African Journal of Science, Technology, Innovation and Development

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/rajs20

Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach

Murimo Bethel Mutanga & Abdultaofeek Abayomi

To cite this article: Murimo Bethel Mutanga & Abdultaofeek Abayomi (2020): Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach, African Journal of Science, Technology, Innovation and Development, DOI: 10.1080/20421338.2020.1817262

To link to this article: https://doi.org/10.1080/20421338.2020.1817262

African Journal of Science, Technology, Innovation and Development, 2020
https://doi.org/10.1080/20421338.2020.1817262

Routledge
Taylor & Francis Group

Check for updates

Tweeting on COVID-19 pandemic in South Africa: LDA-based topic modelling approach

Murimo Bethel Mutanga and Abdultaofeek Abayomi*

Department of Information and Communication Technology, Mangosuthu University of Technology, Durban, South Africa
**Corresponding author email: taofeekab@yahoo.com*

The advent of COVID-19 has disrupted all facets of human lives. As of September 2020, there is no effective viral therapy for the disease, thus necessitating research efforts toward providing solutions to the diverse areas where the pandemic has wreaked havoc. As a way of reducing the spread of the disease, the South African government declared COVID-19 a national disaster and implemented nationwide lockdowns with several regulations. Nevertheless, the success of such synergized efforts primarily depends on the people's attitudes and perceptions toward the multifaceted management of the pandemic. Therefore, this current study aims to discover what topical issues relating to the pandemic are being discussed by the populace and what impacts these issues have on compliance with regulations, including how they can aid in the implementation of the measures put in place by the government, as we analyze discussions relating to COVID-19 using data harvested from Twitter – social media and opinion mining platform. The Latent Dirichlet Allocation (LDA) algorithm was applied for the extraction of noteworthy topics. From the experiments conducted, it was observed that alcohol sale and consumption, staying home, daily statistics tracing, police brutality, 5G and vaccines conspiracy theories were among the topics discussed and around which attitudes and perceptions were formed by the citizens. The findings also revealed people's resistance to measures that affect their economic activities, and their unwillingness to take tests or vaccines as a result of fake news and conspiracy theories. These findings can assist the government and policymakers in redirecting their efforts by addressing the citizens' concerns and reactions to the instituted measures toward an anticipated overall success.

Keywords: COVID-19, LDA, social media, topic modelling, Twitter, vaccine

Introduction

COrona **VI**rus **D**isease – 20**19** (**COVID-19**) is a novel viral pandemic that broke out in December 2019 in Wuhan, Hubei province, China (Anastassopoulou et al. 2020) and later spread across the globe majorly through human-human transmissions. These could mostly be through respiratory droplets from sneezes and coughs within a 1-metre distance of an infected person and via contact with contaminated surfaces (Chan et al. 2020). According to the Centers for Disease Control and Prevention (CDC), the symptoms include fever, fatigue, dry cough and shortness of breath after a period believed to range from 2 to 14 days following infection. Some patients may also experience a runny nose, nasal congestion, sore throat, muscle pain and diarrhoea. These symptoms which could be mild or severe can lead to pneumonia, kidney failure, severe acute respiratory syndrome (SARS) and even death (Centers for Disease Control 2020). The World Health Organization declared the disease a global pandemic on 30 January 2020 after 7818 global active cases were recorded. Consequently, countries across the globe have called for collaborative efforts towards fighting the spread of this pandemic, the main objective, being the reduction of casualties and easing the burden on the already overwhelmed health workers and facilities.

As of 5 October 2020 (12:23:00), the global confirmed cases of COVID-19 stood at 35,220,166 with 1,037,604 deaths recorded while 24,533,728 infected people requiring no further hospitalization were discharged from the hospitals (JHU 2020; WHO 2020). However, in South Africa, at the same juncture, there were 681,289 confirmed cases and 16,976 deaths. In this study we focused on South Africa because it has the highest documented confirmed cases and deaths on the continent while it ranked fifth position (as at 15 August 2020) in global confirmed cases (JHU 2020; WHO 2020). The country also topped the list of percentage of global daily new infection cases and was second in the percentage of daily deaths on 14 June 2020 (Levitt 2020; Worldometers 2020). The negative impacts of this pandemic on the populace and the toll it has taken on all aspects of human lives necessitate research efforts capable of stemming the tide, managing both the healthy and infected people as well as their various concerns, reactions and perceptions, while offering useful lessons to handle the current outbreak and similar ones in the future.

Although, to date – 5 October 2020 – there is no effective viral therapy or vaccine for COVID-19, many countries including Russia, the United Kingdom, the United States of America, China, France, etc. are now approaching the final phase of clinical trials and testing of various vaccine prototypes before obtaining licences for a roll-out. However, governments across the globe are adopting sanitization and other recommendations by the WHO as preventive measures to control the spread of the virus. These measures include avoiding physical contact with persons displaying symptoms of the disease, washing of hands regularly with soap and water or an alcohol-based hand-sanitizer and, most importantly covering the mouth and nose when sneezing or coughing (Tomasi 2020).

In most countries, these measures alone were found to be insufficient. Consequently, many governments instituted lockdown procedures and forced the general population to stay at home.

Although these measures come with adverse effects on the nations' and the world's economy at large, their success in containing the spread of the virus is evident.

However, the lockdown measures allowed rapid clinical tests to be conducted on the populace, quarantined programmes to be monitored while also enabling contacts tracing of those that might have come into contact with infected persons. Furthermore, some countries such as Italy and China reported huge success after implementing such measures. In South Africa, however, these measures were met with mixed feelings due to the resultant social and economic impacts they have on the populace. Consequently, enforcing lockdown measures have been quite challenging within some communities. To further exacerbate the problem, a lot of conspiracy theories with regards to the origin, causes and management of the pandemic have emerged. What is critical to note in this context is that the success of the implemented measures, to a large extent, depend on the concerns and perceptions of the populace being addressed and their reactions to the instituted measures. Therefore, we intend in this study to discover what topical issues relating to the pandemic are being discussed by the South African populace and what impacts these issues have on their compliance with the regulations, including how they can aid in the implementation of the measures put in place by the government to manage the pandemic.

