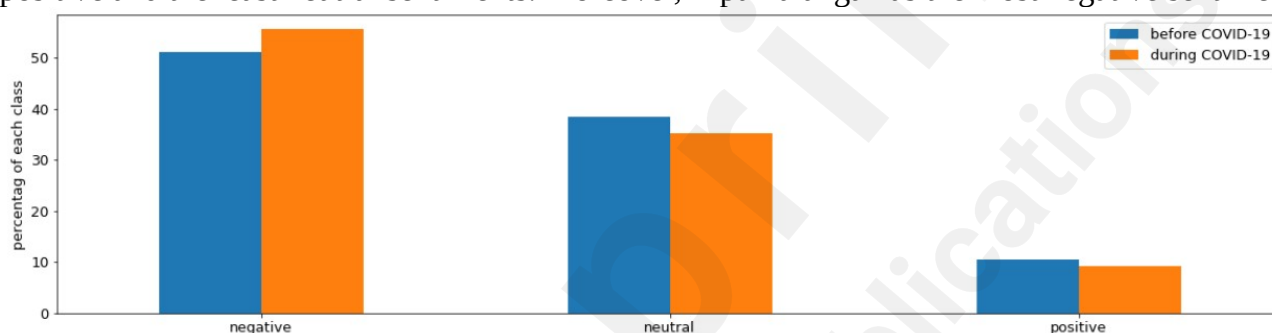


Figure 8: Percentage of tweets from different sentiment classes

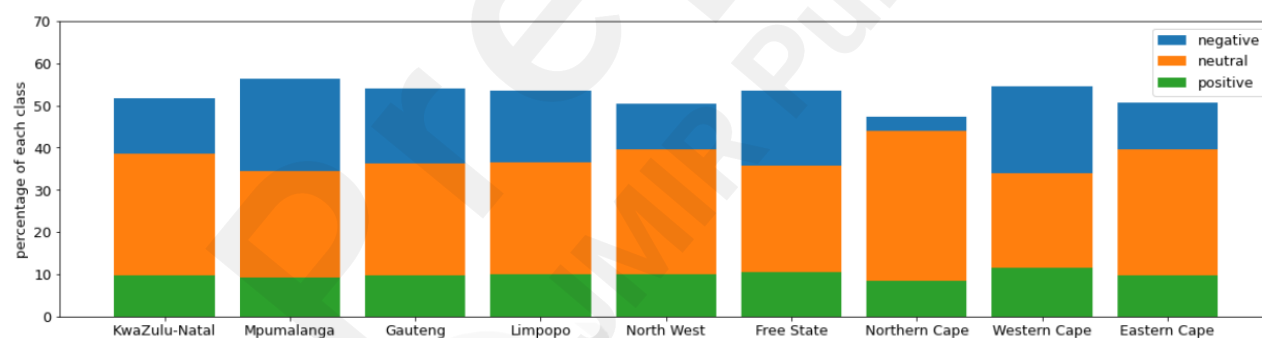


Figure 9: The distribution of sentiment classes on different provinces in percentage

We also calculated the sum of sentiment scores divided by the number of tweets, over time in urban and rural areas of South Africa. We associated Gauteng, KwaZuluNatal, and Western Cape as more urban and Eastern Cape, Northern Cape, Free State, Mpumalanga, and Limpopo as more rural areas of South Africa [31]. Figure 10 shows the sum of sentiment scores divided by the number of tweets in urban and rural areas.

