

Article

An Infodemiology and Infoveillance Study on COVID-19: Analysis of Twitter and Google Trends

Reem Alshahrani ¹ and Amal Babour ^{2,*} ¹ Department of Computer Science, Taif University, Taif 26571, Saudi Arabia; rashahrani@tu.edu.sa² Department of Information Systems, King Abdulaziz University, Jeddah 21589, Saudi Arabia

* Correspondence: ababor@kau.edu.sa

Abstract: Infodemiology uses web-based data to inform public health policymakers. This study aimed to examine the diffusion of Arabic language discussions and analyze the nature of Internet search behaviors related to the global COVID-19 pandemic through two platforms (Twitter and Google Trends) in Saudi Arabia. A set of Twitter Arabic data related to COVID-19 was collected and analyzed. Using Google Trends, internet search behaviors related to the pandemic were explored. Health and risk perceptions and information related to the adoption of COVID-19 infodemic markers were investigated. Moreover, Google mobility data was used to assess the relationship between different community activities and the pandemic transmission rate. The same data was used to investigate how changes in mobility could predict new COVID-19 cases. The results show that the top COVID-19-related terms for misinformation on Twitter were folk remedies from low quality sources. The number of COVID-19 cases in different Saudi provinces has a strong negative correlation with COVID-19 search queries on Google Trends (Pearson $r = -0.63$) and a statistical significance ($p < 0.05$). The reduction of mobility is highly correlated with a decreased number of total cases in Saudi Arabia. Finally, the total cases are the most significant predictor of the new COVID-19 cases.



Citation: Alshahrani, R.; Babour, A. An Infodemiology and Infoveillance Study on COVID-19: Analysis of Twitter and Google Trends. *Sustainability* **2021**, *13*, 8528. <https://doi.org/10.3390/su13158528>

Academic Editor: Andrei P. Kirilenko

Received: 10 June 2021
Accepted: 27 July 2021
Published: 30 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: coronavirus; COVID-19; Google Trends; infodemiology; infoveillance; social media; Twitter

1. Introduction

In late December 2019, news of a novel coronavirus began to emerge from Wuhan, China. The new virus, named by World Health Organization (WHO) as Coronavirus Disease 2019 (COVID-19) causes severe acute respiratory infection leading to death in many cases [1]. This virus has continued to spread aggressively around the world, causing a global panic and posing serious challenges to public health and policy makers. Many countries have enforced rigorous protective measures to combat the pandemic, such as mandatory quarantine and massive closures. However, efforts to slow down the transmission rate have been undermined by the COVID-19 'infodemic' [2–5].

An important public health response to the COVID-19 infodemic is proactive communication, with the objectives of alleviating confusion, avoiding misunderstandings, minimizing adverse consequences, and saving lives [6]. In March 2020, WHO published a recommendation to provide guidance for countries on how to implement effective Risk Communication and Community Engagement (RCCE) strategies to improve public awareness and perception [7]. RCCE can help prevent infodemics by building trust in public health institutions and spokespersons, thereby increasing the chance that public health guidelines will be followed. Determining what people know, how they feel, and what they do to keep the outbreak under control is essential for public health workers to address the public's perception of risk and uncertainties. Thus, designing effective RCCE strategies to stop further spreads of misinformation and rumors must be based on public health surveillance. Social media and web searches have become essential sources of information about the public and are widely utilized for health-related information. The COVID-19

pandemic provides a starting point to discuss how social media and web searches can be leveraged to design beneficial RCCE to address this unprecedented crisis.

In the past decade, the internet has become an integral part of people's lives. Online sources that provide real time or near-real time data are becoming increasingly available, consequently changing the pattern of spread of health-related information [8]. This paradigm shift can be useful for understanding population health concerns and needs, the diffusion of health-related information/misinformation, and the public's reaction to health events [9]. There is a large body of studies utilizing social media and internet traffic to understand the prevalence and diffusion of information and misinformation about COVID-19, providing insights for public health surveillance [6] and health policy makers [10,11].

This use of the internet has formed two new concepts: infodemiology, defined as the science of distribution and determinants of information on the Internet, and infoveillance, defined as the longitudinal tracking of infodemiology metrics for surveillance and trend analysis [9]. Infodemiology and infoveillance have contributed to public health knowledge by analyzing a range of topics such as chronic diseases and influenza. More specifically, infodemiology provides the necessary tools for understanding health-related infodemics, which can make it hard for people to find reliable sources and trustworthy guidance during public health crises [12,13].

Twitter Analytics and Google Trends are two of the most effective infoveillance tools for assessing the dissemination of information on health issues and topics [14,15]. Twitter functions as a convenient source of information about COVID-19 [11]. Google Trends is a tool that provides both real-time and archived information on Google queries worldwide. Thus, information from internet searches that are performed anonymously, enabling the analyzing and forecasting of health-related topics can be obtained. The social media search index has been identified as a promising predictor of COVID-19 transmission rates [16,17]. Although Twitter has some limitations as a tool for disease prediction and containment, its potential for communicating peoples' stories and news sharing may profoundly impact public health outcomes. Both Google Trends and Twitter can serve as viable resources to understand people's perception and to monitor their reaction to the pandemic over time [14,18].

The use and roles of social media and web searches amid the COVID-19 pandemic have been widely studied [14,16,17]. However, to the best of our knowledge, no systematic research has yet been conducted on Arabic infodemiology and infoveillance. Given the fact that 41% of the online population in Saudi Arabia uses social media, a higher percentage than any other country in the world, it is vital to conduct a study concerning the Arabic language [19]. Considering the impact that the spread of information and misinformation of COVID-19 may have upon the transmission rate, determining the public reactions to tweets and investigating the nature and diffusion of COVID-19-related information on the internet can provide important insights into the beliefs and concerns of Arabic users. Furthermore, analyzing the spread of information found in tweets may help decision makers to intervene in limiting the wide spread of misinformation or fake news, such as setting laws to prevent the spread of unreliable information.

In this paper, the infodemiology of COVID-19 in terms of socially disseminated information are addressed. First, the magnitude of misinformation and the quality of information sources regarding the COVID-19 epidemic that is being spread on Twitter in the Arabic language was analyzed. Second, information prevalence indicators (search volume) about COVID-19 was analyzed by using data from Google Trends to examine the correlation between information prevalence and daily new cases in different provinces in Saudi Arabia. Third, the relationship between mobility activity in Saudi Arabia and COVID-19 prevalence was examined [20]. Fourth, an infodemiology study with a multiple regression analysis was conducted to examine the relationship between new COVID-19 cases and various potential predictors, namely overall mobility, total confirmed cases, and information prevalence. In this paper, Saudi Arabia was chosen as a case study since it is one of the top Arabic countries that is highly affected by the pandemic [21]. Moreover,

to the best of our knowledge, no study has examined infodemiology and infovalliance in Saudi Arabia.

Figure 1 shows the status of the pandemic in Saudi Arabia according to the regions. The cumulative number of cases in Saudi Arabia reached 275,000 as of 3 June 2020, with confirmed cases in each Saudi province [22]. This study serves as a starting point for designing strategic messages for health campaigns and establishing an effective risk communication channel.