In today's world, social media platforms allow for free interaction and exposing viewpoints; hence, there are a lot of discussions on the COVID-19 pandemic on social media platforms. In essence, the emergence of mobile and ubiquitous computing technologies has led to overabundant usage of online social and news media (Oppegaard and Rabby 2016). Furthermore, the advent of big data analytics enables the storage and analysis of more and more user-generated content on blogs and social networks. Some of the most popular social media include Facebook, Twitter, Instagram and Whatsapp. Twitter, established in 2006, currently stands as the leading microblogging platform that allows sending or receiving of about 280 characters per tweet and also to upload photos or short videos. The microblog has more than a billion visits per month, 500 million tweets per day and over 330 million active users (Java et al. 2007; Omnicore 2020). Information sharing, news reporting and conversations are the various communication forms on Twitter, and users discuss various health-related issues (Yin et al. 2015), among other trending topics on the platform. This makes the microblog a viable source of data for predicting trends and to analyze human sentiments. Besides, traditional survey methods of collecting and analyzing data can be costly, resource-intensive, and time-consuming. However, with natural language processing (NLP), the task of gathering and analyzing textual data is feasible and affordable. Analysis of natural language, however, comes with unique challenges due to diverse linguistic properties including abbreviation, spelling mistakes, punctuations, stop words and non-standard text.

Several research studies have applied NLP techniques to social media content to extract and classify trending topics. For instance, the research works in (Culotta 2012; Lamb, Paul, and Dredze 2013; Aramaki, Maskawa, and Morita 2011) analyzed Twitter data for the detection of influenza epidemics. In addition, topic modelling has proven to be effective for analyzing similar trends in various applications. This current study, therefore, applies the same to the COVID-19 pandemic in South Africa using Twitter data.

Topic models are statistical and unsupervised models capable of discovering the latent topics within large corpora of documents. In essence, topic modelling algorithms classify groups of documents into coherent themes without user intervention. Topic modelling is considered a clustering problem consisting of a random mixture of topics in a document, and each word presumed to belong to at least one of those topics (Kanan et al. 2015; Ostrowski 2015). The resulting topics can be very general or very specific, depending on the chosen input parameters (Blair, Bi, and Mulvenna 2020). The LDA algorithm (Kanan et al. 2015) has been used successfully in many application areas. Its pattern-discovery power with user-provided guidance proved to be very attractive to practical users of topic modelling as we applied it in this current study to extract latent topics relating to the COVID-19 pandemic in South Africa.

Further to this introductory section, the rest of this paper is structured as follows. Research studies relating to topic modelling are discussed in the next section, with the section on materials and methods following, wherein the dataset utilized and the LDA method applied, including the experiments, are described. The penultimate section presents the results obtained and the corresponding discussions, while the conclusions and future directions are given in the final section.

Literature review

With the increase in the use and adoption of social networks and online blogs, mechanisms to analyze unstructured data are becoming popular. For instance, hotel or restaurant customer reviews can be a good source of feedback from customers (Chatterjee 2019; Tsai et al. 2020), although the data is unstructured. Furthermore, the increasing use of Twitter for echoing political views and sentiments by both politicians and ordinary citizens has made it an essential part of the networked sphere in which political issues are discussed (Bode and Dalrymple 2016). It has, thus become crucial for politicians and political analysts to extract and analyze social media data. Topic modelling has become a de-facto technique for achieving this. Some of the topic modelling approaches that have been applied in the literature include; the Latent Dirichlet Allocation (LDA), Document-pivot topic detection (Doc-p) (Winarko and Pulungan 2019), Graph-based feature-pivot topic detection (GFeat-p), Frequent Pattern Mining (FPM), Soft Frequent Pattern Mining (SFPM) (Zendah and Maghari 2019) and the BNgram approach (Ibrahim et al. 2018).

The LDA algorithm is the most popular and widely used topic modelling method for extracting main topics, thus we interchangeably use topic modelling and topic extraction to mean the same thing in this current study. The LDA is an unsupervised machine learning algorithm which identifies latent topic information among large document collections. This technique relies on a 'bag of words' approach, which treats each document as a

vector of word counts. When generating a topic model using the LDA, many parameters have to be set, which affects the coherence of the generated topics. Without prior knowledge of the corpus being modelled, setting a small number of topics will result in broad topics that contain mostly common words and stop words (Blair, Bi, and Mulvenna 2020). It is, therefore, imperative to optimize the number of extracted topics. Popularized in text retrieval applications, the LDA has been applied with excellent performance among entity resolution systems, fraud detection in telecommunication systems and image processing. The Doc-p approach utilizes Locality Sensitive Hashing (LSH) (Petrović, Osborne, and Lavrenko 2010) to extract the nearest neighbour in a document and enhances the clustering task. The LSH has the advantage of quickly producing the nearest neighbours using cosine similarity in a large collection of documents.

On the other hand, the feature-pivot method, such as GFeat-p achieves feature clustering by using the structural clustering algorithm for networks (SCAN). The SCAN algorithm (Xu et al. 2007) can detect the number of nodes and also offers a number of hubs that the groups of nodes can connect to. In a feature-pivot method, these groups represent the topics while the nodes are the terms. Similarly, the FPM approach was developed to address the generic or merged topics drawback of the GFeat-p. For topic modelling, the FPM utilizes the frequent itemset mining technique which determines the items that are likely to co-occur in a set of transactions (Goethals 2005). The SFPM approach (Zendah and Maghari 2019) investigates co-occurrence patterns between sets of terms such that the cardinality is greater than two but without the need that all the terms in these sets co-occur frequently. A set of terms would, therefore, be built and new terms added greedily based on how often they co-occur with the term. However, for the BNgram approach, a similar result could be achieved just like those of the FPM and SFPM techniques that considered co-occurrence between more than two terms, but in an easier way. BNgram approach naturally groups together co-occurring terms, and it may be considered to offer a first level of term grouping. It is thus suited for Twitter data since a large number of status updates are just copies or retweets of previous messages, so relevant n-grams tend to be frequent. The method is also capable of yielding good performance (Ibrahim et al. 2018). These topic modelling approaches have been utilized in diverse application areas including business, medicine, computational journalism and politics. Some of the relevant works are here discussed.