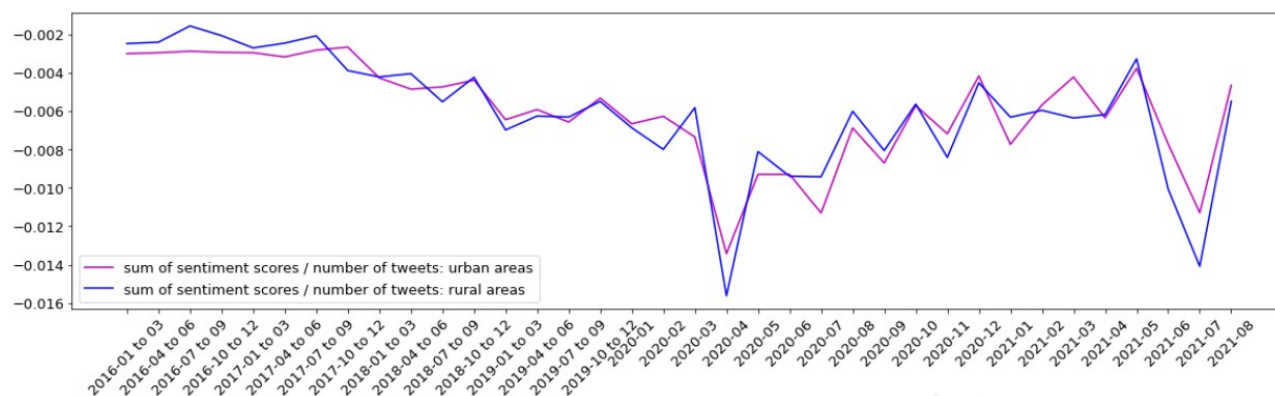


Figure 10: Sum of sentiment scores divided by number of tweets for urban and rural areas

According to Figure 10 sentiments for rural and urban areas are very close to each other over time. However, during the first lockdown in April 2020, and the country-wide riots in July 2021, the sentiments of rural areas were more negative than the urban areas. This may indicate that during these times, people from rural provinces suffered more and were more dissatisfied with the conditions.

Next, we compared the normalized sum of sentiment scores with the unemployment rate, during and before COVID-19. Figure 11 shows the distribution of the normalized sum of the sentiment scores during the COVID-19 pandemic, over time, and Figure 12 shows that for the pre pandemic period. Table 5 shows the correlation and the p-value of sentiment scores with different keywords for the first and second part of the dataset, i.e. before and during the COVID-19 pandemic.

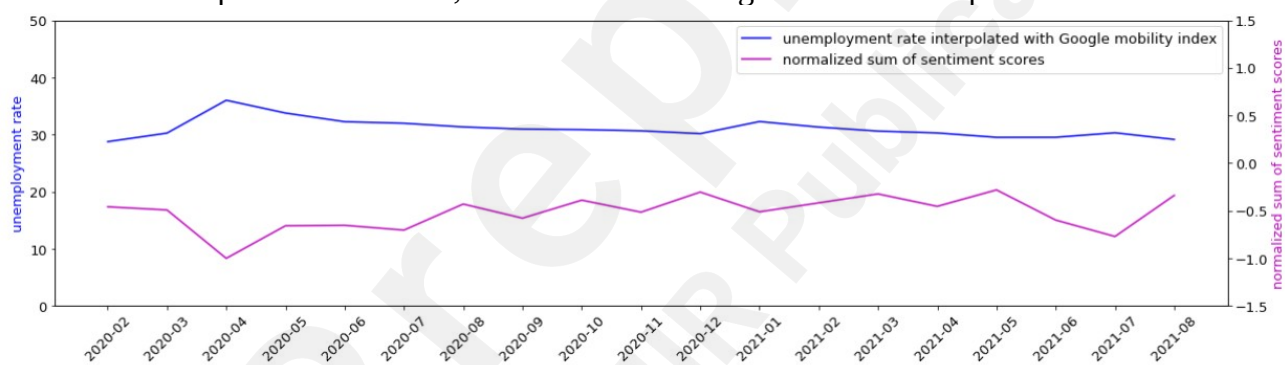


Figure 11: The distribution of sum of sentiment scores over time during the COVID-19 pandemic