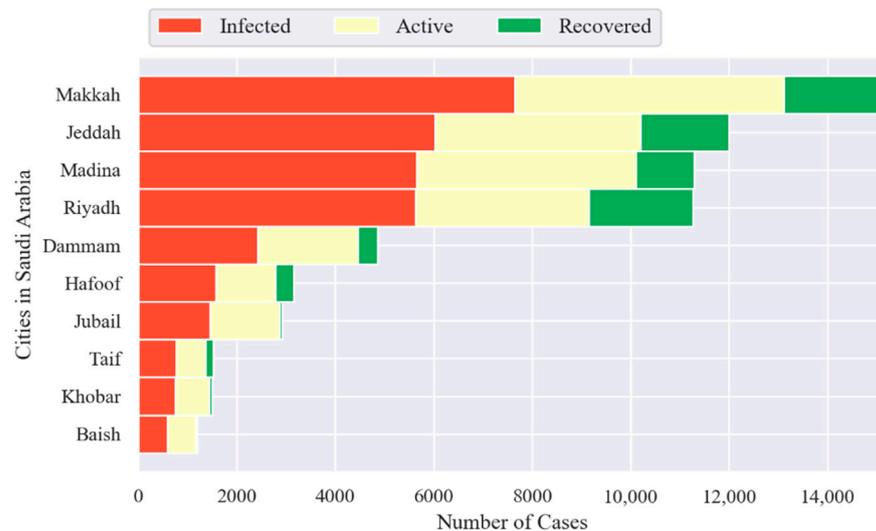


Figure 1. The status of COVID-19 cases in Saudi Arabia's cities as of 3 June 2020.

2. Materials and Methods

To answer the research questions of this study, data from Twitter [23], and Google Trends [24] and Google COVID-19 community mobility data [20] were collected. In this section, the data collection process and analysis are explained.

2.1. Twitter

Searches about the novel coronavirus related tweets written in the Arabic language and geolocated in Saudi Arabia were performed between 13 June 2020 and 12 July 2020 using Python software 3.7.0, Twitter standard search Application Programming Interface (API), and Tweepy Python libraries. In the search, a set consisting of predefined search terms that were most widely used as media terms for the novel coronavirus remedy from Twitter trends were used. Table 1 shows the English name for each of the terms used in the search with a brief description.

Table 1. List of terms used in collecting tweets.

Term	English Name	Description
كورونا	Coronavirus	The most common term for the novel coronavirus
ديكساميثازون	Dexamethasone	Corticosteroid used as anti-inflammatory
القسط الهندي	Saussurea costus	Common folk remedy for respiratory complaints
سماق	Sumac	Common folk remedy for a variety of complaints
رمديسيفير	Remdesivir	An antiviral medication

A set of 6541 Arabic tweets was collected. The data contains text and metadata of the tweets, including tweet id, username, hashtags, and number of retweets. The data preprocessing involves three phases which are the data cleaning phase, normalization phase, and lemmatization phase. In the data cleaning phase, all non-Arabic terms, stop

words, numbers, punctuations, emojis, hashtags, and URLs were automatically removed by coding. In the normalization phase, multiple forms of a letter were converted into one uniform letter, numbers, spaces, repeated letters, and elongation were removed using the Tashaphyne Python library [25]. In the lemmatization phase, words were converted to their roots using the Farasa toolkit [26]. An example of the preprocessing phases is shown in Figure 2.

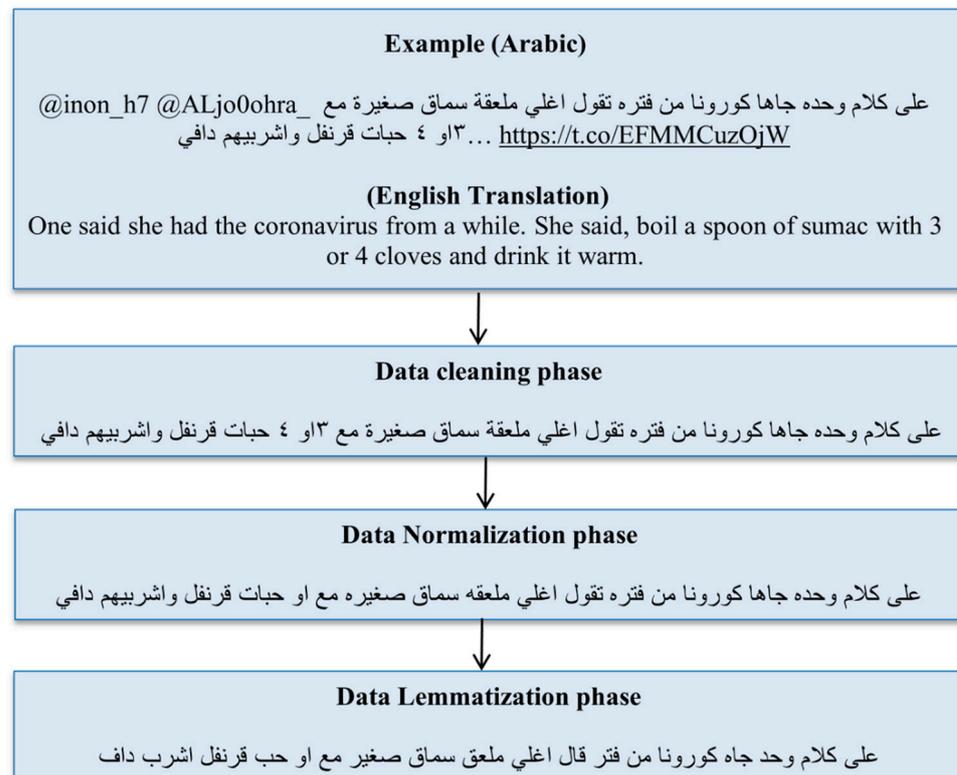


Figure 2. Example of preprocessing phases.

A set of metrics were selected to analyze the collected tweets. Information prevalence was computed by counting the number of conversations mentioning 'كورونا' ('coronavirus') in combination with one or more of the other terms listed in Table 1 within the collected tweets. The information occurrence ratio was determined by calculating the information prevalence for each of the terms divided by the total number of the collected tweets. The information prevalent quality was found by examining the source of each tweet in the collected tweets to determine its reliability. The accounts of Saudi Arabia's Ministry of Health, official TV channels, TV news channels, TV programs, official newspapers, and online newspapers were considered as high-quality sources (HQS); usernames of personal accounts, personal groups, and unofficial accounts were coded as low-quality sources (LQS). Then, the information prevalent quality was calculated as the number of the tweets that were retweeted from LQs divided by the total number of the tweets that were retweeted from both HQSs and LQs. Information incidence was calculated by determining the number of conversations about each of the defined four terms combined with 'كورونا' ('Coronavirus') in the collected tweets by units of time, where the unit of time in this study is four weeks. To visualize the most frequent words in the conversations, word clouds was used via RStudio Version 1.3.1056 [9,15].

2.2. Google Trends Data

Google Trends is an open online tracking tool that provides real-time and archived information for internet hit search volumes [19]. It normalizes and scales data in the form

of search volume numbers to reflect search popularity on a scale relative to the total number of queries carried out on Google overtime. The scale ranges from 0 (low) to 100 (highly popular) for a specific search term. The information prevalence indicator [19] in this study is represented by the popularity scale provided by Google Trends.

In order to find the information prevalence in different provinces, the framework in [14] was followed. The framework provides a systematic way to extract information from Google Trends. In this study, the country was set to “Saudi Arabia”, and a default of “All categories” and “Web search” were selected. Since “Coronavirus or كورونا” is the official name used by the Saudi Ministry of Health, the topic “كورونا” (“Coronavirus”) was selected as the most widely used term for the novel coronavirus. Besides this, other popular search terms/phrases were added. The terms/phrases and the English name for each of them are provided in Table 2. Note that only Arabic words were examined; no English words were included in this study.

Table 2. English translation for the most Arabic terms used to describe COVID-19.