Kagashe, Yan, and Suheryani (2017) utilized Twitter data during the 2012/2013 flu season and applied the Latent Dirichlet Allocation (LDA) technique to mine trending topics on widely used therapeutic drugs. The authors constructed a machine learning classifier to extract tweets from which the most consumed flu drugs were identified while using dependency words as features. Similarly, Ostrowski (2015) explored topic modelling using the LDA technique and evaluated the performance of the proposed approach using a filtered collection of customer relationship management related Twitter messages. The method was found effective for the identification of sub-topics and to support topics classification for a huge corpus. In another study, Robertson and Yee (2016) also detected topics at the peak of an avian influenza outbreak by utilizing official reports mined using Twitter data. The authors proposed a real-time risk surveillance system by developing methods capable of analyzing data as it arrives. Cole-Lewis et al. (2015) extracted topics from tweets relating to electronic cigarettes for sentiment and content analysis using supervised machine learning method. However, Maskeri, Santonu, and Kenneth (2008) applied LDA to business-related topics from programme source codes to assist in software maintenance. Text mining and sentiment analysis techniques were utilized by Wu et al. (2016) to analyze online reviews of selected hotels in China to gain insights for informed decision making, strategic planning, enhanced operation and service delivery. On the other hand, the LDA, incremental Gibbs samplers and particle filters were used by Canini, Lei, and Thomas (2009). In this research, the authors applied inference algorithms to extend a batch algorithm into a series of the online algorithm while maintaining a higher level of flexibility. In another study, an ensemble model comprising of LDA and LSA (Latent Semantic Analysis) was proposed by (Shen et al. 2010). In this model, various document-word tuples co-occurring in documents and their assigned topics were utilized to derive topic distribution over documents subsequently. Experiments were conducted using both real and synthetic data and evaluated with the perplexity metric, which indicates the LDA ensemble outperforming the LSA method but in terms of efficiency, the LSA performance is better. Furthermore, research studies have also been conducted to bring together two separate LDA models in a two-fold method, one model for aspects and another for the sentiment (Burns et al. 2011). That method was extended further in a related study to allow for multiple aspects in a sentence as the initial version assumed only one aspect per sentence (Burns et al. 2012). Hong, Ovidiu, and Brian (2011) also applied LDA to extract topics from Twitter data by training a topic model on aggregated measures. The topic mixture distributions learned by topic models was shown to offer additional features in classification tasks and capable of improving performance. Meanwhile, Rider and Chawla (2013) combined multiple topic models using medical data to predict a patient's risk of disease. The ensemble utilized component models and trained them with various datasets such as race, gender, poverty level and age while aggregating the topics discovered in each model into a single matrix. The aggregated topic matrix was then used to obtain a distribution over new patients to predict the disease risk they are susceptible to. Also, Tian, Meghan, and Denys (2009) compared the performance of LDA with other machine learning methods to classify 41 software into their respective domain categories to enable definitions based on programming languages, architectures and libraries. In addition, an online LDA-based topic model was developed by Wang, Eugene, and Michele (2012) and applied on 30 million social media postings. Findings indicate that the method

revealed good topic transitions, including work to life rhythm of cities and factors that are associated with the particular complaints.

In this current study, we investigated: (i) what topical issues relating to the pandemic are being discussed by the South African populace and how they are extracted; (ii) what impacts the content of the discussions have on the populace's compliance with the regulations; and (iii) how the issues discussed can aid in the implementation of the measures put in place by the government to manage the pandemic. The LDA algorithm was applied for topic extraction and we subsequently obtained a word cloud indicating the trending words relating to the COVID-19 pandemic in South Africa among the populace. Though some of the literature reviewed also applied the LDA algorithm but among the gaps that we intend filling is that our study focuses particularly on COVID-19 pandemic in South Africa. To the best of our knowledge, this current research is among the few (if at all there are any earlier such studies) that have been conducted on the COVID-19 in the South African context.

Materials and method

Data acquisition

As the dataset utilized for our experiments, we mined Twitter messages between 15 March 2020 – 30 April 2020. During this period, the virus had spread to all the continents and had killed more than 220,000 people (WHO 2020). Since our focus is on South African populace, we combined two different hashtags – #COVID19SA, or #CoronaVirusSA to extract the initial data. To ensure that the data was exhaustive and an accurate representation of the discussed topics, we further obtained other related hashtags from the initially mined data. These new hashtags which include #LOCKDOWNFOOD, #LOCKDOWNINSOUTHAFRICA, #LOCKDOWNNOW, #SHUTDOWNSOUTHAFRICA, #STAYATHOMEORDER, #STAYATHOMESA, #STAYATHOMESAVELIVES, #STAYATHOMESOUTHAFRICA, #STAYATHOMESTAYSAFE, #STAYATHOMESAFE, #COVID19INSA, #COVID19INSOUTHAFRICA, #21DAYLOCKDOWN, #21DAYLOCKDOWNSA were used to mine more data, and the process was repeated until a comprehensive collection of tweets was gathered. In all, 68,000 randomly filtered messages from twitter were harvested using the Twitter streaming API, and we applied the KNIME data analytics software as well as the Python programming software for data analysis and topic modelling tasks. The messages were filtered using the most common hashtags used during the coronavirus outbreak and included those related to the origins, spread and the responses to the pandemic.

Data pre-processing

The initial stage of data pre-processing was done to remove noise from our text data. In the context of Twitter data for natural language processing, we consider the following as a noise: Non-standards text and symbols, URLs & Whitespaces, Hashtags, Names of Users and Duplicate tweets. After the noise was removed, the following pre-processing steps were further conducted:

a) lo*wer casing*: All words were converted to lower case to reduce dimensionality and increase topic coherence.
b) *stop words removal*: Commonly used words that do not change the meaning of the data were also eliminated from the corpus. In essence, by eliminating such words from the text, the LDA algorithm could only take as input, the most important words hence producing more accurate results.
c) *text normalization*: Our mined data contained a lot of abbreviations, misspellings and use of out-of-vocabulary words (oov). In addition, the data had a lot of non-English words such as people's names and nicknames that were written in local African languages. Before normalization, we extracted a set of all such words and identified the ones that required no normalization. A rule-based method was then applied to normalize the remainder of the text (Miłkowski 2010).

Experimental setup

Using python code, the pre-processed tweets were converted into a corpus of text, and a word cloud was used to verify that there were no unwanted data within the corpus. The word cloud also served as a way of verifying the output from the pre-processing stage. The bag of words feature representation technique was used to represent the corpus. Consequently, the Latent Dirichlet Allocation (LDA) topic modelling technique was implemented on the bag of words created from a text corpus. The LDA classifies documents into topics by building a topic per document model and words per topic model, modelled as Dirichlet distributions (Kanan et al. 2015). Each pre-processed tweet is modelled as a multinomial distribution of topics, and each topic is modelled as a multinomial distribution of keywords. The assumption is that the bag of words used to feed the model contains words that are related. In our dataset, common hashtags relating to the COVID-19 pandemic were used to mine the tweet; hence this assumption is valid.

The LDA further assumes that the data used in analysis are made up of a mixture of topics, and the topics then generate words based on their probability distribution. To address limitations of this assumption, tweets were mined using different, but related hashtags, each addressing a specific aspect of trending issues on the COVID-19 pandemic. Initially, a total of 20 topics were extracted from the data, but further experiments were conducted to determine the optimal number of topics in the corpus, which was thereafter found to be 10.