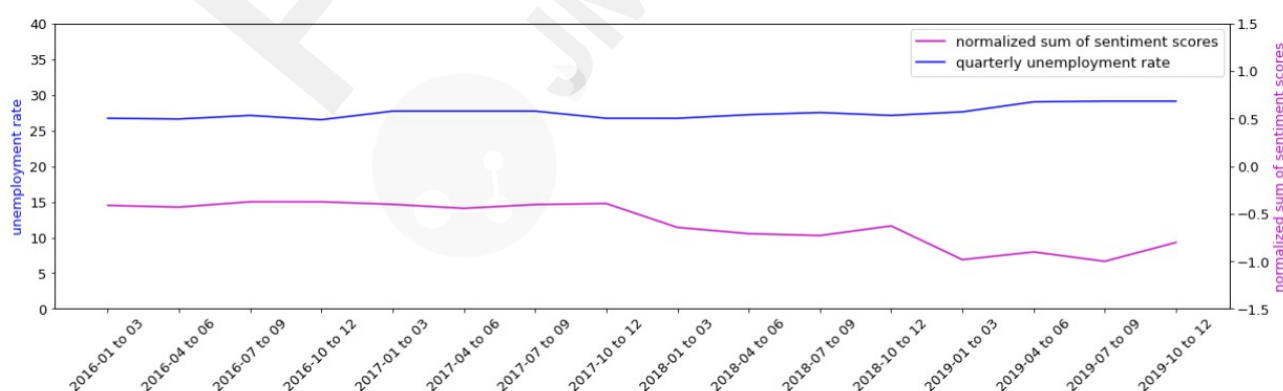


Figure 12: The distribution of sum of sentiment scores over time before the COVID-19 pandemic

Table 5: Correlation and p-values unemployment rate and different keywords with the sum of sentiment scores

Keyword	employed	unemployed	different forms of retrench	total dataset	Unemployment Rate
---------	----------	------------	-----------------------------	---------------	-------------------

	Corr .	P-value	Corr .	P-value	Corr .	P-value	Corr. .	P-value	Corr .	P-value
Sentiment scores during COVID-19	-0.9	.001	-0.92	.001	-0.56	.01	-0.93	.001	-0.71	.001
Sentiment scores before COVID-19	-0.98	.001	-0.98	.001	-0.86	.001	-0.99	.001	-0.67	.001

According to Figures 11 and 12 and Table 5, the sentiment scores have a high negative correlation with the unemployment rate, during and before the COVID-19 pandemic. This outcome can show that the higher the unemployment rates are, the more negative the sentiments will be or the more dissatisfied people will become.

Finally, the pre pandemic and during the pandemic datasets were concatenated and used to nowcast the unemployment rate. Table 6 shows the correlation and p-values of the concatenated dataset with the unemployment rate. According to table 6 the concatenated dataset has a strong correlation with the unemployment rate.

Table 6: Correlation and p-values of unemployment rate with the concatenated dataset

Keyword	employed		unemployed		different forms of retrench		total dataset		sentiment scores	
	Corr .	P-value	Corr .	P-value	Corr .	P-value	Corr .	P-value	Corr .	P-value
Unemployment Rate	0.77	.001	0.76	.001	0.78	.001	0.83	.001	-0.84	.001

The sentiment scores were next inverted to have a positive correlation with the unemployment rate. Two-thirds of the dataset were used for training the PCR model and the remaining portion was used for testing. According to Figure 13, the predicted values of unemployment rate were very well correlated with the actual values. The RMSE and Mean Absolute Error (MAE) metrics stated in Figure 13 were calculated using Equations 2 and 3.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (A_i - P_i)^2}{n}} \quad (2)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n \frac{|A_i - P_i|}{A_i} \quad (3)$$

Where  $n$  is the number of tested values,  $A$  is the actual unemployment rate, and  $P$  is the predicted value. To verify that the unemployment rate during the COVID-19 pandemic can indeed be estimated using the number of tweets on the selected keywords combined with their sentiments, we then used the first 20 samples of the dataset for training the PCR and the remaining samples for testing. Figure 14 shows that the estimated unemployment rate moves closely alongside the actual unemployment rate. Figure 15 shows that the estimated unemployment rate is correlated with the actual unemployment rate, during the COVID-19 pandemic.