Arabic Term/Phrase	English Name
كورونا	Coronavirus
كوفيد ١٩	COVID-19
كوفيد	COVID
كرونا	Corona
عدد حالات كورونا	Coronavirus new cases
كورونا في السعودية	Coronavirus in Saudi Arabia

Google Trends data were collected as time series queries during the period from February to July 2020 and were retrieved in the csv format. Since there is a lag between the action of Google search and new cases confirmation, a delayed effect of 14 days was considered. The collected data were aligned with official data on daily COVID-19 new cases and deaths per one million people for each province in Saudi Arabia during the period from May to July with a lagged difference of 14 days, all of which was retrieved from the Saudi Ministry of Health [27]. The data was then arranged as (date, province, demand prevalence indicator, number of new cases). The aim is to examine the correlation between Google search queries for different Arabic cities and the daily increase in COVID-19 cases. Pearson correlation coefficients were used to examine the association between daily new COVID-19 cases and the hit search volume in each province using the same data. Autocorrelation function (ACF) (the correlation of a parameter with itself over the time) was also performed using Google Trends data to test whether the number of cases of a specific day would impact the search volume the next day [28]. Both crude and partial autocorrelations were computed. All of the calculations were performed using a Python 3.7.0 environment. Note that for all analyses, an alpha level of 0.05 was used to determine statistical significance.

2.3. Google Mobility Data

Daily community mobility data provided by Google covers 130 countries starting from 15 February 2020 [20]. Saudi Arabia was picked as a case study to examine whether or not less mobility is associated with fewer COVID-19 total cases per one million. The data are grouped by Google into five categories based on the most popular activities among people. these categories are work, grocery and pharmacy, parks, residential, and retail. To describe mobility, Google uses a percentage change in relation to previous values or baseline [20], as seen in Figure 3. The baseline is calculated by Google as the average value, for the same day of the week, during the five-week period between 3 January and 6 February 2020. For example, on 21 February 2020 in Saudi Arabia compared to the baseline, grocery mobility decreased by 43%, park mobility decreased by 60%, retail and transit mobilities decreased

by 100%, and residential mobility decreased by 100%. The government imposed a curfew to control the spread of the pandemic. During the curfew, most retail stores were closed and residential activities were prohibited. However, there were specific hours during the day when people were allowed to go to grocery and pharmacy or parks. Noteworthy, a positive score in mobility indicates increased mobility and a negative score indicates a reduction in mobility. Google collects only the data from the devices whose users allow their location to be used anonymously.

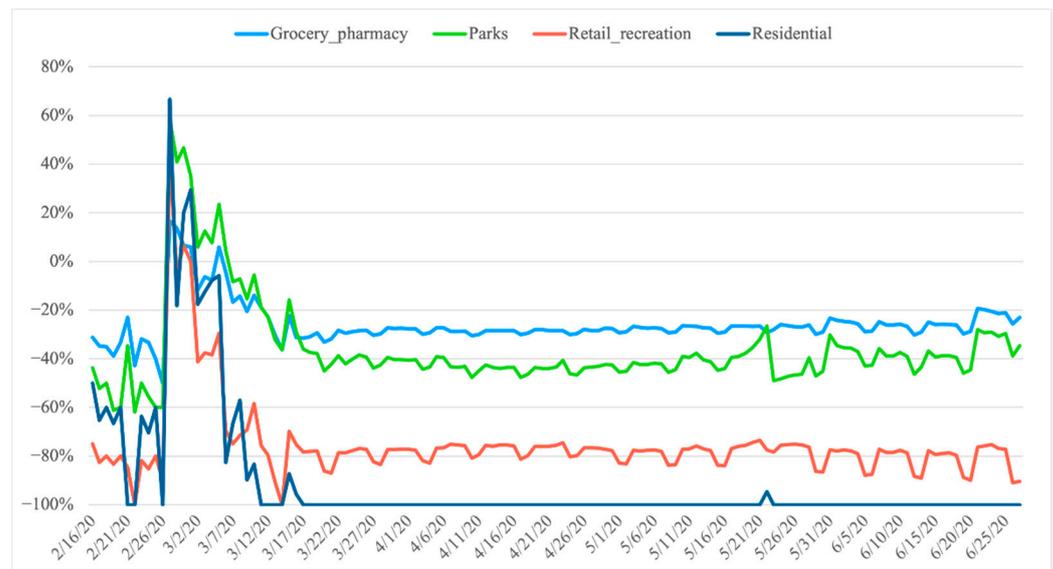


Figure 3. Mobility of Google users in Saudi Arabia.

To capture daily trends in movement patterns, the data used in this analysis covers the period from 16 February to 25 June 2020. Since the Saudi Arabia government ordered all workers to work from home throughout this period [16], the “work” category was excluded and mobility was assessed based on the four remaining categories to find any association between the new COVID-19 cases and people’s mobility. These associations were investigated using Pearson correlations (if both variables entered into the correlation were normally distributed) or non-parametric Spearman rank correlations (if one/both of the variables entered into the correlation were not normally distributed). The Pearson Correlation Coefficient was applied to specify how two variables vary together (the new COVID-19 cases and people’s mobility).

Overall mobility (i.e., the average mobility across all categories) from the same data set were used to conduct a multiple regression analysis to investigate how mobility in general could predict new coronavirus cases. This analysis also included total cases and information prevalence as predictors of new cases as well. The data was examined for the entire country and was not disaggregated by province. To conduct this analysis, Rstudio Version 1.3.1056 was used. Using the same data, a multiple regression analysis was carried out to investigate whether patterns of travel to grocery and pharmacy, parks, residential, and retail areas could significantly predict new coronavirus cases (i.e., the four mobility categories were used as predictors in this regression, as well as total cases and information prevalence).

3. Results

3.1. Twitter

During the study period, the total number of tweets about the most popular media terms for the novel coronavirus remedy was 6541. Figure 4 shows the differences among the information prevalence, the information occurrence ratio, and the quality of the information prevalence for each of the identified terms. As seen, the majority of tweets with an

information prevalence of 3807 and an information occurrence ratio of 58.2% was about 'ديكساميثازون' ('Dexamethasone'), 89% of which were retweeted. This was followed by 1694 (26%) tweets about the folk remedy 'القسط الهندي' ('Saussurea costus'), 64% of which were retweeted. The use of 'رمديسيڤير' ('Remdesivir') comes third with 1021 (15.6%) tweets, 90% of which were retweeted. The use of the folk remedy 'سماق' ('Sumac') is considered the least prevalent with 19 tweets (0.3%) and 100% retweets.

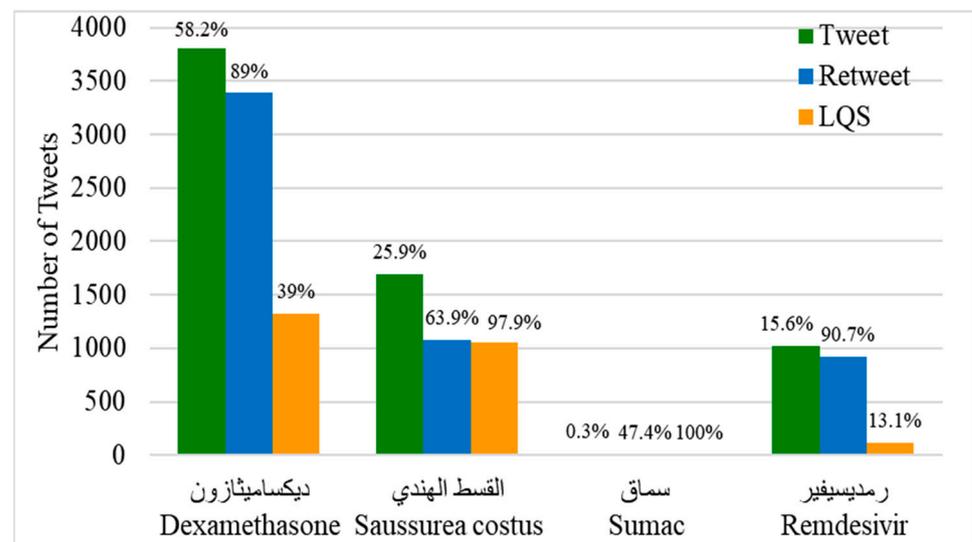


Figure 4. The information prevalence, information occurrence ratio, and the quality of the information prevalence of coronavirus remedy terms in tweets and retweets.