Results and discussion

The word cloud that was obtained from the experiment conducted is shown in Figure 1. From the South African populace's perspective as harvested and processed from the tweets, the current study's findings indicate that most of the words in the word cloud are related to the conspiracy theories and fake news being circulated during the period (Kaur 2020; Schraer and Lawrie 2020). Some of these words include 'microchips', 'trust', 'Bill Gates', 'testing of the vaccines on Africans' and 'vaccines' could be linked to the 'wrong' perceptions by the populace.

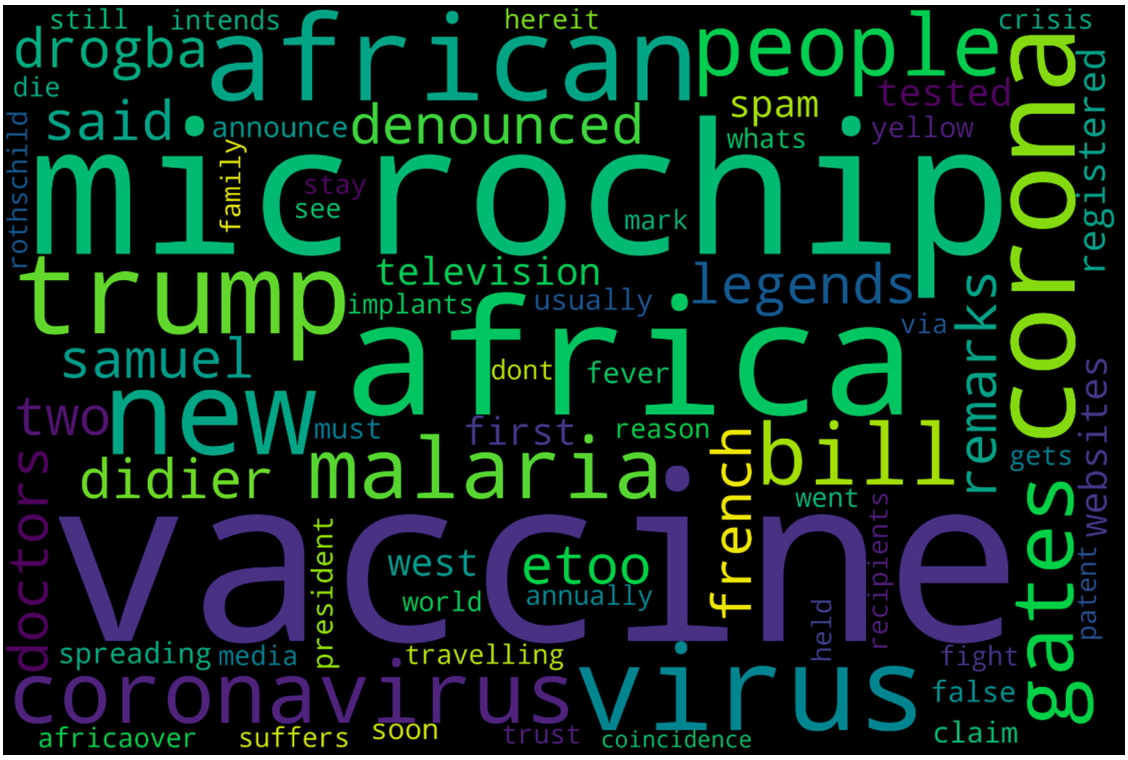

Figure 1: Uni-gram word cloud from COVID-19 related tweets in South Africa.

Such perceptions are capable of reducing the people's trust and cooperation in concerted efforts put in place by the South African government to tame the spread of the disease. The government must, therefore, intensify their awareness campaigns about the disease and interventions to correct these 'wrong' perceptions. In addition, as shown in Figure 1, some other words related to the symptoms (fever), the spread of the disease (stay, travelling, don't), getting tested (tested), monitoring information on COVID-19 (television) as well as the frontline health workers (doctors) are also being discussed among South Africans which could indicate their intention to protect themselves and stay safe.

The list of the extracted topics is shown in Table 1, and the most common words under each topic, including the frequency of occurrence, are shown in Figure 2. As shown in Figure 2, 'lockdown' is more popular than any other words and has a frequency of above 10,000. The initiative of the lockdown was introduced by the South Africa government on 26 March 2020, following similar measures in other nations across the world and supported by the World Health Organization (WHO 2020). The aim is to reduce human to human transmission of the disease by restricting public interactions, conduct more tests and screening while also isolating potential asymptomatic infections. Economic activities are suspended, offices, schools and markets closed while movements are strongly restricted among others. These would definitely have an adverse effect on the populace's livelihood as their psychology and mindset are seriously plagued. Hence, the issue of lockdown is linked to topic 0, which on the inter-topic distance plot (in Figure 4) shows that the topic is heavily discussed.

The media has been awashed with conspiracy theories on the COVID-19 pandemic. Consequently, unproven theories linking the revolutionary 5G mobile networks and the spread of COVID-19 are fuelling tensions and arson attacks on cell towers in some parts of the world, including Europe. These theories are reflected in words such as 'radiation', which is mentioned about 5600 times, as shown in Figure 2. This is linked to topic 1: the 5G and issue of conspiracy theories. Words such as 'virus', 'towers', 'don't', 'cause', etc. are the most common words suggesting that the perception of the populace is that the 5G networks might be the cause of the virus, and the erections of towers or communication masts should not be allowed.

Topic 3 is linked to alcohol and some of the most common words under this include 'SANDF (South Africa National Defence Force)', 'president', 'listen', news, etc. Further to the ban on the sale and consumption of alcohol during the lockdown period as announced by the president, the SANDF and police were deployed to enforce these directives among others. These deployments of security personnel have been met with mixed feelings by the populace, hence the high rate of discussion; as SANDF and police were discussed in more than 4000 tweets processed.

In addition, as seen in Table 1, topic 4 is linked to SANDF, and police violence as quite a number of high-handedness of security personnel were recorded including gender-based violence and abuse at the home fronts in some communities. Some of the common words under this topic include 'people', 'police', 'beat', 'lockdown', 'law', etc. and reflects the people's responses for the government to intervene appropriately.

Table 1: Topics extracted from tweets related to COVID-19 in South Africa.

