Topic Modelling for Characterizing COVID-19 Misinformation on Twitter: A South African Case Study

Irene Francesca Strydom(✉) and Jacomine Grobler

Department of Industrial Engineering, Stellenbosch University, Stellenbosch, Western Cape, South Africa
strydomfrancesca@gmail.com, jacomine.grobler@gmail.com
https://ie.sun.ac.za/

Abstract. The COVID-19 pandemic has recently shed light on the potential for social media as a means of spreading mis-, dis-, and mal-information. This paper investigates embedding and cluster-based topic modelling to characterise the COVID-19 infodemic on South African Twitter, which has largely remained unstudied during the COVID-19 pandemic. The best performing model is able to identify specific misinformation narratives, but these narratives are mostly found within more general topics. A more fine-grained model is trained, and is able to much better isolate rumour/misinformation topics from more general topics. Finally, the paper makes several suggestions for dealing with the multilingual and code-switched nature of South African Twitter, as well as for the exploration and development of new dynamic topic modeling approaches that could be especially valuable for tracing the development of specific misinformation or rumour narratives over time. The paper presents novel insights and results on the application of a combination of data mining, machine learning and optimisation for addressing the pressing issue of misleading information on social media.

Keywords: Misinformation · Transformers · Social media · Twitter · Topic modelling · Natural language processing

1 Introduction

The COVID-19 pandemic has brought to light the potential of social media as a platform for the widespread and rapid dissemination of misinformation. In South Africa, misinformation has also made an appearance. Examples from early 2020 include rumours that 5G radio signals cause COVID-19 and rumours about COVID-19 relief funds provided to citizens by the South African government. Nearly 3 years later, in January 2023, some of these rumours are still circulated in one form or another.

Unlike many social media narratives in countries such as the United States and the United Kingdom, the South African social media narratives about

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023, LNCS 13957, pp. 289–304, 2023.
https://doi.org/10.1007/978-3-031-36808-0_19

COVID-19 have remained largely unstudied, especially regarding their development over time. In January 2020, internet penetration in South Africa was at 62% and the number of social media users was approximately 37% of the total population. By the start of 2022, South Africa's internet penetration rate had increased to 68.2% of the total population, and the number of social media users to 46.4% of the total population [11]. The number of social media users in South Africa is likely to continually increase as more South Africans gain access to the internet, making it all the more important to understand the dynamics of information dissemination on these platforms, since they are likely to play important roles in information dissemination and public sensemaking during future crises and emergencies. Without knowing what role these platforms played at the height of the COVID-19 pandemic, it is impossible to plan effectively or mitigate potential risks related to social media misinformation in future health crises, or any large scale crisis for that matter.

Machine learning (ML) and natural language processing (NLP) are useful tools to study the large amounts of social media being generated at a rapid pace. Research efforts devoted to analysing social media narratives about COVID-19, including misinformation narratives, have successfully utilized supervised ML techniques for classifying misinformation, while others have opted for unsupervised learning and data mining techniques to study the characteristics related to misinformation. Since supervised methods require labelled training data, and such training data are not yet available for the South African context, supervised methods are not yet applicable. Unsupervised learning techniques, such as topic modelling, are therefore considered instead.

This paper uses the BERTopic topic modelling algorithm [7] to extract salient themes from South African Twitter data. The study evaluates the use of this unsupervised topic modelling approach specifically for characterising mis- /disinformation on South African Twitter. This study is the first in-depth assessment of the performance of topic modelling on South African COVID-19 Twitter data. Consequently, the analysis provides new insight to the relevant misinformation topics discussed on South African Twitter during COVID-19. Additionally, topics unrelated to misinformation provide insight to the perceptions and experiences of South Africans during the pandemic. The paper also highlights specific challenges faced when working with multilingual code-switched data for which there are limited NLP resources available.

The rest of the paper is structured as follows: Sect. 2 discusses past research in the relevant fields of study, Sect. 3 describes the methodology applied in the study, and the results are discussed in Sect. 4. The study is concluded and recommendations for future work are made in Sect. 5.

2 Related Work

Following the outbreak of COVID-19 in 2020 and the multiple resulting global crises, there has been an explosion of research on methods for addressing misinformation on social media. ML and NLP have proven to be especially popular

techniques to apply to social media data at scale. Many studies follow a supervised learning approach to address the problem, in which the aim is to train a supervised classification model to detect misinformation among general information. Supervised methods naturally require labelled data for training, and social media misinformation datasets (specifically related to COVID-19) have been curated over the past three years precisely for this purpose [4,9,14,16]. Studies by [1] and Kaliyar *et al.* [10], for example, used labelled Twitter datasets to train models, and achieve F1-scores of more than 95% on their test data.

