
IDENTIFYING INFORMATION SUPERSPREADERS OF COVID-19 FROM ARABIC TWEETS

Sarah Alqurashi*

Center of Innovation and Development in Artificial Intelligence
Umm Al-Qura University
Makkah, Saudi Arabia
s43980127@st.uqu.edu.sa

Abdulaziz Alashaikh

Computer and Networks Engineering Department
University of Jeddah
Jeddah, Saudi Arabia
asalashaikh@uj.edu.sa

Eisa Alanazi

Center of Innovation and Development in Artificial Intelligence
Umm Al-Qura University
Makkah, Saudi Arabia
eaaanazi@uqu.edu.sa

July 9, 2020

Abstract

Since the first confirmed case of COVID-19, information was spreading in large amounts over social media platforms. Information spreading about the COVID-19 pandemic can strongly influence people's behavior. Therefore, identifying information superspreaders (or influencers) during the COVID-19 pandemic is an important step towards understanding public reactions and information dissemination. In this work, we present an analysis over a large Arabic tweets collected during the COVID-19 pandemic. The presented study construct a network from users' behaviors to identify information superspreaders during the month of March, 2020. We employed both HITS and PageRank algorithms to analyze the influence of information spreading, and compared the ranking of the users. The results show that both HITS and PageRank discovered a similar subset of superspreaders with 40% were found to be verified Twitter accounts.

1 Introduction

The emerging coronavirus disease (COVID-19) poses a great challenge to humanity. The first confirmed case of the novel coronavirus (SARS-CoV-2), was recognized in the city of Wuhan, China in December 2019 [1]. With its high spread rate, the coronavirus has affected people across the globe with more than 10 million confirmed cases [2]. Since then, healthcare officials rushed to identify and contain the spread of COVID-19, and in the meanwhile information was spreading in large amounts over social media platforms at unprecedented levels. Due to the impact of the disease and the absence of continuous official and validated information associated with the disease, medical misinformation has been spreading and disseminating rapidly on different social media platforms. Information spreading about the COVID-19 pandemic in social media can strongly influence people's behaviors and impact the effectiveness of the measures established by governments and health officials to contain the virus. In recent years, researchers' interest has shifted to focus on whether the diffusion of information about a topic can be maximized by spreading information

*Corresponding author

via specific individuals, often known as “influencers” [3] or superspreaders. Influencers² have characteristics that distinguish them from others such as expertise or connectivity. These characteristics allow them to influence a large number of people in their network. Influencers (or Superspreaders) in social media can contribute to the spread of misinformation about COVID-19 pandemic. Therefore, identifying them during the COVID-19 pandemic is important to understand public reactions and information dissemination.

Twitter platform is well suited to studying the role of influencers during COVID-19. With over 300 million monthly active users [4], Twitter provides direct access to an unprecedented amount of content which may inflate rumors and questionable information. Twitter includes various types of users that communicate in the same way through tweets to their followers. Another unique feature of Twitter is the ability to propagate posts to mass audiences. Therefore, it is possible to measure and compare the influence of individuals in a standard form by the activity that is obvious on Twitter. In addition, it is possible distinguish users and posts by their location, time, age, sex, and/or language, albeit some users’ information may be private or missing. This enables researchers to focus on studying and exploring a specific pattern of interest in the Twitter network.

In this paper, we investigate the problem of finding information superspreaders of COVID-19 on Arabic Twitter and apply two ranking algorithms known as HITS and PageRank [5, 6] that are used for web page ranking to rank users’ influence. We focus on the analysis of the dynamic interactions among users where actions such as retweets, replies, mentions, and quotes are represented and taken into account. In nutshell, we analyze a large directed weighted graph where nodes represent twitter users and edges are the dynamic interactions happening in Twitter.