Topic number	Description of extracted topics and words
Topic 0:	***Lockdown***
	Income, lockdown, work, food
Topic 1:	***5G Conspiracy theories***
	Radiation, people, coronavirus, virus, towers, think, don't, cause, corona, like
Topic 2:	***Staying home***
	Home, day, stay, lockdown, safe, days, time, today
Topic 3:	***Alcohol***
	SANDF, president, nothing, listen, alcohol, ramaphosa, news, remember, hands, die
Topic 4:	***SANDF and police violence towards blacks***
	SANDF, saps, people, police, beat, lockdown, law, black, distancing social
Topic 5:	***Tracing of daily statistics***
	Cases, workout, home, new, number, confirmed, time, COVID, today, total
Topic 6:	***SA president's address – Expectations***
	South, Africa, covid, covid_sa, coronavirus, president, africans, african, ramaphosa
Topic 7:	***Essential workers***
	Lockdown, food, people, essential, workers, time, health, need, government, services
Topic 8:	***Bill gates and conspiracy theories***
	Vaccine, people, virus, don't, gates, us, bill, want, like, test

We extended our experiments to obtain the most common tri-grams words or phrases tweeted and discussed by the populace as related to the disease. This is shown in Figure 3 and aimed at broadening the scope and meanings of the words and issues being discussed. Some of these words include:

- *President Cyril Ramaphosa (PCR*): This might be attributed to the fact that PCR has addressed the nation so many times (SONA 2020), and the citizens always look forward to his addresses of the situation report and what direction the government is heading. Hence more than 300 tweets mentioned PCR.
- *Case – South Africa*: This is a real-time update on the disease, it shows that people are actually concerned and tracking the relevant statistics including the rate of spread of the disease.
- *Vodafone boss blows & boss blows whistle*: These are related to the conspiracy theory that links 5G to COVID-19.
- *COVID-19 Vaccine Africa*: is also another conspiracy theory that is related to topic 1.

Topic coherence and perplexity

After applying the LDA algorithm to extract 10 topics relating to the COVID-19 pandemic as shown in Table 1, we evaluated the performance of the algorithm on our dataset using the coherence and perplexity metrics. We obtained a coherence score of 56% and a

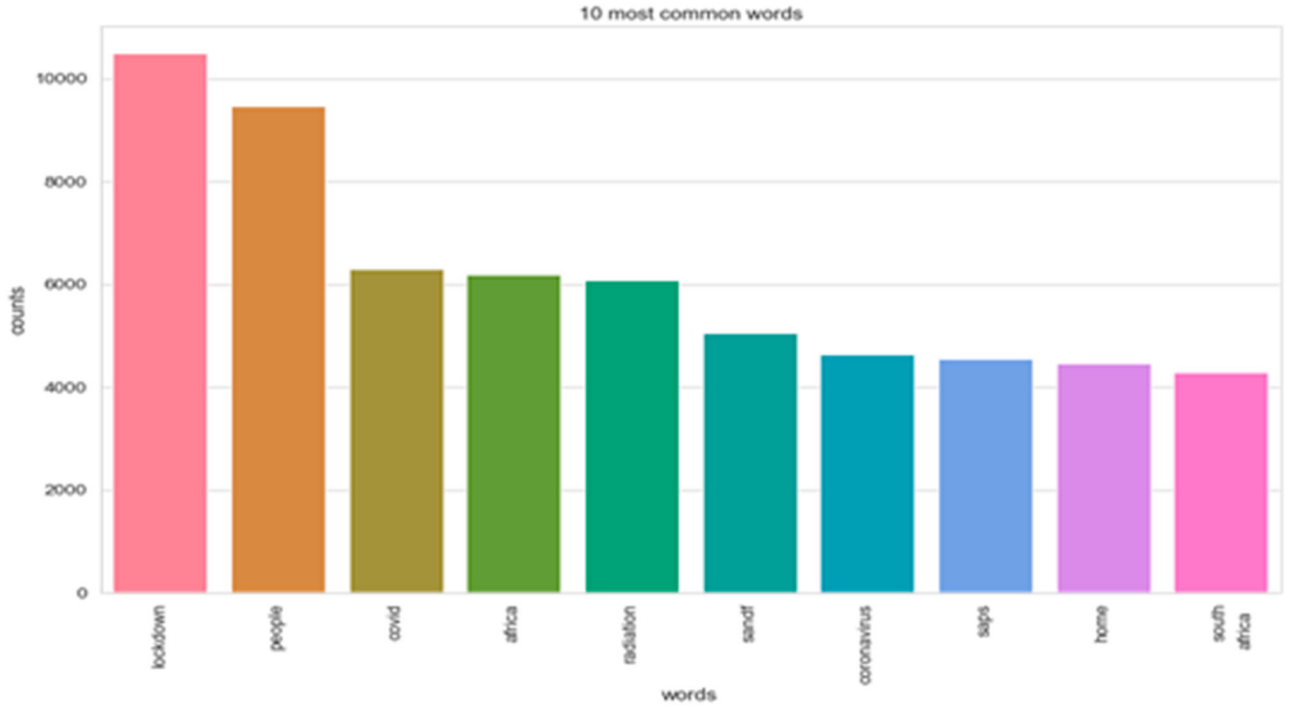

Figure 2: Histogram of the 10 most common uni-gram words from COVID-19 related tweets in South Africa.

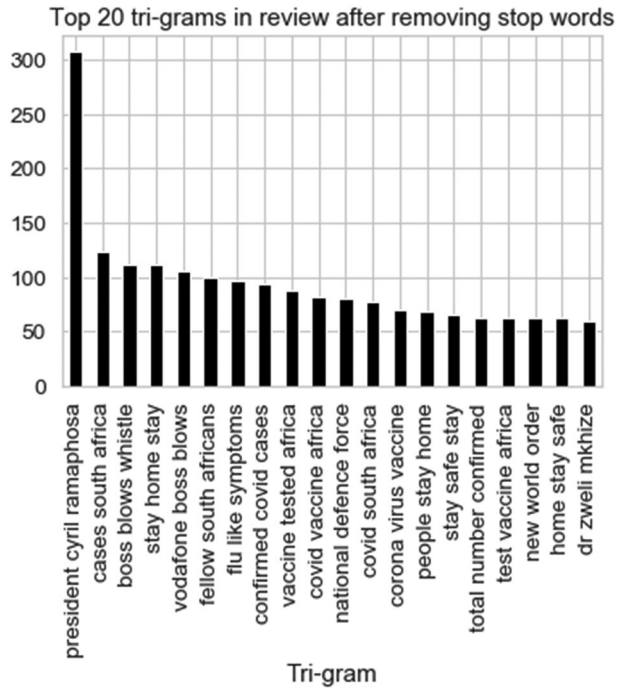

Figure 3: Histogram of the 20 most common tri-gram words from COVID-19 related tweets in South Africa.

perplexity score of −18. The coherence metric measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable and topics that are artefacts of statistical inference. At the same time, perplexity is the measure of how well a model predicts a sample and is measured as the normalized log-likelihood of a held-out test set. A lower perplexity score indicates a better generalization performance.