In the South African context, however, there are no labelled COVID-19 misinformation datasets, and so researchers must rely on publicly available misinformation datasets from other countries for training data. Strydom and Grobler [19] collected labelled data from four publicly available Twitter datasets and trained a neural network misinformation classifier. Although their best performing model attained good results on their test data (an F1-score of 89.9%), when applied to South African Tweets, their classifier did not perform as well. The reasons cited for this are that the differences in datasets relating to events, entities involved, and use of language are too substantial.

In addition to the aforementioned difficulties in finding appropriate, suitable training datasets, the often complex and nuanced nature of misinformation makes using static, labelled datasets problematic in any real-life setting. In many datasets what is labelled as misinformation is not necessarily false, it is simply 'unverifiable' or not backed by any evidence at that particular moment in time. Consequently, situations could arise where information that is questionable at one moment in time (and thus labelled as misinformation) may be verified at the next moment in time, immediately making certain labels outdated. The recently reported events related to claims about COVID-19 vaccines and their efficacy are one example of such a situation. Another issue with labelled datasets is that they do not necessarily reflect the official policy of local governments for what constitutes instances of mis-/dis-/malinformation. In South Africa, Real411, is an online platform for reporting digital misinformation. Developed by Media Monitoring Africa, the platform relies on technical and legal experts, alongside journalists to judge whether reported content constitutes disinformation, hate speech, incitement, or harassment. In this manner Real411 ensures that its assessments are in line with the South African constitution.

An unsupervised approach that identifies topics that are potentially *controversial/contentious* or *risk topics*, and that leaves the final evaluation of veracity to human experts in the field, is therefore a much more attractive option for dealing with misinformation in a holistic and sustainable manner. Topic modelling is an unsupervised text mining technique that extracts salient themes or topics from large corpora of text and has been applied in a variety of COVID-19 social media studies. In their South African study, [15] apply the latent Dirichlet allocation (LDA) topic modelling algorithm to a corpus of COVID-19 Tweets and identify 9 broad topics, two of which are related to misinformation narratives. These studies are not focused on misinformation, however, and so the opportunity to study misinformation on South African social media in more depth

remains. Furthermore, LDA was proposed nearly 20 years ago, in 2003; the state of the art for NLP has advanced drastically since then and many alternative, advanced topic modelling techniques have been developed accordingly.

Word embedding models have played an especially important role in advancing NLP. These models are pre-trained on large corpora of text and convert words into numerical representations (embeddings) that capture the semantic meaning of the words. Earlier models simply return the same embedding for a word regardless of its context, but more recent contextualised embedding models (such as the famous BERT model [5]) are now able to adjust a word's embeddings based on the context it is used in, thus more accurately capturing its semantic meaning.

BERTopic [7], a novel topic modelling technique, makes use of a combination of pre-trained embeddings, dimensionality reduction, clustering and term importance scores for words to generate useful and accurate topic representations. The technique is modular and therefore extremely flexible and simple to implement. Fig. 1 shows the standard configuration of the BERTopic algorithm, which utilizes sentence transformers [18] (for creating embeddings), uniform manifold approximation and projection [13] (UMAP, for dimensionality reduction), hierarchical density-based spatial clustering of applications with noise [2,3] (HDBSCAN, for clustering), CountVectorizer [17] (for creating bag-of-words representations) and class-based term frequency inverse document frequency [7] (c-TF-IDF, for creating topic representations). BERTopic's modular nature allows for easy experimentation and supports a number of variations of topic modelling.

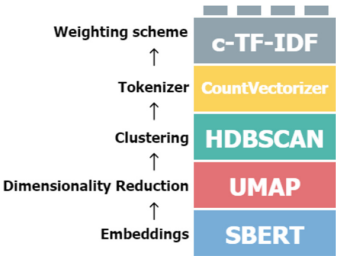

Fig. 1. The default configuration of the BERTopic topic modelling algorithm, as illustrated by [8]

3 Method

This section describes the method followed to apply the BERTopic topic modelling technique to a corpus of South African Tweets about COVID-19. Section 3.1 describes the data collection and pre-processing steps followed to collect the

relevant Tweets. Section 3.2 describes the modelling process in terms of each component of the topic model, including hyperparameter optimization. Section 3.4 presents additional experiments with regards to fine-grained topic models conducted based on observations made from the results in Sect. 3.2.