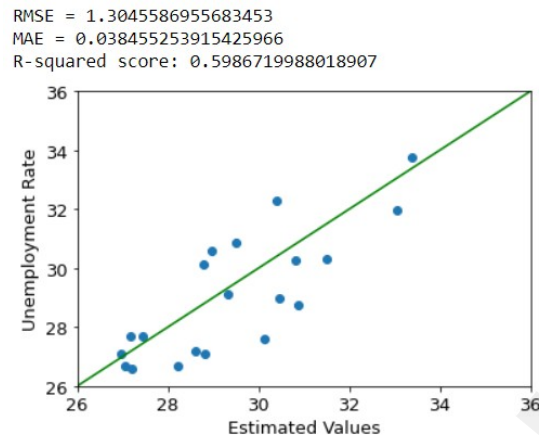


Figure 13: estimated values of the unemployment rate are very well correlated with the actual unemployment rate

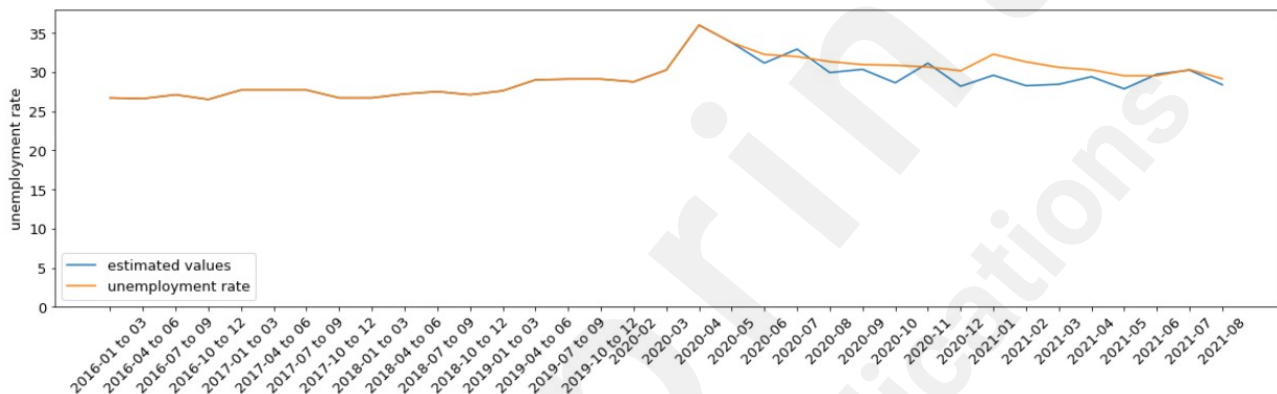


Figure 14: estimated values of unemployment rate during COVID-19 pandemic matches the actual unemployment rate

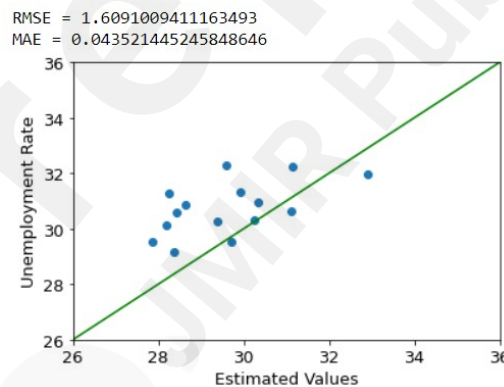


Figure 15: estimated values of unemployment rate during COVID-19 pandemic is correlated with the actual unemployment rate

As shown in Figure 15, the RMSE and MAE of our method are about 1.6 and 0.04, respectively. This error accounts for about 4% of the actual unemployment rate on average. Moreover, according to Figure 13, the r-squared score of our method is about 0.6 which indicates that our method has a good effect size.

### 5. Limitations

The diagrams show that our proposed method underestimates the actual unemployment rate. The reason lies down in the nature of the tweets that we gathered. Figure 16 shows the percentage of tweets of each province.