Figure 4 also shows that 100% of the 'سماق' ('Sumac') and 98% of the 'القسط الهندي' ('Saussurea costus') tweets were retweeted from LQS. These are considered high percentages and the tweets may include added misinformation. This was followed by 39% of 'ديكساميثازون' ('Dexamethasone'), and 13% of 'رمديسيڤير' ('Remdesivir') tweets retweeted from LQS, respectively. Spreading information from unofficial accounts means that they are not posted from HQS (e.g., WHO, Ministry of Health). Noteworthy here is that information prevalence from LQS is mostly for non-medical or folk remedy. This is evident by the high percentage of retweets from LQS for 'sumac' and 'Saussurea costus'.

Next, a sense of the conversation incidence of the defined terms over the study period was calculated. The results, as summarized in Figure 5, show how conversation volume for each term changes over time. It can be seen that a high conversation incidence about 'ديكساميثازون' ('Dexamethasone') and 'القسط الهندي' ('Saussurea costus') during the first week of the study period that then decreases through the rest of the study period. In contrast, the conversation incidence about 'رمديسيڤير' ('Remdesivir') is low during the first week, followed by a high incidence during the second week, then followed by a low incidence through the last two weeks. As for 'سماق' ('Sumac'), the incidence of its conversations is the least from the beginning to the end of the study period when compared with the other three terms.

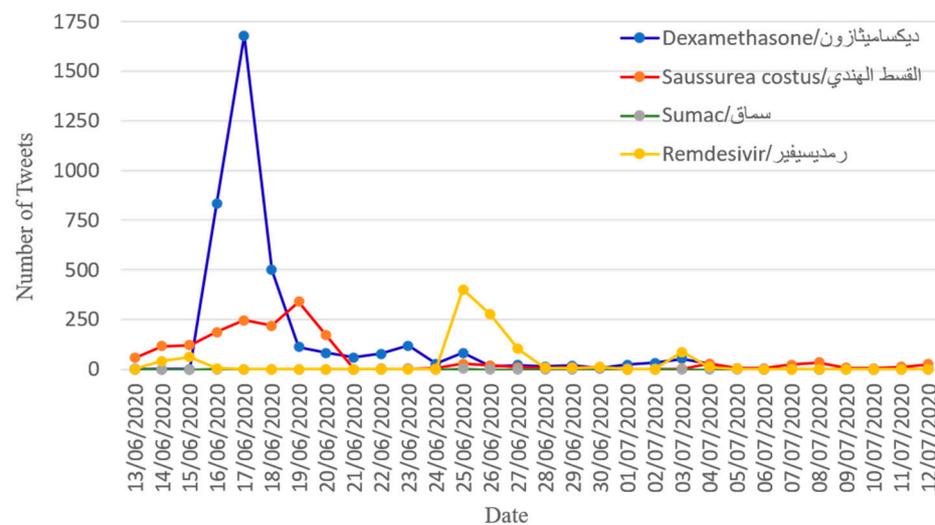


Figure 5. Information incidence over time.

Figure 6 shows a word cloud providing an intuitive overview of the most prevalent words that appear at least 150 times in the tweets, where words are sized based on how frequently they appear in the whole tweets. As seen, the most prevalent word was 'كورونا' ('Coronavirus') numbering (4466) times. The other top 10 most prevalent words and their frequency were as follows: 'ديكساميثازون' ('Dexamethasone') (3807), 'علاج' ('Treatment') (3203), 'القسط الهندي' ('Saussurea costus') (1694), 'الصحة' ('The health') (1398), 'مرضى' ('Patients') (1338), 'رمديسيفير' ('Remdesivir') (1021), 'بريطانيا' ('Britain') (1011), 'فيروس' ('Virus') (931), 'عقار' ('Drug') (912), and 'دواء' ('Medicine') (790).



Figure 6. Word Cloud of frequently mentioned words in novel coronavirus in Arabic tweets in (a) Arabic and (b) English language.

3.2. Google Trends

Since words alone give us limited insight into people's perception and topics of conversation, the pattern of information prevalence about COVID-19 by using an autocorrelation analysis was identified to show how the prevalence of these terms change over time. Search volume and time series data from Google Trends [19] and their relation to the number of total cases are observed. The pattern of internet search volume shown in Figure 7 did not reveal a cyclic trend, as can be seen from the autocorrelation diagram in Figure 8. No seasonal trends were found ($ACF = -0.2$). The research volume remained constant from March to June 2020, apart from a peak in late March and early April.

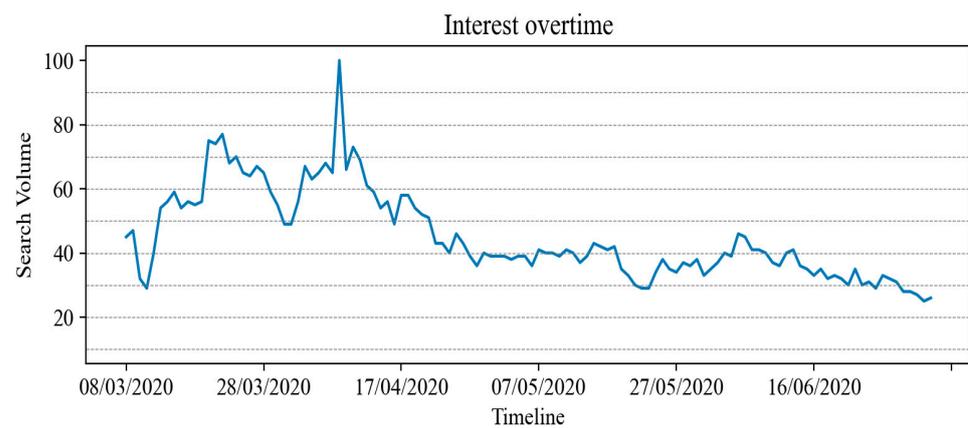


Figure 7. Google Trends-based COVID-19 hit search volume in Arabic from March to June 2020.