3.1 Data Collection and Pre-processing

The Twitter dataset contains 143 228 Tweets about COVID-19 that were posted in South Africa from 1 March 2020 to 30 April 2020. The Tweets were collected from the Twitter API using a combination of search parameters. The collected Tweets (1) contain either one of the keywords "covid", "coronavirus", "covid-19", "pandemic", or the Tweet context annotation "COVID-19", and (2) are geo-tagged with the "ZA" place country tag. The "ZA" tag ensures that only Tweets posted from locations within South Africa, and presumably Tweets by South Africans, are returned. Not all Tweets necessarily have location data, but the "ZA" place country tag is the only way to filter Tweets by location, since sensitive user profile information (such as country of residence) is kept private. The corpus of Tweets was pre-processed before being passed to the BERTopic model. Duplicate Tweets, user mentions, hyperlinks, emojis, and Tweets with two or less words were all removed. Tweets referencing videos or images were also removed, since they either contain no text at all, or contain text that references the videos or images. The cleaned text is then passed to the "all-MiniLML6-v2" pre-trained sentence transformer, which converts each Tweet's text to a 384-dimensional embedding.

3.2 Modeling

The embedded Tweets are then passed to the BERTopic model. First UMAP reduces the dimensionality of each embedding from 384 to five, and these reduced embeddings are then passed to the clustering algorithm, HDBSCAN. The topic representations are then calculated from the clusters of documents output by HDBSCAN.

The BERTopic algorithm, as described in Sect. 2, has a number of parameters that influence its performance and the manner in which topics are formed and represented. A brief description of each parameter and the value selected for it follows. *top_n_words* refers to the number of terms that are extracted to represent each topic. This value was set to ten, as is best practice in topic modelling. The higher this number, the less coherent topic representations tend to become. *n_components* refers to the number of dimensions that UMAP must reduce the text embeddings to. This value is set to five. *n_gram_range* is the range that the number of words which can count as a single term can fall within. Extending the range beyond one allows for phrases like "Cape Town" or "President Ramaphosa" to be represented as a single term in the corpus, while also preserving the individual words that make up the phrases. The wider the range, the more computationally expensive modelling becomes. This value is set to a minimum of one and a maximum of two. *umap_metric* refers to the distance

metric used to calculate the similarity between embeddings. The cosine distance metric was selected, since it is known to perform well with high dimensional data, such as the 384-dimensional embeddings output by the sentence transformer model. *hdbscan_metric* refers to the distance metric used to calculate similarity between the reduced embeddings output by UMAP, i.e. the distance metric to be used for clustering. The Euclidean distance metric was selected, since it is computationally efficient when working with lower dimensional data.

In addition to the fixed parameter values, the values of two additional parameters, *nn* and *mts*, were determined by evaluating the performance of the models for different combinations of values for these two parameters. The first of these, *nn*, refers to the number of neighbouring points that UMAP takes into account when approximating the manifold. The smaller this number, the more local structure is preserved; the larger the number, the more global structure is preserved. The second parameter, *mts*, simply refers to the minimum number of data points that may constitute a cluster.

Coherence is most often the metric of choice when evaluating and comparing the performance of topic models. The BERTopic modelling process is stochastic, and, therefore, produces a slightly different output every time it is run, even if all parameters are fixed. Therefore, a set of 30 models was trained for each combination of *nn* and *mts* values, and the average performance score for each set calculated. The metric used to assess the performance and topic representation of trained topic models is the UMass coherence score. Table 1 shows the different combinations of the two parameters tested, while Fig. 2 shows the average UMass coherence score obtained for each combination of values.

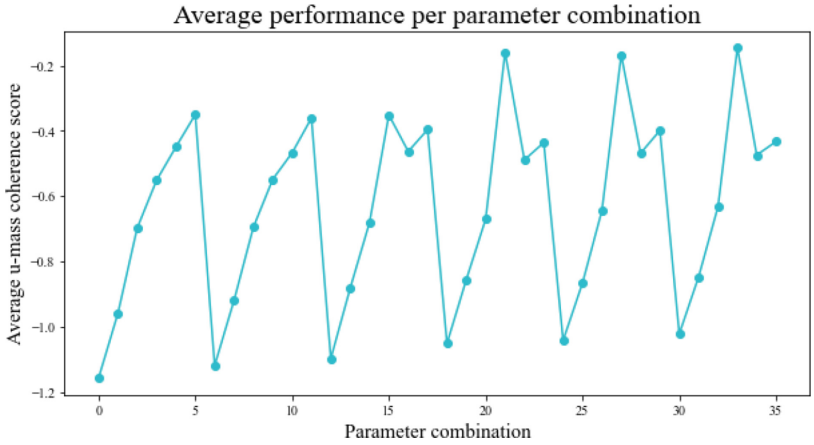

Fig. 2. The average UMass coherence score obtained for different combinations of *n_neigbours* and *mts*