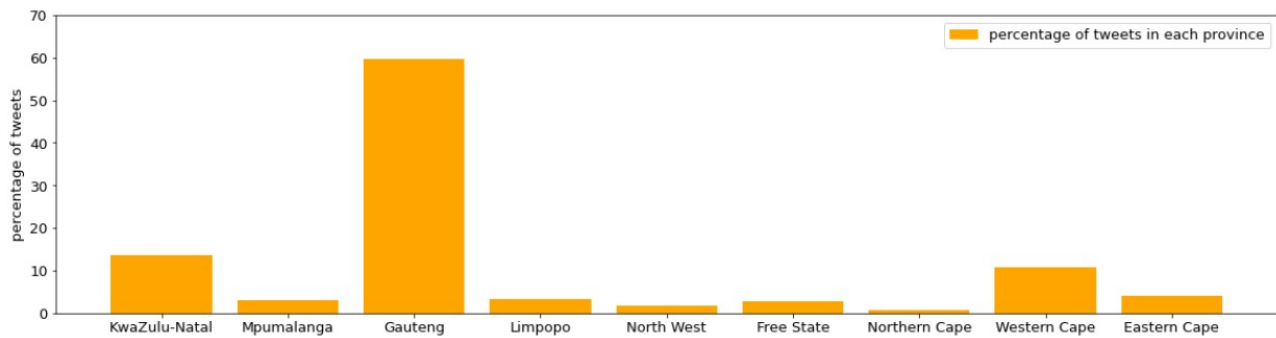


Figure 16: percentage of tweets from each province

Figure 16 clearly shows that more than 50% of the tweets come from Gauteng, the major industrialized province of South Africa. Moreover, tweets from KwaZulu-Natal and Western Cape, the other two urban areas, are more than other provinces. One reason for this is that we gathered English tweets. South Africa has eleven official languages [32]. English is mostly spoken in Urban areas of South Africa [31]. For this reason, most of the users of our dataset are from wealthier populations and belong to mid to high income groups. As a result, our PCR model under-estimates the unemployment rate of South Africa in at least some areas.

## 6. Discussion

Metrics such as unemployment rate are an indication of people's lives, concerns, and ideals. Social media is turning into the primary place where people share their thoughts and daily activities. It is very rational that the discussions on social media reflect statistical measurements such as unemployment rate. We can use this reflection to understand and estimate the unemployment rate. We just need to find out where to look for it. In other words, we need to find the right keywords that will allow us to catch this reflection.

In this study, we showed that keywords such as employed, unemployed, retrench, retrenched, retrenching, retrenches, retrenchment, and retrenchments enable us to receive the unemployment rate in South Africa. The selected keywords correlate with the unemployment rate during and before COVID-19. Therefore, it is very likely that the number of tweets gathered with these keywords will keep on correlating with the unemployment rate of South Africa, in the future. Moreover, the fact that the normalised sum of the sentiment scores of the tweets gathered with these keywords have a strong negative correlation with unemployment rate verifies that these keywords are able to reflect the unemployment rate of South Africa in social media. As the unemployment rate increases, people begin to talk about it in the social media, in a negative way, and the selected keywords are able to pick this reflection.

Our PCR method for estimating unemployment rate using the number of tweets on the selected keywords and the normalised sum of their sentiments has a fair RMSE and a reasonable r-squared score. The reason for this under-estimation is that we gathered English tweets only. However, South Africa has eleven official languages, among which English is mostly spoken in urban areas. For this reason, our dataset comes mostly from the wealthier population. This has caused our PCR model to under-estimate the unemployment rate of South Africa.

In conclusion, our PCR method is able to estimate the unemployment rate of South Africa reasonably well. This is very valuable as it allows us to remove the barriers and difficulties of census methods and estimate the unemployment rate in real-time, especially during the COVID-19 pandemic that unemployment has turned into a crisis around the world.

## 7. Conclusion

In this paper, social media, particularly, twitter was traced to estimate the unemployment rate of South Africa in real-time. Since in South Africa unemployment rate is measured quarterly, this method can be used to find the missing information on unemployment rate, as well. Moreover, this method can provide the unemployment rate statistics in real-time, and without the difficulties of

census taking methods. Finally, this information can be highly valuable for analyzing labor market flow when facing disasters such as a pandemic.

In our method, we found employed, unemployed, retrench, retrenched, retrenching, retrenches, retrenchment, and retrenchments to be the keywords that correlated the most with the unemployment rate of South Africa before and during COVID-19. The normalized sum of sentiment scores over time, before and during the COVID-19 pandemic had a strong negative correlation with unemployment rate. We combined the number of tweets on different keywords, and the sentiment scores and used PCR to nowcast the unemployment rate. The results show that the estimated unemployment rate was well correlated with the actual unemployment rate.