The rest of the paper is organized as follows. In the next section, we provide an overview of related work. section 3 introduces common algorithms used in influencers analysis. section 4 details our experiment. We discuss out analytical results on superspreaders in section 5. Conclusion and foreseeable work is presented in section 6

2 Literature Review

Twitter has over 330 million monthly active users worldwide generating, on average, hundreds of millions of tweets per day [7]. This rapid flow of daily information that is posted on Twitter through the interaction between users is impressive. Twitter has a large number of users, but finding the most influential ones is of great benefit. The most common algorithms to find and rank influential users are PageRank and HITS algorithms [5, 6] which are used to measure the importance of webpages on the Internet. Both of these algorithms have been widely applied to the Twitter social network [8]. The PageRank algorithm is a widely used influence measure. Kwak et al. [9] evaluated influential users on Twitter by ranking users based on the number of followers and PageRank. However, they ignored the impact of user’s communication in the Twitter network. Tunkelang [10] proposed the TunkRank algorithm which is an adaptation of PageRank to Twitter. The TunkRank algorithm strongly depends on follow-up relationships among users. The algorithm defines the influence score based on how much attention a user can get from his/her followers and how much attention the followers bring to the user from their own networks of followers. The influence score is a numerical estimation of readers that will ultimately read a particular user’s posts. In an extension to PageRank, Weng et al. [11] proposed TwitterRank that considers both the topic similarity between users and the follow-up structure of the social network into account. In TwitterRank the user influence varies in different topics because a Twitter user can have interests in many distinct topics. The U-R model proposed by Huang and Xiong [12] evaluates user’s influence based on the PageRank algorithms. The users influence in the follow-up networks can be ranked by the U-R value. Yang et al. [13] proposed a modified version of the HITS algorithm for Twitter social network that can be applied on both User Graph and Tweet Graph with retweet and follower-follower relationship. However, the follow-up social network alone does not reflect the influence a user gets when the users interact with other users. The users’ interactions and tweets are important to identify the most influential users and demonstrate values of influence in information diffusion. Therefore, many efforts applied PageRank and HITS algorithms in identifying influential users according to user interactions[14, 15, 16, 17].

Twitter has become a rich source of data to study various topics during the novel Coronavirus. Many studies focused on understanding general behaviors during the pandemic. Which usually depends on the analyzes of the content of posts on social media. Abd-Alrazaq et al. [18] identified the main English topics discussed by Twitter users during COVID-19 pandemic using the Latent Dirichlet allocation (LDA) topic modeling technique. The tweets were analyzed using word frequency techniques unigram and bigram. The work categorized the topics into four categories, which are topics related to the source of the virus, the origin of the virus, the impact of the virus on people and countries, and methods to reduce the spread of the virus. Sentiment analysis was applied to the identified topics, only two topics had negative sentiment which are the topics related to death by COVID-19 and increase racism. Saad et al. [19] investigated public awareness regarding COVID-19 in countries that are most affected by the corona-virus pandemic using Twitter data.

²In this paper, we use influencers interchangeably with (information) superspreaders

To study the response of each country to the pandemic, the authors identified temporal patterns of COVID-19 trends. The authors used topic modeling techniques and sentiment analysis to understand user reactions to the measures. The results show that countries with lower COVID-19 cases have high information flow and subsequently contribute to public awareness. An analysis of how decision makers influence each other regarding COVID-19 on Twitter was presented by Sha et al. [20] using topic modeling and network analysis. A social network created for U.S. governors and presidential cabinet members was created to identify the most influential politician during the pandemic using granger causality [21]. They observed that the influence of the governors was affected by the size of the state.

With the rising spread of rumors and misinformation about COVID-19 in social media, it is difficult to obtain accurate information. This leads to more fear and obstructs measures taken to tackle the virus. Therefore, most of the research has focused on identifying influencers who influence information dissemination and on studying the impact of infodemic. Infodemic is a combination of two words "information" and "epidemic" that usually indicates a rapid spread of accurate and inaccurate information about something like a disease.

Escuredo and Tarazon [22] identified and ranked influencers on Twitter during COVID-19 based on the centrality metrics for an undirected behavioral social graph that consists of over 10k nodes and 100k edges. The flow property has been added to each edge of the network, where the flow increases when adding an edge that already exists in the network. The study discussed the difference in data flow between the centrality metrics.

Ahmed et al. [23] presented a study of the conspiracy theory that links 5G to the spread of COVID-19 on Twitter by identifying the influential users in spreading this misinformation using behavioral social networks and betweenness centrality measure. The author have identified 10 influential accounts related to the conspiracy, most of the influential users in spreading the conspiracy were citizen accounts. The author also identify five of the most refer to websites relating to 5G and COVID-19 conspiracy. Most of these websites were not trusted sources. Also, they manually analyzed the content of the tweets to determine whether people really think COVID and 5G are related only, a small percentage of the users believed in the conspiracy.