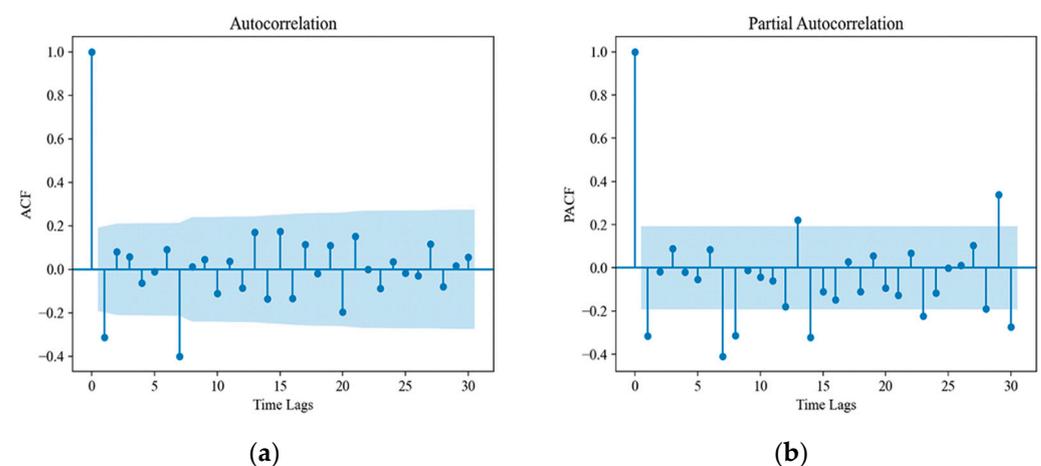
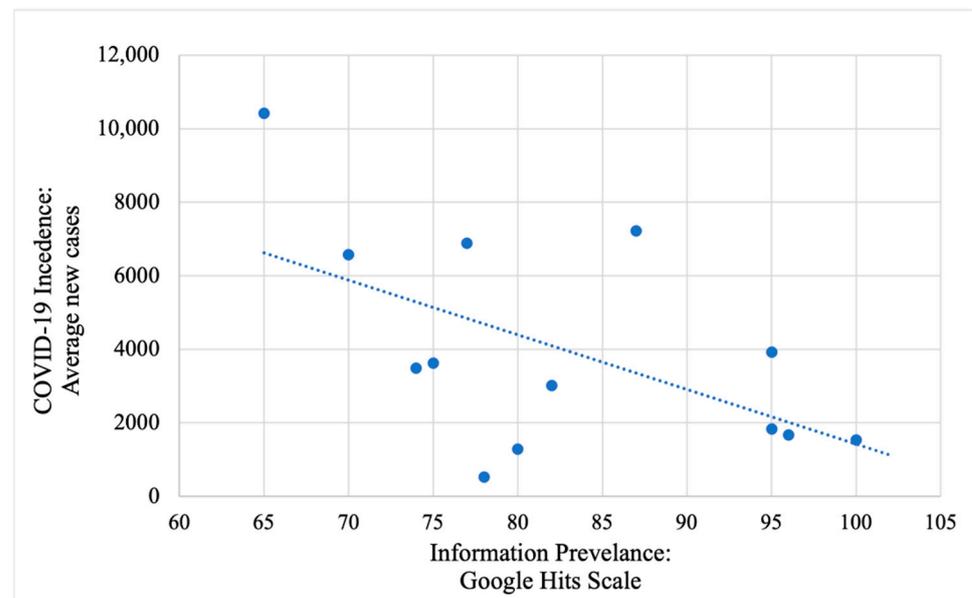


Figure 8. Autocorrelation plot for Coronavirus hit search in Arabic (a) and partial autocorrelation plot (b), showing no cyclical pattern or regular trend ($\lambda = 1$, $d = 0$, $D = 0$, $CI = 0.95$, CI type = white).

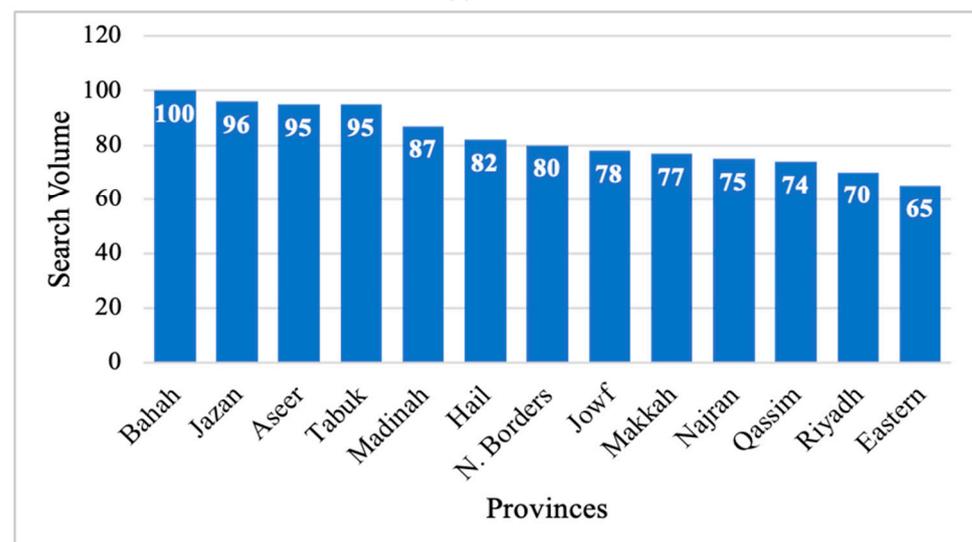
COVID-19 cases and symptoms are the most searched coronavirus terms by the Arabic users in Saudi's provinces. Our analysis revealed that provinces with a higher number of COVID-19 cases per 1 million people had lower Google search interest related to COVID-19 (e.g., Makkah, Riyadh, and the Eastern provinces). Figure 9 shows the information prevalence indicator represented by the popularity scale provided by Google Trends. The COVID-19 related search queries have a significant strong negative correlation with the incidence of the average of new COVID-19 cases per 1 million in these provinces (Pearson $r = -0.63$, $p < 0.05$).

3.3. Google Mobility Data

Shapiro–Wilk tests found that all mobility data categories were normally distributed, with the exception of Parks. Since COVID-19 virus has an incubation period of 5 to 14 days, mobility and potential exposure will have a delayed effect on the new confirmed cases. Thus, mobility data with 14 days lag difference was correlated to current new cases. Moreover, the Saudi government imposed a curfew on 23 March 2020. Thus, two separate analysis were conducted. The first one is for mobility before curfew from 16 February to 23 March, and its impact on new cases from 1 March to 5 April. The second analysis is from 24 March to 14 June, which represents the time period after curfew, and its impact on new cases from 6 April to 25 June.



(a) Disease Incidence.



(b) Information Prevalence.

Figure 9. Disease incidence scatterplot (a) versus Information prevalence (b).

The analysis of Google mobility data before imposing curfew shows that the increased reduction of mobility in all categories (e.g., grocery & pharmacy, parks, retail and recreation, and residential) is highly and negatively correlated with decreased number of total cases in Saudi Arabia (see Table 3). In other word, the less the movement, the lower the number of new cases. However, after the curfew, the analysis shows a positive weak relationship between new cases and mobility in grocery & pharmacy and retail and recreation, which indicates that reduction in mobility has a weak relationship with new cases. On the other hand, there is a negligible correlation between parks and residential mobility and new COVID-19 cases (Pearson $r = 0.2$). Noteworthy, all these correlations are statistically significant with $p < 0.05$. It is noteworthy that these correlations are all very large [29]. Figure 10 shows the scatterplots of these four correlations.

Table 3. Correlation coefficients for different community mobility data with new COVID-19 cases per one million before and after imposing curfew by the Saudi government, and their statistical significance according to their two tailed p -value.

	Mobility Categories							
	Grocery & Pharmacy		Parks		Retail & Recreation		Residential	
	r	p -Value	r	p -Value	r	p -Value	r	p -Value
Before curfew	−0.8	0.007 *	−0.6	0.04 *	−0.8	0.003 *	−0.8	0.01 *
After curfew	0.4	0.0001 *	0.2	0.006 *	0.4	0.0002 *	0.2	0.003 *

* Statistically significant. Note: Parks mobility data was not normally distributed, so a Spearman rank correlation was conducted. All other correlations were Pearson correlations.