One contribution to the future work of this project is to use social media to estimate other economic metrics such as inflation rate, job vacancy rate, labour force participation rate, and part time working rate. Another work that can be done is to use social media to forecast economic metrics such as unemployment rate. Different methods or techniques of time series prediction or data mining and machine learning algorithms can be used to forecast these metrics. This can be extremely useful for disaster management response and recovery. Finally, since other media, especially images and videos make up a large portion of the social media, new methods need to be proposed to process social media content further.

### References

- [1] Google COVID-19 Community Mobility Reports, See how your community is moving around differently due to COVID-19, <https://www.google.com/covid19/mobility/>
- [2] International Labour Organization, ILO Monitor: COVID-19 and the world of work. Seventh edition Updated estimates and analysis, ILO, Jan 2021, [https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/briefingnote/wcms\\_767028.pdf](https://www.ilo.org/wcmsp5/groups/public/---dgreports/---dcomm/documents/briefingnote/wcms_767028.pdf)
- [3] J. Devlin, M.-W. Chang, K. Lee, et al, BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding, arXiv, Computation and Language (cs.CL), May 2019, <https://arxiv.org/abs/1810.04805>
- [4] H. Xu, L. Shu, P. S. Yu, et al, Understanding Pre-trained BERT for Aspect-based Sentiment Analysis, Computational Linguistics, Spain, Dec 2020, PP. 244-250,
- [5] Hugging Face, Twitter-roBERTa-base for Sentiment Analysis, <https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment>
- [6] F. Barbieri, J. Camacho-Collados, L. Neves, L. Es[inos]a-Anke, TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification, arXiv, Oct 2020.
- [7] D. Antenucci, M. Cafarella, M. C. Levenstein, et al, Using Social Media to Measure Labor Market Flows, National Bureau of Economic Research, March 2014, <https://www.nber.org/papers/w20010>
- [8] V. Tzinovits, Using Social Media to Measure Labour Market Flows in Greece, Thesis, Master of Science, Applied Economics and Data Analysis, School of Business Administration, Department of Economics, Aug 2016.
- [9] D. Rizio, T. Suryavanshi, M. Yahya, et al, Can We Use Twitter to Track COVID-caused Unemployment in the USA?, Data Science Report, May 2021.
- [10] R. Mokhlesur, A. Nawaz, J. L. Xue, et al, Socioeconomic factors analysis for COVID-19 US reopening sentiment with Twitter and census data, Heliyon, Feb 2021, doi: 10.1016/j.heliyon.2021.e06200.
- [11] Y. Su, A. Venkat, Y. Yadav, et al, Twitter-based analysis reveals differential COVID-19 concerns across areas with socioeconomic disparities, Comput Biol Med, May 2021, doi: 10.1016/j.combiomed.2021.104336.
- [12] D. Altig, S Baker, J. M. Barrero, et al, Economic uncertainty before and during the COVID-19 pandemic, Science Direct, Journal of Public Economics, V. 191, Nov. 2020, doi: 10.1016/j.jpubeco.2020.104274.