The impact of infodemic on social networks was investigated by Mourad et al. [24] where they analyzed the content of twitter data to discover the misleading tweets about COVID-19 and identify the source of the tweets using natural language processing (NLP) techniques. The users' profile were analyzed to determine whether or not users have a medical specialty. 83% of users have no medical specialty. The study describes the dangerous impact of the dissemination of misinformation and the importance of using influencers from outside the medical field to help spread the correct information about the virus to a large number of people.

The study in [25] presented a comparative study of infodemic in a wide range of social media platforms which are Twitter, Instagram, YouTube, Reddit, and Gab. The content of the posts and comments from these platforms were analyzed. The author measured interaction and engagement in COVID-19 using the cumulative number of posts and the number of reactions of each post. The epidemic models were used to model the spread of information and analysis the spreading of misinformation for all the five platforms. They found out that the spreading of reliable or questionable information doesn't present notable differences.

Most of the previous studies focused on analyzing the English content of Twitter. In this work, we provide the first analysis on identifying COVID-19 information influencers (e.g., superspreaders) from Arabic tweets.

3 Twitter Social Network

During the past few years, Twitter has been utilized as an important tool to communicate emergencies to the public in real-time [26]. In fact, during emergency situations, Twitter users provided detailed information that warns others from dangerous situations such as epidemic outbreak and severe weather conditions which helps others to make better/careful decisions. As a platform, Twitter enables fast and concise information exchange in various circumstances. Twitter allows users to publish posts that called tweets with a limit of 140 characters. There are no rules or formats on what an acceptable tweet is. Furthermore, the post created by the user becomes public and is immediately accessible and available to others in their social network. Unlike other social network services, Twitter doesn't require users to grant friends accepts to other users before the latter can see his/her posts. To follow conversations on Twitter, users need to follow other members for their tweet to appear on their newsfeed. Moreover, Twitter users receive instant and automated feedback when they post a Tweet from other members of their social network. Therefore, Twitter has been a target platform for researchers to test interests, latest events, and users' attitudes as well as identifying the most influential members of a group. It is essential to know which individuals have the power to alter people's decisions on a specific topic because of their authority, position, or relationship [27]. These individuals are known as influencers. Identifying the influence of users can be helpful in many fields, such as marketing [28], political sciences [29],

and rumor spreading [8]. Identifying influencers during a pandemic is important to understand public reactions and information dissemination.

To identify influencers on Twitter, we look for influence by means of interconnections with others so it is natural to consider the follow-up and behavioral networks when evaluating users' social influence. The Twitter social network normally represented with a directed graph $G = (N, E)$ where N is a set of user nodes and E is a set of directed edges that express/depict how the nodes are connected. The follow-up relationship is the most traditional representation of the Twitter social network. In the follow-up graph, N is a set of Twitter users, and the edges express the follow-up relationship among users. Such that an edge $(A, B) \in E$ indicates that user A is a follower of user B . However, since the behavioral networks are based on the dynamic interaction between two users such as to reply, mention, quote, and retweet a tweet, this dynamic interaction can be represented by a directed graph with a set of user nodes and the edges represent the different interactions between users (for example, a replay from one user to a tweet posted by a different user). The follow-up network represents the potential for information flow between accounts whereas the behavioral network represents actual information flow. Therefore, the behavioral network represents a stronger relationship between accounts compared to the follow-up network. This makes behavioral networks ideal for studying the role of influencers in information dissemination. For the detection and ranking of influential users, there are essentially two well-known algorithms that must be highlighted. Therefore, a description of these two algorithms is described in the following subsections.

3.1 PageRank algorithm

A well-known algorithm developed at the early stage of Google's search engine is the PageRank algorithm to determine a page's relevance or importance [5]. The algorithm assigns each page a relative numeric score of importance and authority by estimating the quality and quantity of the links it contains. The algorithm is calculated using a simple iterative algorithm and is based on the random surfer model which is the probability of a random surfer of the web reaching their target page starting from links at random from one page to another.