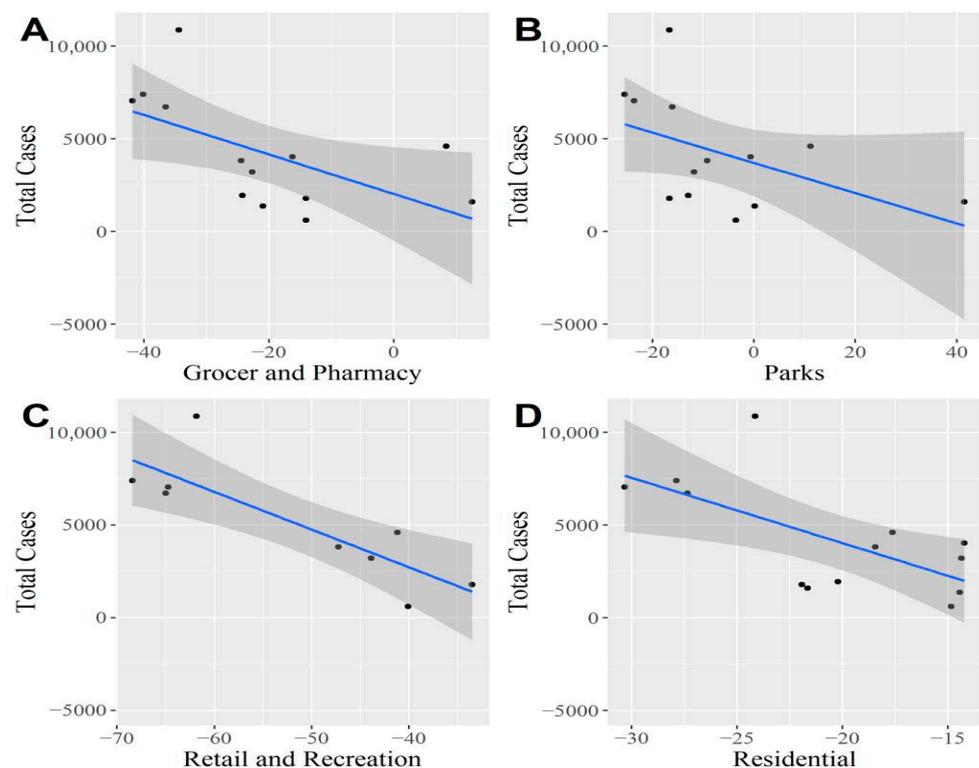


Figure 10. Scatter plots of total cases and (A) grocer and pharmacy, (B) parks, (C) retail and recreation, and (D) residential mobility data in Saudi Arabia. Blue lines represent the linear line of best fit, and the shaded areas are the 95% confidence interval of the line of best fit.

A multiple regression analysis was conducted to examine the relationship between new cases of COVID-19 and various potential predictors, namely overall mobility, total confirmed cases, and information prevalence. The results of the multiple regression indicated that the model explained 88.31% of the variance and that the model was a significant predictor of the number of new cases, $F(3, 105) = 272.9$, $p < 0.001$. While all three independent variables were significantly predictive of new cases (total cases $B = 0.02$, $p < 0.001$; mobility $B = -12.62$, $p < 0.001$; information prevalence $B = -16.82$, $p < 0.001$), the proportion of variance in new cases is uniquely explained by each predictor varied substantially: total cases explained 44.32% of the variance, mobility explained 14.11% of the variance, and information prevalence explained 7.28% of the variance (22.60% of the explained variance was therefore not unique to one predictor, but shared between at least two). When this analysis was repeated with the four separate mobility dimensions as opposed to one average mobility score, the results were similar. The model explained 89.16% of the variance, and it was a significant predictor of the number of new cases, $F(6, 102) = 149.0$, $p < 0.001$ (see Table 4). Total cases, information prevalence, grocery and pharmacy mobility, and parks mobility were all statistically significant predictors of the number of new cases, but

retail and recreation mobility and residential mobility were not. This is might be due to the impact of the curfew, since recreation and residential mobility have the highest percentage of reduction (80% and 100%, respectively) [20]. Therefore, they were not significant predictors. In contrast, the reduction in mobility in grocery and pharmacy and parks is low (25% and 40%, respectively), which explains why these two are significant predictors. It should be noted, however, that 58.22% of the variance in the number of new cases was not unique to one predictor, likely due to the four mobility variables all correlating very highly with each other (all $r_s \geq 0.80$). Therefore, the low percentages of unique variance explained are likely due to multicollinearity.

Table 4. Multiple regression results for total cases, information prevalence, and four mobility scores predicting number of new cases.

Predictor	B	p	% Variance Explained
Total cases	0.024	<0.001 *	20.48%
Information prevalence	−22.06	<0.001 *	9.69%
Grocery and pharmacy mobility	−46.57	0.009 *	0.11%
Parks mobility	28.01	0.012 *	0.46%
Retail and recreation mobility	11.54	0.335	0.05%
Residential mobility	4.43	0.402	0.15%

* Statistically significant.

4. Discussion

Social media and online platforms have become key distribution channels for information surrounding COVID-19. One of the main advantages of these data sources is that they can be obtained both anonymously and easily, and early in the epidemic at a low cost. Although governments and health organizations have used these platforms as communication channels to reach the public, populations can become overwhelmed with the propagation of misinformation and disinformation, as it is increasingly prevalent. Inveillance can monitor how people react to the evolution of the pandemic over time, as well as identify common beliefs, concerns, or hopes regarding prevention, treatment, and vaccines.

COVID-19 is the first global pandemic of the digital era. Our results provide a means for understanding people's perceptions of risk and recognizing the prevalence of misinformation. It also provides insights into the public's information seeking behaviors and the role of mobility in contributing to the number of cases. Designing an effective risk communication strategy based on digital health solutions can play a major part in the fight against COVID-19 [13]. Understanding the population's perception and compliance with health guidelines can be utilized to develop more effective RCCE and digital health solutions and health campaigns to improve public awareness and perception.

The retweet was the most common form of interaction. It is important to note that most of the popular folk remedies are shared from unreliable sources or low-quality sources. Nevertheless, it must be noted that most retweets are for information from reliable sources or high-quality sources. Twitter users are passionate in notifying their followers with any information about possible coronavirus treatments, whether they are mainstream medical treatments or folk remedies, and whether they are from reliable or unreliable sources.

The conversation volume for each remedy changes over time based on the global news and trends. For example, retweets related to "Dexamethasone" spiked when it was announced as a cure to treat COVID-19 [30]. This is an important finding, suggesting that people are aware and up-to-date with any news related to COVID-19. Thus, the use of Twitter as a source of information can cause a wide spread of misinformation. These results build on existing evidence of the internet search behavior and the extent of the results in [31,32].

The results of this study suggest a mean through which public health officials can identify the most common forms of misinformation and indicate that the best remedy

for the COVID-19 infodemic is to broadcast timely and correct information to Twitter audiences. On the other hand, people confronting the COVID-19 pandemic face unfamiliar circumstances and are anxious for any source of information. Therefore, it is important to deliver accurate and appropriate information through risk communication and digital health campaigns [33]. Identifying the top examples of misinformation in Arabic in Twitter helps government authorities and experts debunk comments on misinformation and rumors. Specifically, Twitter analytics can enable them to choose what keywords and hashtags are more prevalent among the audience. For example, governments can effectively communicate accurate information on how to deal with the symptoms of coronavirus. These implications can be utilized to enhance government fact-checking services, risk communication, and health campaigns.

While this finding is encouraging, there are many more types of misinformation that need to be analyzed to understand the total impact of myths on the number of COVID-19 cases. It is crucial to monitor infodemics and build a system to limit the spread of misleading information and potentially harmful Arabic content [34]. Twitter and Google have updated their approaches to misleading information about COVID-19 by classifying its related content for English-speaking audiences [35]. However, it is important to consider filtering the Arabic language content as well. Therefore, there is a need for a strong system to detect rumors or misinformation posted on social media, possibly under the supervision of the government and health monitors such as the work in [21]. Systems such as the one described in [21] could help governments to set laws to prevent the spread of misinformation such as imposing penalties, fines, blocking accounts, or omitting the blogs that have misinformation related to health or health treatment [36].