Intertopic distance plot

We also plotted an intertopic distance plot from the topics generated. It provides a global view of the topics and how they differ from each other, while at the same time allowing for a deep inspection of the terms most highly associated with each individual topic.

Each topic's overall prevalence is encoded using the areas of the circles. The right panel, as shown in Figure 4 depicts a horizontal bar chart whose bars represent the individual terms that are the most useful for interpreting the currently selected topic on the left.

We monitored the mainstream media including the online, print and electronic during the period of this current study in order to determine the burning issues relating to the COVID-19 pandemic that are being discussed, reported or broadcasted and our findings indicated that they align with the results of the topics and words generated by the LDA method that we applied on the COVID-19 tweets. As shown in Table 1, trending topics and words including lockdown, food, home, income, South African police service (saps), South African national defence force (SANDF), radiation, corona, safe, die, 5G, workout, vaccines, etc. were among the terms that the South African populace formed perceptions and attitudes around based on their knowledge and awareness of the COVID-19 pandemic.

Even if some of these perceptions are 'wrong', it could be argued that the populace are (in)advertently responding to the Abraham Maslow's hierarchy of needs motivation theory. A humanistic psychologist, Maslow (1948, 1970)

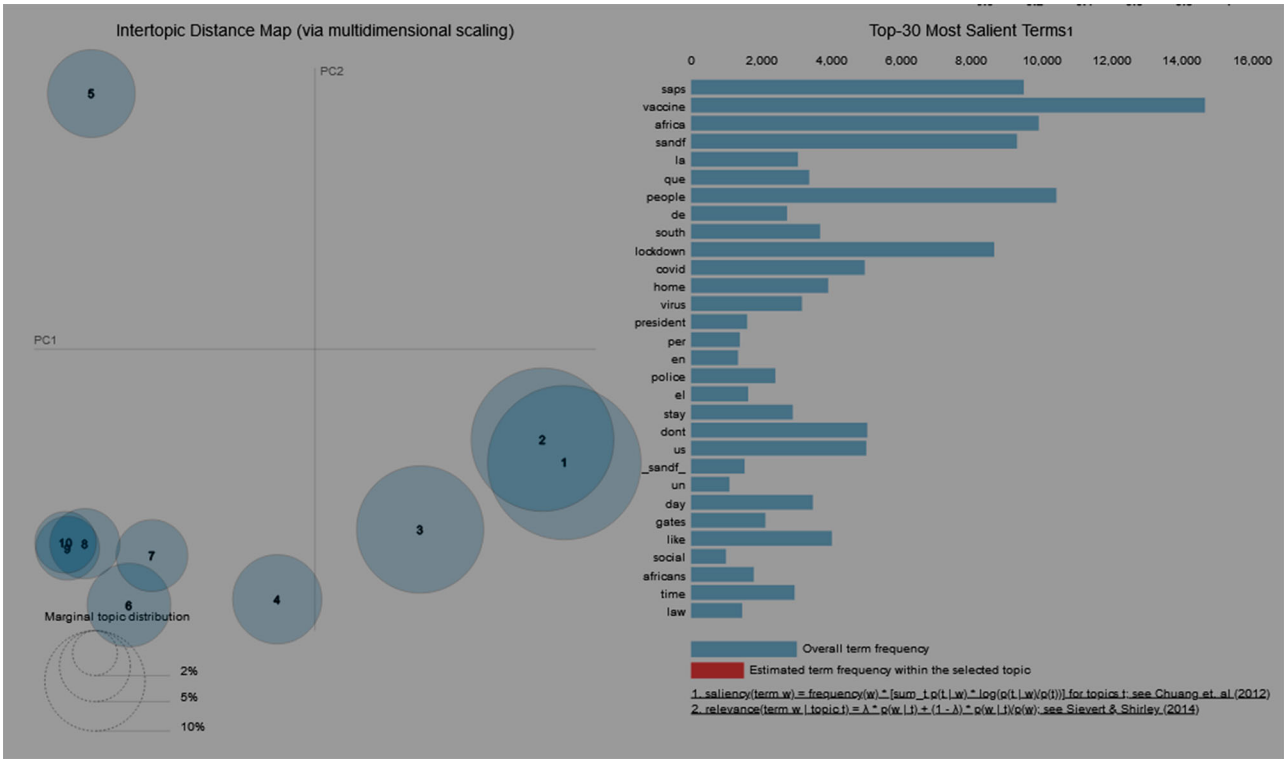

Figure 4: Inter-topic distance plot of the 10 topics from COVID-19 related tweets in South Africa.

opined that human needs are boundless and indeed have a particular hierarchical order hence no need or drive can be considered independently as they are related to one another. Every need is connected to the satisfaction or dissatisfaction of other needs and human beings are thus greatly motivated by unsatisfied needs. The national government's regulations on COVID-19 with a view to containing the spread, directly impacts on the populace's stages of how people progress in life which is termed hierarchical needs as postulated by Maslow (1948, 1970) and includes (i) physiological and basic needs consisting of food, water, shelter and warmth, etc. as the lockdown or movement restrictions affect sources of incomes and foods for a large group of people even as the government and other non-governmental agencies attempt to provide relieve materials such as foods, water, blanket and sundry items to some identified citizens. Evictions from residences were also halted to cushion the effects of the restrictions amidst a phased easing; (ii) safety needs – all needs to feel secure and free from dangers including security of health and property as well as physical safety are listed here, as people are very disturbed about their health status, the risk of getting infected as well as falling victims of security agencies' brutalities and domestic violence which escalated during the period. Hence 5G, vaccines, SAPS, SANDF trended as a result of these factors. Provision of nose masks, sanitizers, personal protective equipment, improved enlightenment campaigns and testing, sanctioning government officials and persuasions rather than brute force, etc. are some suggested ways that the government responded to some of the 'wrong' perceptions of the populace; (iii) social needs – consists of friendship, belongingness, bonding, association, mingling, etc. on which the populace are strictly regulated under the physical and 'social' distancing rules which was also extended to gathering and crowding restrictions; (iv) esteem needs – include approval, competence, respect and self-confidence, etc. The populace's self-confidence is deeply eroded here amidst the fear of contracting the COVID-19 virus; and (v) self-actualization needs – this include mental growth, pride and realizing one's potential, etc., which are severely impacted as psychological stress affects mental health and hinders actualizing full potential. Thus, some of the populace's resistance and perceptions are borne out of these human behavioural reactions towards meeting and satisfying their numerous needs.