To analyze the twitter social graph, the PageRank score is considered for each user. Assuming a list of users U that have interaction such as retweeted or replied to user V tweet, the PageRank score of each user in U is equally distributed to user V . For example, user U_i has a PageRank score of X , and it has interacted with a total number of five users' tweets. The PageRank Score of X from U_i will be send to each users that U_i has interacted with. It can be mathematically expressed as follows [5]:

$$PR(A) = \frac{1-d}{N} + d \sum_{U_i \in P(A)} \frac{PR(U_i)}{P(U_i)} \quad (1)$$

where $PR(A)$ is the PageRank calculated for certain user node A and N corresponds to the total number of nodes in the graph. $P(A)$ is the set of nodes pointing to node A and $P(U_i)$ is the number of outlinks from node A to nodes $U_1 \dots U_n$. Lastly, the parameter d is a damping factor which represents the probability of jumping to a random user node and is usually set to 0.85.

3.2 The Hyperlink-Induced Topic Search (HITS) algorithm

The HITS algorithm is a Web page ranking algorithm developed by Kleinberg [6]. The algorithm is based on the idea of authorities and hubs, where the authorities are the pages that have a greater amount of incoming links and the hubs are the pages that have outgoing links to many related authorities. A page is considered a good hub if it points out to many good authorities and a good authority page is pointed by many good hubs. Unlike the PageRank algorithm HITS algorithm starts with a subset of the network consisting of pages that relevant to the search query and any pages that link to a relevant page. As for Twitter behavioral social graph analysis, we consider each node N as a user node and each node contains a set of Hub score Nh and Authority score Na . Given that S is the set of nodes that point to N , while T is the set of nodes that N point to. The hub and authority scores are calculated in the iterative procedure as follows: The authority score $auth(N)$ of user N is the normalized sum of all the hub scores of nodes S that point to user N

$$auth(N) = \sum_{S \in (N_{to})} hub(S) \quad (2)$$

The hub score $hub(N)$ of user N is the normalized sum of all the authority scores of nodes that T points to.

$$hub(N) = \sum_{T \in (N_{from})} athou(T) \quad (3)$$

4 Experiment

4.1 Data Collection

To identify Arabic tweeters influencers during the COVID-19 pandemic, we use tweets from our large Arabic dataset [30]. The dataset contains more than 4,514,136 million tweets from January 1, 2020 to April 30, 2020. We created a list of the most common Arabic keywords associated with COVID-19 to collect the tweets. Using Twitter’s streaming API and Tweepy Python library, we extracted and filtered the tweets based on the keywords list. We store the full object of the tweets including the timestamp of the tweet, the id of the tweet, user profile information including the number of followers, and geolocation of the tweet in a MongoDB NoSQL database. The stream contained information about how many times a tweet is retweeted. The retweets is identified by tweets beginning with “RT @user”. The dataset also contains information about mentions, replies, and quotes which represent a public conversation between users on Twitter [30]. However, in our experiment, we only considered the tweets retrieved in March (from March 1, 2020, to March 30, 2020). Our choice of March is motivated by the fact that almost all Arabic-speaking countries confirmed their first cases of COVID-19 in early March, 2020. Table 1 shows some statistics of March tweets. The daily distribution of these tweet is shown in Figure 1.

Number of Tweets :	1,708,788
Total Unique Users:	521,598
Number of original tweets:	4,513,576
Total Verified Users :	4932
The Number Average Followers Count:	4779

Table 1: Summary statistics of March tweets

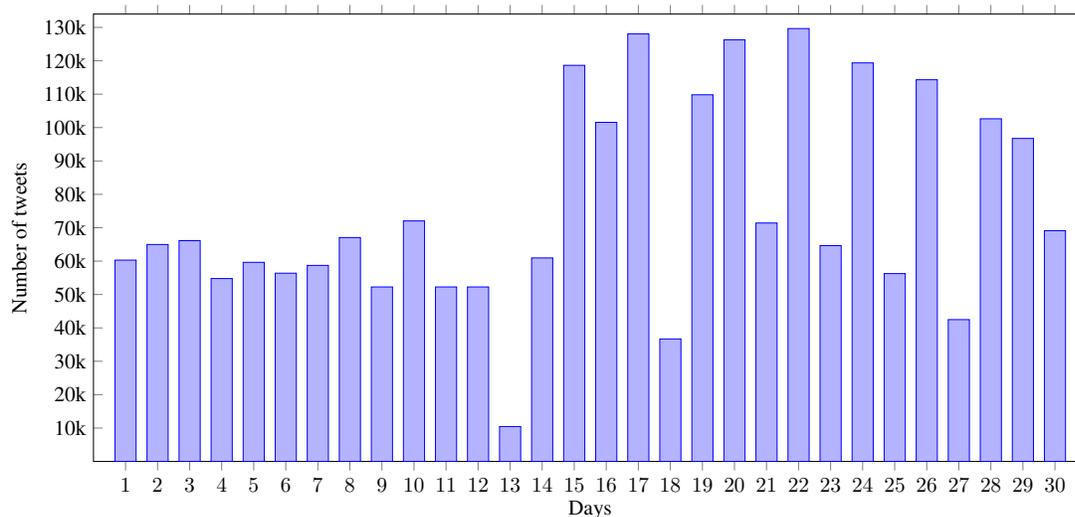


Figure 1: Number of daily tweets collected over the month of March 2020.