The study provides a new insight into the relationship between the increased perceptions and the number of new cases. The results demonstrate how the internet search volume can be used to measure peoples' perception of the risk of the pandemic. In line with the hypothesis, the Saudi provinces where people do less research about COVID-19 tend to have more cases. These results build on an existing evidence in [37]. This result emphasizes the importance of encouraging people to stay up to date about the pandemic. Besides, future digital health initiatives should focus on reaching people using different communication channels to make a greater impact on the epidemic, specifically by identifying digital communication channels and influencers with the potential to reach larger audiences.

A further finding in our analysis of public mobility and the spread of COVID-19 reveals that increased mobility leads to an increased number of cases. A significant strong negative correlation between the reduction in mobility and the number of daily new cases was found, especially within the grocery and residential categories. Specifically, a low mobility score means more reduction in mobility. Therefore, as people move about more, more cases arise. The results demonstrated match the state-of-the-art methods used in [38] and [39]. This is an important finding in the understanding of the effectiveness of quarantining and social distancing during the pandemic. This result aligns with the findings in [40] that emphasizes a high correlation between social distancing and daily new cases. In general, when enforcing new health guidelines to combat the pandemic, it is crucial to convey the reasoning behind, and aims of, such guidelines. This will help manage people's fear and increase the likelihood that they will adhere to the measures imposed on them during the crisis.

Among all COVID-19 predictors (e.g., total cases, mobility, and information prevalence), the total cases are the strongest predictor, while the information prevalence is the weakest. This result highlights the strongest predictor of the number of daily new cases. Our findings suggest that the transmission rate increases as the total number of cases increase, therefore emphasizing the importance of imposing more restrictive measures in places where the total number of cases is higher. Moreover, this result can be used to design customized risk mitigation measures for different provinces or cities based on mobility data and total cases [41]. This could contribute to the controlling of the pandemic and supporting economic activities without applying extreme restrictions in unnecessary

areas. Applying unnecessary precautions can delay the economy re-opening, which result in a lower economic activity levels and a slower economic growth. Although this does not mean that information prevalence should be ignored since it is statistically significant, total cases and mobility increase the transmission rate of the virus, better explaining the results. Besides, among the four mobility categories, grocery and pharmacy and parks were significant predictors since the mobility reduction in these two categories was lower than the others. This result emphasizes the importance of curfews during the pandemic.

Study Limitations

Some limitations can be attributed to this study in the spread of misinformation in tweets regarding COVID-19. Although multiple search terms can be used to collect COVID-19 related tweets such as “COVID-19”, “COVID”, “SARS-CoV2”, “Wuhan virus”, and “Chinese virus”, the official Arabic name used by the Saudi Ministry of Health and the most widely used term for COVID-19 in Saudi Arabia is “كورونا” (‘coronavirus’). Therefore, the latter term with a set of predefined search terms for COVID-19 remedies from Twitter trends were used to collect the tweets. In addition, the study only analyzed tweets written in the Arabic language and geolocated in the country of Saudi Arabia. Moreover, the Twitter standard search API does not allow for collecting tweets older than one week. Thus, tweets posted before 13 June 2020, could not be collected. Furthermore, the study could not collect tweets from private accounts. Thus, the results only represent organizations and people who use Twitter for trading news and information. Therefore, the results may not reflect the real number of tweets discussed by all kinds of users which may include misinformation related to COVID-19. All these points may restrict the generalizability of the results of this study.

Locally distributed information and socioeconomic status for different cities are important factors to be considered in the analysis. However, the Saudi Ministry of Health is the main source of information nationwide. Thus, there is no information dissemination from local medical resources in different cities or hospitals other than what the Ministry of Health distributes. This is a protection measure from the Saudi government to guarantee the correctness of the disseminated information. Moreover, there is no local data available about socioeconomic status for different cities. Therefore, the data used for the regression analysis are at a national level and does not consider socioeconomic status as a predictor. It should also be noted that many of the results presented in this study are correlational, and as such does not imply any directional associations.

5. Conclusions and Future Work

In this study, we used three different platforms to conduct an infodemiology and infoveillance survey of COVID-19 in Saudi Arabia. First, using Twitter, the study showed the prevalence of misinformation and folk remedies that might hinder people from following medical information from trusted sources. Second, using Google Trends, we investigated the relationship between information prevalence in each province in Saudi Arabia and the number of daily new cases. The results showed that there is a strong negative relationship, which indicates that literacy among people is an important factor in controlling the pandemic. Third, we used Google mobility data to investigate the impact of mobility on the number of daily new cases. The results showed that reduction of mobility can decrease the number of daily new cases. Finally, the study examined the relationship between new cases of COVID-19 and various potential predictors, namely overall mobility, total confirmed cases, and information prevalence. The analysis showed that the total confirmed cases is the most significant predictor.

Governments, policymakers, and healthcare providers can use these results to design effective programs, awareness messages, and community campaigns to increase the perception and knowledge about COVID-19. Moreover, governments can apply customized restrictive measures for different places based on the total cases and mobility.

It is important to emphasize that this study focuses on the spreading of misinformation regarding COVID-19. One direction of future work will focus on the opposite side of this study, which discusses the information prevalence on Twitter to analyze public awareness and perception towards COVID-19 in LQS versus HQS. Moreover, establishing directional associations links between COVID-19 information distribution, mobility, and deaths allows for future research to investigate causal associations.

Author Contributions: Conceptualization, R.A. and A.B.; methodology, R.A. and A.B.; software, R.A. and A.B.; validation, R.A. and A.B.; formal analysis, R.A. and A.B.; investigation, R.A. and A.B.; resources, R.A. and A.B.; data curation, R.A. and A.B.; writing—original draft preparation, R.A. and A.B.; writing—review and editing, R.A. and A.B.; visualization, R.A. and A.B. Both authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ACF	Autocorrelation Function
COVID-19	Novel coronavirus disease
HQS	High-Quality Sources
LQS	Low-Quality Sources
RCCE	Risk Communication and Community Engagement
WHO	World Health Organization

References

- Harapan, H.; Itoh, N.; Yufika, A.; Winardi, W.; Keam, S.; Te, H.; Megawati, D.; Hayati, Z.; Wagner, A.L.; Mudatsir, M. Coronavirus disease 2019 (COVID-19): A literature review. *J. Infect. Public Health* **2020**, *13*, 667–673. [CrossRef] [PubMed]
- World Health Organization. Managing the COVID-19 Infodemic: Promoting Healthy Behaviours and Mitigating the Harm from Misinformation and Disinformation. 2020. Available online: <https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation> (accessed on 22 November 2020).
- Nsoesie, E.O.; Cesare, N.; Müller, M.; Ozonoff, A. COVID-19 Misinformation Spread in Eight Countries: Exponential Growth Modeling Study. *J. Med. Internet Res.* **2020**, *22*, e24425. [CrossRef] [PubMed]
- Cinelli, M.; Quattrocchi, W.; Galeazzi, A.; Valensise, C.M.; Brugnoli, E.; Schmidt, A.L.; Zola, P.; Zollo, F.; Scala, A. The COVID-19 social media infodemic. *Sci. Rep.* **2020**, *10*, 1–10. [CrossRef] [PubMed]
- Tagliabue, F.; Galassi, L.; Mariani, P. The “Pandemic” of Disinformation in COVID-19. *SN Compr. Clin. Med.* **2020**, *2*, 1287–1289. [CrossRef]
- Smith, G.D.; Ng, F.; Li, W.H.C. COVID-19: Emerging compassion, courage and resilience in the face of misinformation and adversity. *J. Clin. Nurs.* **2020**, *29*, 1425. [CrossRef]
- World Health Organization. Risk Communication and Community Engagement Readiness and Response to Coronavirus Disease (COVID-19): Interim Guidance. 19 March 2020. Available online: <https://apps.who.int/iris/bitstream/handle/10665/331513/WHO-2019-nCoV-RCCE-2020.2-eng.pdf?sequence=1&isAllowed=y> (accessed on 22 November 2020).
- Rovetta, A.; Bhagavathula, A.S. COVID-19-Related Web Search Behaviors and Infodemic Attitudes in Italy: Infodemiological Study. *JMIR Public Health Surveill.* **2020**, *6*, e19374. [CrossRef]
- Eysenbach, G. Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet. *J. Med. Internet Res.* **2009**, *11*, e11. [CrossRef]
- Cuan-Baltazar, J.Y.; Muñoz-Perez, M.J.; Robledo-Vega, C.; Pérez-Zepeda, M.F.; Soto-Vega, E. Misinformation of COVID-19 on the Internet: Infodemiology Study. *JMIR Public Health Surveill.* **2020**, *6*, e18444. [CrossRef]
- Kouzy, R.; Jaoude, J.A.; Kraitem, A.; El Alam, M.B.; Karam, B.; Adib, E.; Zarka, J.; Traboulsi, C.; Akl, E.W.; Baddour, K. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. *Cureus* **2020**, *12*. [CrossRef]
- Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). Available online: <https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf> (accessed on 22 November 2020).