Conclusion

In this paper, we used the LDA algorithm to extract topics from text mined from Tweeter. The study was primarily aimed at discovering and extracting the most topical issues relating to the COVID-19 pandemic that are being discussed by South Africans as well as the impacts of the extracted issues on the people's reactions to and compliance with the South African government's efforts in managing the disease. With 681,289 confirmed COVID-19 cases; 16,976 deaths; 614,781 recoveries and 4,269,626 tests conducted (as at 5 October 2020 – 12:23:00) in a population of about 60 million, the majority of whom still live in impoverished townships, government intervention programmes have been met with mixed reactions. This work, therefore, among other tasks extracted from Tweeter messages, how the different sectors of the population reacted to the pandemic and the various government interventions put in place. The LDA algorithm was able to extract coherent topics with a coherence score of 56% and a perplexity score of −18. Perplexity measures how well a probability model predicts a sample, and the lower the score, the better the model. This measure alone is not sufficient unless supported by looking at the topics themselves and check if they align with the realities. The most common tri-grams and topics extracted are coherent enough; hence the perplexity score of −18 is acceptable. Some of the study's findings are trending topics which include discussions on conspiracy theories linking 5G to the cause and spread of COVID-19 disease. Furthermore, the suspension of economic activities due to the lockdown was also a topical issue because of its adverse effects on the populace's livelihood. This is despite the government's response with some financial palliatives as well as a phased easing of restrictions as a result of the people's reactions. In essence, the intertopic distance plot shows that the issue of lockdown regulations and law enforcement agencies' brutalities are heavily discussed while the fake news and conspiracy theories of 5G and vaccines relating to COVID-19 also trended a lot. As lofty as the government regulations might be, these 'wrong' perceptions are capable of reducing the people's trust and cooperation in concerted efforts put in place by the government to tame the spread of the disease; consequently, an increase in the populace's trust is thus required. The implications of the results are relevant and should assist the government and policymakers in redirecting their efforts. This may include improved enlightenment campaigns and populace trust enhancement by addressing their wrong perceptions and reactions to the instituted measures toward an anticipated overall success. For instance, this could be noticed in the easing of some of the lockdown restrictions. However, as the COVID-19 pandemic continues to rage and new areas of damage and opportunities are being discovered, future work will need to include the extraction of sentiments and emotions from the harvested tweets. We are also interested in testing other topic extraction algorithms, including a combination of the natural language processing techniques and machine learning methods toward an automatic classification and prediction of diverse factors relating to the COVID-19 pandemic.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Anastassopoulou, C., L. Russo, A. Tsakris, and C. Siettos. 2020. "Data-Based Analysis, Modelling and Forecasting of the COVID-19 Outbreak." *PLoS ONE* 15 (3): e0230405. doi:10.1371/journal.pone.0230405.

Aramaki, E., S. Maskawa, and M. Morita. 2011. "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter." *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*,

Barzilay and Johnson (eds). Presented at: EMNLP 2011, Edinburgh, Scotland, 1568–1576. Association for Computational Linguistics.

Blair, S. J., Y. Bi, and M. D. Mulvenna. 2020. "Aggregated Topic Models for Increasing Social Media Topic Coherence." *Applied Intelligence* 50 (50): 138–156.

Bode, L., and K. E. Dalrymple. 2016. "Politics in 140 Characters or Less: Campaign Communication, Network Interaction, and Political Participation on Twitter." *Journal of Political Marketing* 15 (4): 311–332.

Burns N., Y. Bi, H. Wang, and T. Anderson. 2012. "Extended Twofold-LDA Model for Two Aspects in One Sentence." *In proceedings*: Greco S., Bouchon-Meunier B., Coletti G., Fedrizzi M., Matarazzo B., Yager R.R. (eds) *Advances in Computational Intelligence. IPMU 2012. Communications in Computer and InformationScience* 298: 265–275. Springer, Berlin, Heidelberg.

Burns, N., Y. Bi, H. Wang, and T. Anderson. 2011. "A two-Fold-LDA Model for Customer Review Analysis." *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology* 1: 253–256.

Canini, K. R., S. Lei, and L. Thomas. 2009. "Griffiths. Online Inference of Topics With Latent Dirichlet Allocation." *International Conference on Artificial Intelligence and Statistics*.

CDC (Centers for Disease Control and Prevention). 2020. *Coronavirus Disease 2019 (COVID-19) Symptoms*. Atlanta, GA: CDC. Accessed March 12, 2020. https://www.cdc.gov/coronavirus/2019-nCoV/index.html.

Chan, J. F., S. Yuan, K. H. Kok, K. K. To, H. Chu, J. Yang, F. Xing, et al. 2020. "A Familial Cluster of Pneumonia Associated With the 2019 Novel Coronavirus Indicating Person-To-Person Transmission: A Study of a Family Cluster." *The Lancet* 395 (10223): 514–523.

Chatterjee, S. 2019. "Drivers of Helpfulness of Online Hotel Reviews: A Sentiment and Emotion Mining Approach." *International Journal of Hospitality Management* 85: 102356.

Cole-Lewis, H., A. Varghese, A. Sanders, M. Schwarz, J. Pugatch, and E. Augustson. 2015. "Assessing Electronic Cigarette-Related Tweets for Sentiment and Content Using Supervised Machine Learning." *Journal of Medical Internet Research* 17 (8): e208.

Culotta, A. 2012. "Lightweight Methods to Estimate Influenza Rates and Alcohol Sales Volume from Twitter Messages." Language Resources and Evaluation, Special Issue on Analysis of Short Texts on the Web.

Goethals, B. 2005. "Frequent Set Mining." In *Data Mining and Knowledge Discovery Handbook*, edited by O. Maimon and L. Rokach, 377–397. Boston, MA: Springer. doi:10.1007/0-387-25465-X_17.

Hong, L., D. Ovidiu, and D. D. Brian. 2011. "Predicting Popular Messages in Twitter." In *Proceedings of the 20th International Conference Companion on World Wide Web*, Sadagopan et al.(eds.): 57–58. ACM. ISBN: 978-1-4503-0637-9.

Ibrahim, R., A. Elbagoury, M. S. Kamel, and F. Karray. 2018. "Tools and Approaches for Topic Detection from Twitter Streams: Survey." *Knowledge and Information Systems* 54 (3): 511–539.