To determine which combinations of values produced *statistically significant* increases in performance over other values, the Mann Whitney U-test, a non-parametric test for independent samples, was used. Each set of models was compared to every other set (with the Mann Whitney U-test) and the number of *wins*, *losses* and *draws* were recorded for each sample. A win is recorded if a sample distribution is statistically different from and has a higher mean coherence than another sample, a loss is recorded if a sample is statistically different from and has a lower mean coherence than another sample, and a draw is recorded if there is no statistically significant difference between the samples. The results of each set of Mann Whitney U-tests are shown in Table 1. The results for combinations 22, 28, and 34 are ignored, since these values produce models with only three topics that do not present any meaningful insights. Combination 6, with 31 wins, 3 losses, and 1 draw, thus provides the best performance. Consequently, the final model was trained with the values of 50 for *nn* and 500 for *mts*, as described in Sect. 3.3.

Table 1. The number of wins, draws and losses obtained for each combination of *nn* and *mts*

Combination (nn, mts)	Wins	Draws	Losses	Combination (nn, mts)	Wins	Draws	Losses
1 (50, 50)	0	0	35	**19 (300, 50)**	3	1	31
2 (50, 100)	6	0	29	**20 (300, 100)**	9	2	24
3 (50, 200)	12	1	22	**21 (300, 200)**	15	2	18
4 (50, 300)	18	1	16	**22 (300, 300)**	33	2	0
5 (50, 400)	23	4	8	**23 (300, 400)**	20	5	10
6 (50, 500)	31	1	3	**24 (300, 500)**	26	2	7
7 (100, 50)	1	0	34	**25 (400, 50)**	3	1	31
8 (100, 100)	7	0	28	**26 (400, 100)**	9	2	24
9 (100, 200)	12	2	21	**27 (400, 200)**	14	3	18
10 (100, 300)	18	1	16	**28 (400, 300)**	33	1	1
11 (100, 400)	21	5	9	**29 (400, 400)**	20	5	10
12 (100, 500)	30	1	4	**30 (400, 500)**	28	2	5
13 (200, 50)	2	0	33	**31 (500, 50)**	5	0	30
14 (200, 100)	8	0	27	**32 (500, 100)**	9	2	24
15 (200, 200)	13	3	19	**33 (500, 200)**	14	3	18
16 (200, 300)	20	12	3	**34 (500, 300)**	34	1	0
17 (200, 400)	21	5	9	**35 (500, 400)**	20	3	12
18 (200, 500)	28	2	5	**36 (500, 500)**	25	3	7

3.3 Final Model Training

The final model was trained with the best performing parameters identified previously, and produced 34 topics and attained a UMass coherence score of -0.39. In addition to the 34 topics generated, an additional group of outlier documents (labelled as -1) was produced. These outliers are not assigned to any topic and correspond to 'junk topics' often generated by other kinds of

topic models. There are two approaches for addressing the outliers generated by the model, which are: (1) to ignore the outliers in the analysis of results (this approach is recommended by the authors of BERTopic and HDBSCAN); (2) to pass outliers to the model and assign them to the most similar clusters based on a set threshold. In this study, the first approach was followed and the outliers were not considered in the analysis of the model's output.

3.4 Additional Experiments

An additional experiment was performed in addition to the one described in Sect. 3.2. A fine-grained model with a larger number of topics (and thus smaller topics) was trained to test if it would better separate and isolate potential misinformation topics from general topics. The model was trained with a mts of 100 and nn of 50 and produced 157 topics, attaining a UMass coherence score of -1.04.

4 Discussion

The topics produced by the final model (described in Sect. 3.3) were inspected and labeled, as shown in Tables 2 and 3. Topics were labeled according to their top ten representative terms, as well as sample Tweets belonging to each topic. Out of the 34 identified topics, three could be directly linked to COVID-19 misinformation. These are topics 15, 20, and 23, whose word clouds are shown in Fig. 3. A number of COVID-19 rumours were identified by inspecting each of these topics' Tweets, hashtags, mentions, entities, and other related metadata.

The three identified misinformation topics are 15 (China), 20 (News and leadership), and 23 (Ethnicity).