13. Ho, Y.-H.; Tai, Y.-J.; Chen, L.-J. COVID-19 Pandemic Analysis for a Country's Ability to Control. The Outbreak Using Little's Law: Infodemiology Approach. *Sustainability* **2021**, *13*, 5628. [CrossRef]
14. Mavragani, A.; Ochoa, G. Google Trends in Infodemiology and Infoveillance: Methodology Framework. *JMIR Public Health Surveill.* **2019**, *5*, e13439. [CrossRef]
15. Singh, L.; Bansal, S.; Bode, L.; Budak, C.; Chi, G.; Kawintiranon, K.; Padden, C.; Vanarsdall, R.; Vraga, E.; Wang, Y. A First Look at COVID-19 Information and Misinformation Sharing on Twitter. *ArXiv* **2020**, arXiv:2003.13907v1.
16. Sahni, H.; Sharma, H. Role of social media during the COVID-19 pandemic: Beneficial, destructive, or reconstructive? *Int. J. Acad. Med.* **2020**, *6*, 70. [CrossRef]
17. Bernardo, T.; Liang, C.; Sun, L.; Marlicz, W.; Mavragani, A. Infodemiology and Infoveillance: Scoping Review. *J. Med Internet Res.* **2020**, *22*, e16206. [CrossRef]
18. Chaves-Montero, A.; Relinque-Medina, F.; Fernández-Borrero, M.Á.; Vázquez-Aguado, O. Twitter, Social Services and Covid-19: Analysis of Interactions between Political Parties and Citizens. *Sustainability* **2021**, *13*, 2187. [CrossRef]
19. Saudi Arabia Social Media Statistics 2020. 8 September 2020. Available online: <https://www.globalmediainsight.com/blog/saudi-arabia-social-media-statistics/> (accessed on 22 November 2020).
20. Google. COVID-19 Community Mobility Reports. Available online: <https://www.google.com/covid19/mobility/> (accessed on 22 November 2020).
21. "Covid19 Infodemic Observatory" from FBK-WHO Partnership. 2020. Available online: <https://covid19obs.fbk.eu/#/> (accessed on 22 November 2020).
22. COVID-19 Dashboard. Available online: <https://covid19.moh.gov.sa/> (accessed on 27 March 2021).
23. Twitter. Search Tweets. Available online: <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview> (accessed on 13 June 2020).
24. Google Trends. Available online: <https://trends.google.com/trends/explore> (accessed on 31 July 2020).
25. Zerrouki, T. Tashaphyne 0.3.4.1. Available online: <https://pypi.org/project/Tashaphyne/> (accessed on 22 November 2020).
26. Abdelali, A.; Darwish, K.; Durrani, N.; Mubarak, H. Farasa: A fast and furious segmenter for arabic. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA, USA, 12–17 June 2016.
27. Alahdal, H.; Basingab, F.; Alotaibi, R. An analytical study on the awareness, attitude and practice during the COVID-19 pandemic in Riyadh, Saudi Arabia. *J. Infect. Public Health* **2020**, *13*, 1446–1452. [CrossRef]
28. Bence, J.R. Analysis of Short Time Series: Correcting for Autocorrelation. *Ecology* **1995**, *76*, 628–639. [CrossRef]
29. Cohen, J. A power primer. *Psychol. Bull.* **1992**, *112*, 155–159. [CrossRef]
30. World Health Organization. Coronavirus Disease (COVID-19): Dexamethasone. 2020. Available online: <https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-dexamethasone#:~:text=Dexamethasone%20is%20a%20corticosteroid%20used,for%20critically%20ill%20patients> (accessed on 22 November 2020).
31. Rovetta, A.; Bhagavathula, A.S. Global Infodemiology of COVID-19: Analysis of Google Web Searches and Instagram Hashtags. *J. Med. Internet Res.* **2020**, *22*, e20673. [CrossRef]
32. Jang, H.; Rempel, E.; Roth, D.; Carenini, G.; Janjua, N.Z. Tracking COVID-19 Discourse on Twitter in North America: Infodemiology Study Using Topic Modeling and Aspect-Based Sentiment Analysis. *J. Med. Internet Res.* **2021**, *23*, e25431. [CrossRef]
33. Mheidly, N.; Fares, J. Leveraging media and health communication strategies to overcome the COVID-19 infodemic. *J. Public Health Policy* **2020**, *41*, 1–11. [CrossRef] [PubMed]
34. Centers for Disease Control and Prevention. Stop the Spread of Rumors. Available online: <https://www.cdc.gov/coronavirus/2019-ncov/daily-life-coping/share-facts.html> (accessed on 22 November 2020).
35. Van Der Linden, S.; Roozenbeek, J.; Compton, J. Inoculating Against Fake News About COVID-19. *Front. Psychol.* **2020**, *11*, 2928. [CrossRef] [PubMed]
36. Baraybar-Fernández, A.; Arrufat-Martín, S.; Rubira-García, R. Public Information, Traditional Media and Social Networks during the COVID-19 Crisis in Spain. *Sustainability* **2021**, *13*, 6534. [CrossRef]
37. Commodari, E.; La Rosa, V.; Coniglio, M. Health risk perceptions in the era of the new coronavirus: Are the Italian people ready for a novel virus? A cross-sectional study on perceived personal and comparative susceptibility for infectious diseases. *Public Health* **2020**, *187*, 8–14. [CrossRef]
38. Fang, H.; Wang, L.; Yang, Y. Human mobility restrictions and the spread of the Novel Coronavirus (2019-nCoV) in China. *J. Public Econ.* **2020**, *191*, 104272. [CrossRef]
39. Engle, S.; Stromme, J.; Zhou, A. Staying at Home: Mobility Effects of COVID-19. *SSRN Electron. J.* **2020**. [CrossRef]
40. Qureshi, A.I.; Suri, M.F.K.; Chu, H.; Suri, H.K.; Suri, A.K. Early mandated social distancing is a strong predictor of reduction in peak daily new COVID-19 cases. *Public Health* **2021**, *190*, 160–167. [CrossRef]
41. Ceccato, R.; Rossi, R.; Gastaldi, M. Travel Demand Prediction during COVID-19 Pandemic: Educational and Working Trips at the University of Padova. *Sustainability* **2021**, *13*, 6596. [CrossRef]