4.2 Data Preprocessing

Before we contracted the network, the data was cleaned by removing the null values and considering only the necessary columns. We have over 1.7 million tweets in our dataset some of these tweets are written by the same user. Therefore, we only included unique users. Each tweet in the dataset was analyzed to extract users who mentioned, retweeted, quoted, or replied to a tweet in the dataset. By exploring the JSON fields, retweeted status, and quoted status in our database, we recovered user id of source tweet form each retweet and the user id of a quoted tweet of each quote. We included the users who mentioned a tweet by exploring the JSON field mentions embedded within the JSON field entities. Finally, from the JSON field in reply to, we extracted the user id in which the tweet was a reply to. The fields such as retweeted by, mentioned by, quoted by, replied by were added. These fields contain the id numbers of users that interact with the tweet in an array format. We extracted users locations from the user profile and we removed deleted or disabled accounts.

4.3 Network construction

We construct a behavioral directed graph $G = (V, E)$ where V represent the users and E is the set of weighted edges that represent the dynamic interaction between users which can be a retweet, a mention, a reply, or a quote. Every connection in the graph involves two accounts: the source and the target. The direction of the edge varies depending on the type of interaction. For instance, if account A is mentioning account B , an edge is constructed from the authoring account A to account B that is mentioned. Indeed, further exchanges of the mention produce edges pointing in opposite directions. However, not all interactions are the same. When a tweet is retweeted or quoted it's shared by another user, therefore a bigger network of users is exposed to the tweet. Hence, retweet and quote edges have a higher interaction level than mentions or replies. To quantify the interactions, we assign to mentions and replies a weight of 0.3, and 0.5 for retweets and quotes, respectively. Indeed, the result is a directed weighted graph where every edge is annotated with a weight that represents the interaction weight.

Example: Consider, the following hypothesized collection of tweets:

1. @userA: "According to @saudiMOH COVID-19 symptoms are Fever, Cough, and Shortness of breath". Generated an edge from @userA to @saudiMOH.
2. @userB: Retweeted @userA tweet. Two edges from userB to userA and saudimOH are generated.
3. @userC: "@userA @saudiMOH @userE did you see this?" Three edges generated from @userC to @userE, @userA, and @saudiMOH account.

Figure 2 shows the resulted weighted graph out of the above interactions.

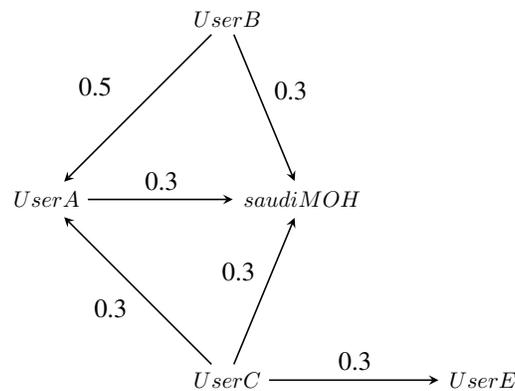


Figure 2: The graph generated from the example discussion.

Table 2 presents general properties of each ranking algorithm we used. The type of graph used by each ranking algorithms is represented. The number of interactions represents the total number retweet, mentions ...etc, for the behavioral network.