The first official cases of COVID-19 were identified in China and a number of theories regarding the origin of the virus are also associated with China. As seen in Table 4, one theory is that the virus was manufactured and is a form of biological warfare. Donald Trump famously called COVID-19 the 'Chinese virus', and this is a term that evidently has spread to South African social media as well, evidenced by the hashtags found in the topic. Interestingly, a link between 5G mobile technology, China, and COVID-19 is also found in this topic, as illustrated by Tweets 1 and 2 in Table 4. A sense of mistrust towards China and the World Health Organization's (WHO) reported number of COVID-19 infections and deaths is also reflected in the Tweets. Additionally, calls for boycotting Chinese products and businesses are also present.

Although topic 20 does address fake news, it is (in most cases) difficult to tell what the fake news is about from only the Tweet text, since many Tweets are comments on URL links contained within the Tweets, which were removed when training the model. Inspecting the URL links contained in these texts does, however, trace the comments to other Tweets or websites that contain the alleged fake news. Examples of these original Tweets are shown in Table 5.

One rumour identified in topic 23 is that Africans are immune to COVID-19, as shown by example Tweets in Table 6.

Table 2. The number of Tweets and top ten terms for topics 0–17

Topic	Label	Count	Top ten terms
0	Working from home	11201	home, working home, working, work, work home, data, stay, stay home, know, time
1	Lockdown 1	8819	lockdown, lockdown, ke, le, day, end, extended, ka, day lockdown, going
2	Business responses	6280	cases, coronavirus, confirmed, covid_19sa, deaths, covid19southafrica, covid19insa, fund, world, relief
3	Quarantine	4898	quarantine, quarantined, self, self quarantine, quarantine quarantine, day, day quarantine, chill, quarantine chill, stream
4	South Africa	3950	south, africa, south africa, cape, africans, south africans, western, african, cases, lockdown
5	Corona	3780	corona, corona corona, time, le, know, time corona, going, ke, want, le corona
6	Coronavirus	2635	virus, corona virus, corona, coronavirus, coronavirussa, coronavirusinsa, infected, coronavirusoutbreak, immune, cure
7	Entertainment	2491	lockdown, music, party, lockdownhouseparty, togetherathome, dj, house party, house, lockdown house
8	Sanitary measures	2198	hands, wash, masks, wash hands, mask, sanitizer, water, hand, hand sanitizer, gloves
9	Food & groceries	2161	food, lockdown, chicken, cook, bread, eat, eating, buy, cooking, weight
10	Addresses	2012	lockdownsa, cyrilramaphosa, lockdownsouthafrica, covid19southafrica, coronavirusinsa, covid_19sa, lockdown, covid19sa, covid19insa, shutdownsouthafrica
11	Alcohol ban	1657	alcohol, liquor, wine, lockdown, drink, drinking, alcohol lockdown, bottle, ban, stores
12	Religion	1656	god, church, pray, jesus, lord, prayer, churches, amen, prayers, pastors
13	Global situation	1550	pandemic, global pandemic, pandemic pandemic, global, world, time, country, going, need, middle
14	Education	1492	school, schools, students, teachers, education, kids, grade, learners, learning, parents
15	China	1446	china, chinese, chinese virus, virus, wuhan, coronavirus, world, trump, corona, korea
16	Social distancing	1275	distancing, social distancing, social, distancing social, practice, practice social, practicing social, practicing, isolation, socialdistancing
17	COVID-19 testing	1207	tested, test, testing, positive, tested positive, tests, kits, results, symptoms, screening

The analysis of topics and related Tweets reveals both the strengths and weaknesses of the unsupervised topic modelling approach. It is evident from the analysis that information disorder is present on South African Twitter. The topics output by the model make sense within the context of the pandemic and correspond with key issues faced by South Africans as identified by other studies, such as the cigarette and alcohol bans [6,12]. In particular, topics identified