Java, A., X. Song, T. Finin, and B. Tseng. 2007. "Why We Twitter: Understanding Microblogging Usage and Communities." In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, Zhang et al. (eds.): 56–65. San Jose, CA: ACM.

JHU, COVID-19 Resource Center. 2020. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University & Medicine, Baltimore, MD. Accessed August 15, 2020.

Kagashe, I., Z. Yan, and I. Suheryani. 2017. "Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data." *Journal of Medical Internet Research* 19 (9): e315.

Kanan, T., S. Ayoub, E. Saif, G. Kanaan, P. Chandrasekarar, and E. A. Fox. 2015. *Extracting Named Entities Using Named Entity Recognizer and Generating Topics Using Latent Dirichlet Allocation Algorithm for Arabic News Articles*. Department of Computer Science, Virginia Polytechnic Institute & State University, Blacksburg, Virginia.

Kaur, H. 2020. "The Conspiracy Linking 5G to Coronavirus Just Will Not Die." Accessed August 15, 2020. https://edition.cnn.com/2020/04/08/tech/5g-coronavirus-conspiracy-theory-trnd/index.html.

Lamb, A., M. J. Paul, and M. Dredze. 2013. "Separating Fact from Fear: Tracking Flu Infections on Twitter." *Conference of the North American Chapter of the Association for Computational Linguistics*.

Levitt, J. 2020. "South Africa Claims a Spot in Top Ten for Global Covid-19 Infections Over Two Days." *Times Live*, June 15, 13:03. Accessed August 15, 2020. https://www.timeslive.co.za/news/south-africa/2020-06-15-sa-claims-a-spot-in-top-ten-for-global-covid-19-infections-over-two-days/.

Maskeri, G., S. Santonu, and H. Kenneth. 2008. "Mining Business Topics in Source Code Using Latent Dirichlet Allocation." In *Proceedings of the 1st India Software Engineering Conference*, Gautam, Jalote, and Rajamani (eds), 113–120. New York, NY.

Maslow, A. H. 1948. ""Higher" and "Lower" Needs." *Journal of Psychology: Interdisciplinary and Applied* 25 (2): 433–436.

Maslow, A. H. 1970. *Motivation and Personality*. 2nd ed. New York: Harper & Row.

Miłkowski, M. 2010. "Developing an Open-Source, Rule-Based Proofreading Tool." *Software: Practice and Experience* 40 (7): 543–566.

Oppegaard, B., and M. K. Rabby. 2016. "Proximity: Revealing New Mobile Meanings of a Traditional News Concept." *Digital Journalism* 4 (5): 621–638.

Ostrowski, D. A. 2015. "Using Latent Dirichlet Allocation for Topic Modelling in Twitter." *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*.

Petrović, S., M. Osborne, and V. Lavrenko. 2010. "Streaming First Story Detection With Application to Twitter." In *HLT: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Kaplan M. R. et al. (eds), 181–189. Stroudsburg, PA: Association for Computational Linguistics.

Rider, A. K., and N. V. Chawla. 2013. "An Ensemble Topic Model for Sharing Healthcare Data and Predicting Disease Risk." *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, Wu and Hannenhalli (eds), 2013: 333-340. Washington, DC, USA. ISBN 978-1-4503-2434-2.

Robertson, C., and L. Yee. 2016. "Avian Influenza Risk Surveillance in North America With Online Media." *PLoS One* 11 (11): e0165688.

Schraer, R., and E. Lawrie. 2020. "Coronavirus: Scientists Brand 5G Claims' Complete Rubbish." Accessed August 15, 2020. https://www.bbc.com/news/52168096.

Shen, Z., P. Luo, S. Yang, and X. Shen. 2010. "Topic Modeling Ensembles." *Proceedings of 2010 IEEE International Conference on Data Mining*, Geoffrey et al. (eds.), 1: 1031–1036. Sydney, Australia. ISBN: 978-0-7695-4256-0.

SONA. 2020. www.stateofthenation.gov.za/message-from-the-president.

Tian, K., R. Meghan, and P. Denys. 2009. "Using Latent Dirichlet Allocation for Automatic Categorization of Software." Mining Software Repositories. MSR'09. 6th IEEE International Working Conference on. IEEE.

Tomasi, D. 2020. *Coronavirus Disease (COVID-19). A Socio-Epidemiological Analysis*. Bennington, VT: Vermont Academy of Arts and Sciences LV(I).

Tsai, C. F., K. Chen, Y. H. Hu, and W. K. Chen. 2020. "Improving Text Summarization of Online Hotel Reviews With Review Helpfulness and Sentiment." *Tourism Management* 80: 104122.

Omnicore. 2020. "Twitter by the Numbers: Stats, Demographics & Fun Facts." https://www.omnicoreagency.com/twitter-statistics/.

Wang, Y., A. Eugene, and B. Michele. 2012. "Tm-Ida: Efficient Online Modeling of Latent Topic Transitions in Social Media." In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Qiang, Deepak and Jian (eds.): 123–131, August 12-16, 2012, Beijing, China. ACM.

WHO (World Health Organization). 2020. "Coronavirus Disease (COVID-2019) Situation Reports." Accessed August 15, 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

Winarko, E., and R. Pulungan. 2019. "Trending Topics Detection of Indonesian Tweets Using BN-Grams and Doc-p." *Journal of King Saud University – Computer and Information Sciences* 31 (2): 266–274.

Worldometers. 2020. "COVID-19 Coronavirus Pandemic: Reported Cases and Deaths by Country, Territory, or Conveyance." Accessed August 15, 2020. https://www.worldometers.info/coronavirus/#countries.

Wu, H., T. Xin, T. Ran, Z. Weidong, Y. Gongjun, and A. Vasudeva. 2016. "Application of Social Media Analytics: A Case of Analyzing Online Hotel Reviews." *Online Information Review*. doi:10.1108/OIR-07-2016-0201.

Xu, X., N. Yuruk, Z. Feng, and T. A. J. Schweiger. 2007. "SCAN: A Structural Clustering Algorithm for Networks." *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12–15, 2007*, Berkhin, Caruana and Xindong (eds.): 824–833, ISBN 978-1-59593-609-7

Yin, Z., D. Fabbri, S. T. Rosenbloom, and B. A. Malin. 2015. "Scalable Framework to Detect Personal Health Mentions on Twitter." *Journal of Medical Internet Research* 17 (6): e138.

Zendah, J. H., and A. Y. Maghari. 2019. "Detecting Significant Events in Arabic Microblogs Using Soft Frequent Pattern Mining." *Journal of Engineering Research and Technology* 6(1), 11–19.