- [13] C. R. Nirmala, G. M. Roopa, K. R. Naveen Kumar, Twitter data analysis for unemployment crisis, IEEE, Applied and Theoretical Computing and Communication Technology, Apr. 2016, doi: 10.1109/ICATCCT.2015.7456920.
- [14] H. Lee, N. Lee, H. Seo, et al, Developing a supervised learning-based social media business sentiment index, Springer, Supercomputing, Jan 2019, doi: 10.1007/s11227-018-02737-x.
- [15] P.-M. Ryu, Predicting the Unemployment Rate Using Social Media Analysis, J of Information Processing Systems, V. 14, no. 4, Aug. 2018, doi: 10.3745/JIPS.04.0079.
- [16] D. Proserpio, S. Counts, A. Jain, et al, The psychology of job loss: using social media data to characterize and predict unemployment, ACM, WebSci, 2016, doi: 10.1145/2908131.2913008.
- [17] E. Bokanyi, Z. Labszki. G. Vattay, Prediction of employment and unemployment rates from Twitter daily rhythms in the US, Springer, EPJ Data Sci, V. 6, no. 14, 2017, doi: 10.1140/epjds/s13688-017-0112-x.
- [18] A. Suomi, T. P. Schofield, P. Butterworth, Unemployment, Employment and COVID19: How the Global Socioeconomic Shock Challenged Negative Perception Toward the Less Fortunate in the Australian Context, Frontiers, Psychology, Oct 2020, doi: <https://doi.org/10.3389/fpsyg.2020.594837>.
- [19] S. Zervoudakis, E. Marakakis, H. Kondylakis, et al, OpinionMine: A Bayesian-based framework for opinion mining using Twitter Data, Elsevier, Machine Learning with Applications, V. 3, March 2021, doi: 10.1016/j.mlwa.2020.100018.
- [20] N. Aguilar-Gallegos, L. Klerkx, L. E. Romero-Garcia, et al, Social network analysis of spreading and exchanging information on Twitter: the case of an agricultural research and education centre in Mexico, Taylor & Francis, J of Agricultural Education and Extension, Apr 2020, doi: 10.1080/1389224X.2021.1915829.
- [21] L. Liu, B. K P Woo, Twitter as a Mental Health Support System for Students and Professionals in the Medical Field, JMIR Medical Education, V. 7, no. 1, 2021, doi: 10.2196/17598
- [22] A. Prada, C. A. Iglesias, Predicting Reputation in the Sharing Economy with Twitter Social Data, MDPI, Applied Science, V. 10, no. 8, March 2020, doi: 10.3390/app10082881.
- [23] D. Valle-Cruz, V. Fernandez-Cortez, A. Lopez-Chau, et al, Does Twitter Affect Stock Market Decisions? Financial Sentiment Analysis During Pandemics: A Comparative Study of the H1N1 and the COVID-19 Periods, Cogn Comput, 2021, 10.1007/s12559-021-09819-8.
- [24] J. Michalak, Does pre-processing affect the correlation indicator between Twitter message volume and stock market trading volume?, APCZ, Economics and Law, V. 19, no. 4, 2020, doi: 10.12775/EiP.2020.048.
- [25] GlobalStats, Social Media Stats in Africa, <https://gs.statcounter.com/social-media-stats/all/africa>, July 2021, Accessed: Aug 2021.
- [26] T. E. Bosch, M. Admire, M. Ncube, Facebook and politics in Africa: Zimbabwe and Kenya, Sage, Media, Culture and Society, V. 42, no. 3, Apr 2020, doi: 10.1177/0163443719895194.
- [27] C. O. Adekoya, J. K. Fasae, Social media and the spread of COVID-19 infodemic, emerald, Global Knowledge, Memory and Communication, V. 70, Apr 2021, doi: 10.1108/GKMC-11-2020-0165
- [28] G. Stevens, H. Ishizawa, D. Grbic, Measuring race and ethnicity in the censuses of Australia, Canada, and the United States: Parallels and paradoxes, Canadian Studies in Population, V. 42, no. 1-2, 2015, doi: 10.25336/P6PW39.
- [29] C. Skinner, Issues and Challenges in Census Taking, Annual Review of Statistics and Its Application, V. 5, 2018, doi: /10.1146/annurev-statistics-041715-033713.
- [30] H. Lee, Y. M. Park, S. Lee, Principal Component Regression by Principal Component Selection, Commun. Stat. Appl. Methods, 2015, doi: 10.5351/CSAM.2015.22.2.173.
- [31] D. F. Gordon, A. Nel, A. S. Mabin, R. Vigne, C. J. Bundy, L. M. Thompson, J R.D. Cobbing, M. Hall, C. C. Lowe, South Africa, Encyclopedia Britannica, Sep. 2021, <https://www.britannica.com/place/South-Africa>. Accessed 9 September 2021.

[32] M. Brenzinger, Eleven Official Languages and More: Legislation and Language Policies in South Africa, *Revista de Llengua i Dret, Journal of Language and Law*, no. 67, 2017, p. 38-54. doi: 10.2436/rld.i67.2017.2945.