Table 3. The number of Tweets and top ten terms for topics 18–33

Topic	Label	Count	Top ten terms
18	Sports	962	league, football, liverpool, sports, cancelled, players, coronavirus, psl, cancel, season
19	Economic repercussions	953	lockdown, pay, money, paid, economy, work, salaries, rent, leave, income
20	News and leadership	879	news, fake news, fake, leadership, president, mr president, spreading fake, leader, mr, good news
21	Italy	877	italy, spain, italian, italians, deaths, coronavirus, france, usa, 10, friends
22	President Ramaphosa	861	ramaphosa, cyril ramaphosa, president, cyril, president cyril, president ramaphosa, address, nation, coronavirussa, address nation
23	Ethnicity	830	black, white, blacks, race, whites, racist, virus, black person, corona, racism
24	Exercise	816	gym, dogs, dog, workout, exercise, walk, lockdown, jogging, day, walking
25	Lockdown 2	780	21, days, 21 days, lockdown, days lockdown, extended, lockdown 21, weeks, lockdown extended, months
26	Law enforcement & crime	742	police, arrested, arrest, lockdown, crime, criminals, law, jail, soldiers, regulations
27	Twitter	655	tweet, twitter, tweets, social media, lockdown, media, tweeting, tl, retweet, social
28	Cigarette ban	621	cigarettes, smoking, cigarette, smoke, smokers, ban, tobacco, lungs, sale, alcohol
29	Birthdays	586	birthday, happy birthday, happy, lockdown birthday, lockdown, birthday lockdown, lockdown happy, hope, day, enjoy
30	Personal care	512	hair, lockdown, nails, barber, haircut, beard, makeup, hair lockdown, skin
31	Relationships	509	lockdown, relationships, relationship, sex, married, partner, love, dating, lockdown lockdown
32	Lockdown 3	506	cyrilramaphosa, cyril, lockdown, bayithemovie, cyrilramaphosa lockdown, extended, bayithemovie bayithemovie, lockdown cyrilramaphosa
33	COVID-19 statistics	505	numbers, 100, number, stats, million, brand, 1000, increasing, gathering, scary

also correspond with specific topics identified by topic models applied in other studies [15]. The BERTopic model also identified topics that contain examples of information disorder, that were not identified in the studies mentioned earlier, namely topics 15 (China), 20 (News and leadership), and 23 (Ethnicity).

The nature of topic modelling, however, makes it difficult to distinguish within a topic between information disorder and general information, opinions, and questions on the level of individual Tweets. That is, the granularity of the model in Sect. 3.4 means that Tweets in a topic, if it indeed includes information disorder, will likely be a mixture of both information disorder Tweets and general



(a) Word cloud for topic 15 (b) Word cloud for topic 20



(c) Word cloud for topic 23

Fig. 3. Word clouds for misinformation topics

Table 4. Examples of Tweets in topic 15

	Tweet
1	Remember those trade bans USA had on China last year? The war for 5G could be linked to Covid19
2	@eugenegu @realDonaldTrump This coronavirus is a hoax just to give China a boost!!-Just look at their stock market, the last 7 https://t.co/6zOtUmP8l9 not forget 5G...
3	You may hate on @realDonaldTrump... and say He is racist, But the truth of the matter is This #COVID-19 is being called wrong it should Be called The #ChineseVirus.
4	I wonder if the UN is going to take action against China fir their creation of the COVID-19. The WHO must also institute charges on China. #UnitedNations #StopCovid19 #ChinaMustPay

Tweets. These results are conceivably still more useful than a classifier that only distinguishes between information disorder Tweets and general Tweets, such as those trained in [19], since information disorder is better described and positioned in relation to other topics and types of information. In this respect, the model proves useful in analysis and determining the overall consensus towards a topic and an appropriate response (or non-response) to the dialogue/information present on Twitter.

To a certain extent, the complexity of information disorder is reflected in this occurrence. A more fine-grained approach, i.e. a topic model with more

Table 5. Examples of Tweets in topic 20

	Tweet
1	Isn't this #FakeNews in violation of the regulations? Hello @SAPoliceService https://t.co/2fDNwF8LBN
2	How come some people want to feed us with #FakeNews ? https://t.co/qdDNvONTrB
3	FAKE NEWS ALERT | WhatsApp voicenote 'was not from me' - Groote Schuur virologist How do we know anything on News24 is not fake news? Who do you believe? https://t.co/1Y4taH5HA2? Shared via @Health24com
4	News24 always with the sensational journalism https://t.co/MuJcUuaCuj

Table 6. Examples of Tweets in topic 23

	Tweet
1	And know that they will not show themselves when they are on the receiving end of things, CoronaVirus is a neumonia like virus attacking the lungs, But it can't do so kuBlack Man! #MelaninOurProtectorAgainstCovid19 That's why they say we have 65 cases, and not say what race is it
2	News is our black skin is resistant to Corona Virus, meaning we as black people are immune to this disease. I have never heard of black person having corona. I hope it's true
3	Is there any black person infected with Corona Virus yet ? Because I'm starting to think this virus is for white people only like they said back then HIV is for black people.
4	Idrise Elba is lying. White people begged him to say this nonsense so that it seem that blacks suffer as much as whites from corona. Blacks don't suffer much from corona.

(and smaller) topics that better subdivides the larger topics identified in Sect. 4, could perform better. These smaller topics would be more representative of the specific narratives and language patterns that are often hallmarks/characteristic of information disorder and could possibly allow for identifying more uniform and 'pure' information topics. The coherence score utilised earlier may not be optimal in such an instance, but what is optimal depends on the context of the problem. Here the results suggest that the context requires another way of measuring performance. The coherence measure may be useful for more traditional uses of topic modelling (i.e. summarising and categorising texts) but may not be ideal for fine-grained analysis if misinformation is anomalous (some niche rumours may be different than mainstream widely spread rumours).

To explore this assumption, the results obtained from the fine-grained topic model with a *mts* of 100 (as described in Sect. 3.4) were inspected. The fine-grained model produced 157 topics, the largest comprising 4 507 Tweets, and the smallest comprising 100 Tweets. Potential mis-/disinformation topics identified from the fine-grained model output are shown in Table 7. It is important to note

that the parameters of this model were not optimised and only serve to illustrate the potential benefit of utilising a more fine-grained topic model specifically for characterising information disorder. The model does particularly well in splitting the larger topics from the model in Sect. 3.3 into smaller coherent themes, and evidently also identifies smaller niche rumours, as expected. Topic 97, for example, contains Tweets that discuss the similarities between films about pandemics and the coronavirus pandemic, and many view the 2011 thriller film, *Contagion*, as a prediction of the COVID-19 pandemic. Topic 114 is specifically focused on rumours about 5G causing COVID-19, and topic 121 is specifically focused on the discussion about the virus's effects on different population groups. Conspiracies related to Bill Gates and his alleged involvement in vaccine development are represented in topic 122. Other conspiracies related to vaccines are found in topic 117.

Table 7. The number of Tweets and top ten terms for each topic

Topic	Count	Top ten terms
9	1463	china, chinese, chinese virus, virus, wuhan, coronavirus, world, trump, korea, virus china
12	1351	corona virus, virus, corona, coronavirus, virus corona, coronavirus corona, virus coronavirus, cure, symptoms, like
26	606	virus, infected, immune, infect, people, virus virus, spreading, spread, infection, spread virus
62	297	fake news, fake, news, spreading fake, spreading, media, news24, arrested, news fake, stop spreading
81	246	whatsapp, whatsapp group, group, whatsapp groups, messages, groups, whatsapp number, 060, 3456, 012 3456
84	235	cuba, cuban, doctors, cuban doctors, cubans, cubandoctors, vaccine, fidel, medical, castro
86	220	flu, covid, covid 19, 19, symptoms, cure, immune, cure covid, disease, flu covid
88	212	vaccine, vaccines, lupus, 19 vaccine, trials, vaccine vaccine, covid, covid 19, 19, patients
89	209	led, ignorance, shame, stupidity, stupid, idiot, ignorance bliss, bliss, dumb, ignorant
97	194	movie, contagion, movie contagion, contagion movie, 2011, watch, sabc, watching, watch movie, watched
114	158	5g, 5g towers, radiation, towers, conspiracy, 5g corona, technology, 5g coronavirus, 5g technology, causes
117	152	ebola, vaccine, vaccines, french, africa, french doctors, tested, africans, test, tested africa
121	147	black, black people, white, white people, blacks, black person, virus, people, corona virus, corona
122	146	gates, vaccine, potential coronavirus, coronavirus vaccine, confident potential, gates confident, vaccines, work africa, africa, vaccine work
134	127	lies, lie, lying, facts, lies lies, lies detected, lied, detected, liar, spreading lies

5 Conclusion

In this study, a topic model, BERTopic, was applied to a corpus of unlabelled South African Tweets about COVID-19. The topics produced by the final model were able to reveal information disorder on South African Tweets. Three topics could be directly linked to information disorder related to origin theories linking China to the creation of the virus (topic 15), the effect of the virus on specific population groups (topic 23), and general discussions about COVID-19 fake news (topic 20). An additional fine-grained model was also trained (without hyperparameter optimization) to test if it would better isolate misinformation topics.

There is still much potential for expanding and refining the topic modelling approach to characterizing misinformation on South African social media. As discussed in Sect. 4, coherence may not be the optimal metric to evaluate the performance of a topic model for characterizing misinformation. Other metrics, such as topic diversity and topic significance, will be utilized in future work.

The Tweets identified as belonging to potential misinformation topics may be used as a starting point for curating labelled COVID-19 misinformation datasets specifically for South African Twitter, and could aid in future research studies that wish to use these datasets for supervised learning. As noted in Sect. 2, the changing and complex nature of health related information must be kept in mind when curating these kinds of datasets.

Additionally, a classification model trained to classify Tweets according to the topics identified by the topic modelling approach may be trained, as well as a stance classifier which can classify whether a Tweet supports, denies, or is neutral towards a specific issue or topic.

An important area for future research is that of dynamic topic modelling (DTM), which models the evolution of topics over time. Some rumours and misinformation seem to spread rapidly, others build up more slowly, and some only surface at specific intervals in time. The temporal aspect of misinformation dissemination has largely been ignored in NLP approaches and therefore needs further attention.

The pre-trained sentence transformer model used to create embeddings was trained on English data only. While the topic model is able to identify salient topics related to misinformation, multilingual Tweets containing primarily non-English words may likely have been labelled as outliers. A more in-depth analysis of the outliers is thus necessary to confirm this supposition. Future work therefore aims to address the multilingual nature of the data by investigating methods for working with low-resource languages, such as isiXhosa, in the South African data.

Acknowledgements. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant number: 129340).

References

1. Abdelminaam, D.S., Ismail, F.H., Taha, M., Taha, A., Houssein, E.H., Nabil, A.: CoAID-DEEP: an optimized intelligent framework for automated detecting COVID-19 misleading information on twitter. IEEE ACCESS **9** (2021). https://doi.org/10.1109/ACCESS.2021.3058066
2. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) PAKDD 2013. LNCS (LNAI), vol. 7819, pp. 160–172. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37456-2_14
3. Campello, R.J., Moulavi, D., Zimek, A., Sander, J.: Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans. Knowl. Discovery Data (TKDD) **10**(1), 1–51 (2015)
4. Cui, L., Lee, D.: CoAID: COVID-19 healthcare misinformation dataset. Comput. Res. Repository (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 4171–4186 (2019)
6. Filby, S., van der Zee, K., van Walbeek, C.: The temporary ban on tobacco sales in south Africa: lessons for endgame strategies. Tob. Control (2021)
7. Grootendorst, M.: BERTopic: neural topic modeling with a class-based TF-IDF procedure. arXiv preprint: arXiv:2203.05794 (2022)
8. Grootendorst, M.: BERTopic algorithm (2023). https://maartengr.github.io/BERTopic/algorithm/algorithm.html. Accessed 20 Oct 2021
9. Hayawi, K., Shahriar, S., Serhani, M., Taleb, I., Mathew, S.: ANTi-Vax: a novel twitter dataset for COVID-19 vaccine misinformation detection. Publ. Health **203**, 23–30 (2022)
10. Kaliyar, R.K., Goswami, A., Narang, P.: A hybrid model for effective fake news detection with a novel COVID-19 dataset. In: ICAART (2), pp. 1066–1072 (2021)
11. Kemp, S.: Digital 2022: South Africa (2022). https://datareportal.com/reports/digital-2022-south-africa. Accessed 30 Aug 2022
12. Matzopoulos, R., Walls, H., Cook, S., London, L.: South Africa's COVID-19 alcohol sales ban: the potential for better policy-making. Int. J. Health Policy Manag. **9**(11), 486 (2020)
13. McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction. arXiv preprint: arXiv:1802.03426 (2018)
14. Memon, S.A., Carley, K.M.: Characterizing COVID-19 misinformation communities using a novel twitter dataset. Comput. Res. Repository (2020)
15. Mutanga, M.B., Abayomi, A.: Tweeting on COVID-19 pandemic in south Africa: LDA-based topic modelling approach. Afr. J. Sci. Technol. Innov. Dev. **12**, 1–10 (2020)
16. Patwa, P., et al.: Fighting an infodemic: COVID-19 fake news dataset. In: Chakraborty, T., Shu, K., Bernard, H.R., Liu, H., Akhtar, M.S. (eds.) CONSTRAINT 2021. CCIS, vol. 1402, pp. 21–29. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73696-5_3
17. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

18. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. Association for Computational Linguistics, Hong Kong, November 2019. https://doi.org/10.18653/v1/D19-1410, https://aclanthology.org/D19-1410
19. Strydom, I.F., Grobler, J.: Transformers for COVID-19 misinformation detection on twitter: a south African case study. In: Giuseppe, N., et al. (eds.) Machine Learning, Optimization, and Data Science: 7th International Conference (LOD 2022), pp. 197–210. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-25599-1_15