The Number of unique users	533,023
The type of the network	behavioral
The Number of Interaction	985,359
Number of iteration	100

Table 2: General properties of the networks

5 Results and Discussion

We classified the accounts into two categories: individual and organizational accounts. We further classified individual accounts into Journalist, Celebrity, or Citizens account. Organizational accounts are classified into media/news, government, charity, or business accounts. The behavioral network expresses a more intimate set of actions. Our behavioral network consists of disconnected components, as not all nodes are connected and only 6% of the nodes are

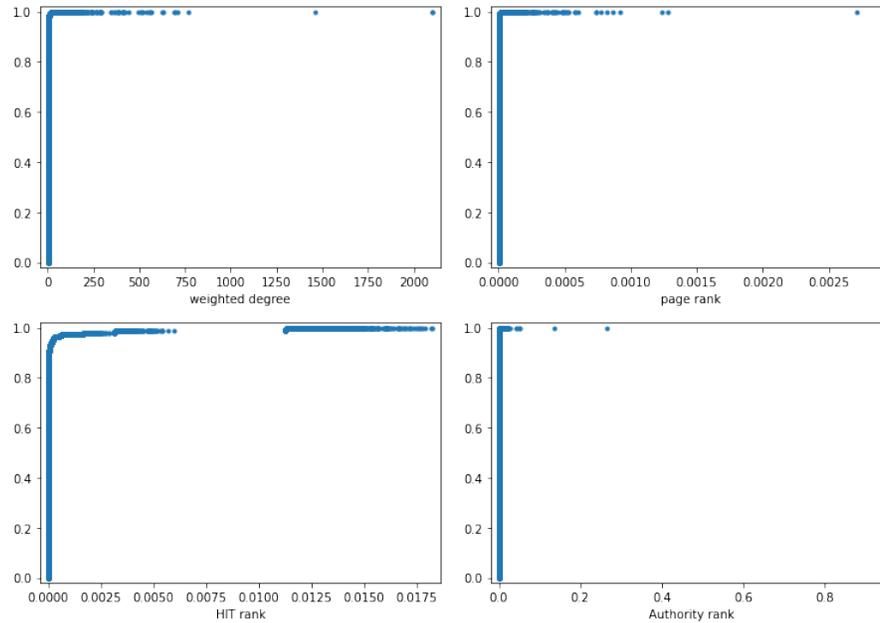


Figure 3: The cumulative distribution functions for the users (or nodes) in our graph based on different scores. The weighted degree for a node is the total weights associated with the incoming and outgoing edges.

strongly connected. According to Smith et al. [31] on Twitter network structures classification, our behavioral network is considered a community clusters network. This type of network forms when a popular topic such as COVID-19 in our case develops multiple smaller groups formed around a few hubs each with its audience. Given the fact that our behavioral network consists of almost one million edges, it is certainly impossible to visualize it. Figure 3 shows the distribution function of different parameters related to our analysis. The y-axis is the fraction of the points that have a value smaller than or equal to the corresponding x-axis value. One can observe that, with few exceptions for the superspreaders nodes, all other nodes have a close to zero value.

Nevertheless, we have selected a subgraph that contains the Top-100 superspreaders in both HITs and PageRank and their direct connections. This subgraph, visualized in Figure 4, shows different modularity classes where the node size is adjusted based on the weighted indegree. Nodes with black color represent the users in the Top-100 superspreaders either in HITs or PageRank.

Table 3 shows information about the top 20 users that were ranked based on PageRank algorithm. Many of the accounts have large number of followers, however, there are accounts in our network with far more followers. This indicates that the top 20 accounts have more interaction with and from their followers.

In our network, we consider the hub account to be a user who frequently interacts with many unconnected accounts resulting in edges pointing away from the hub node and the authority account to be a user who receives many interactions from others resulting in edges pointing to the authority. Thus, the authority accounts can be considered more influential than the hub accounts. Table 4 and Table 5 show the top 20 screen name that were ranked based on the hub and authority scores, respectively. The SaudiMOH937 account, which is the official account of the ministry of health in Saudi Arabia, ranked as the fourth based on authority score and as the twentieth based on hub score. Such a finding is reasonable, since the SaudiMOH937 accounts broadcast information which receives many interactions from others and also provide assistance via Twitter.

It is observed that both the PageRank algorithm and the HITS algorithm authority scores show a similar subset of users in their respective TOP 20 with a percentage of 50%. Notably, the order is significantly different, only 10% of users have the exact same rank for both algorithms and 40% of the top 20 users have verified accounts.

In summary, from the identified top 20 influential Arabic tweeters for both HITS and Page Rank algorithms, there were 19 organization accounts; 8 media/news accounts, 8 Saudi government accounts, 2 charity accounts, and one business account. In addition, there were 7 individual accounts all were male; 3 individual accounts were journalists accounts, 2 celebrity accounts, and 2 citizens accounts.

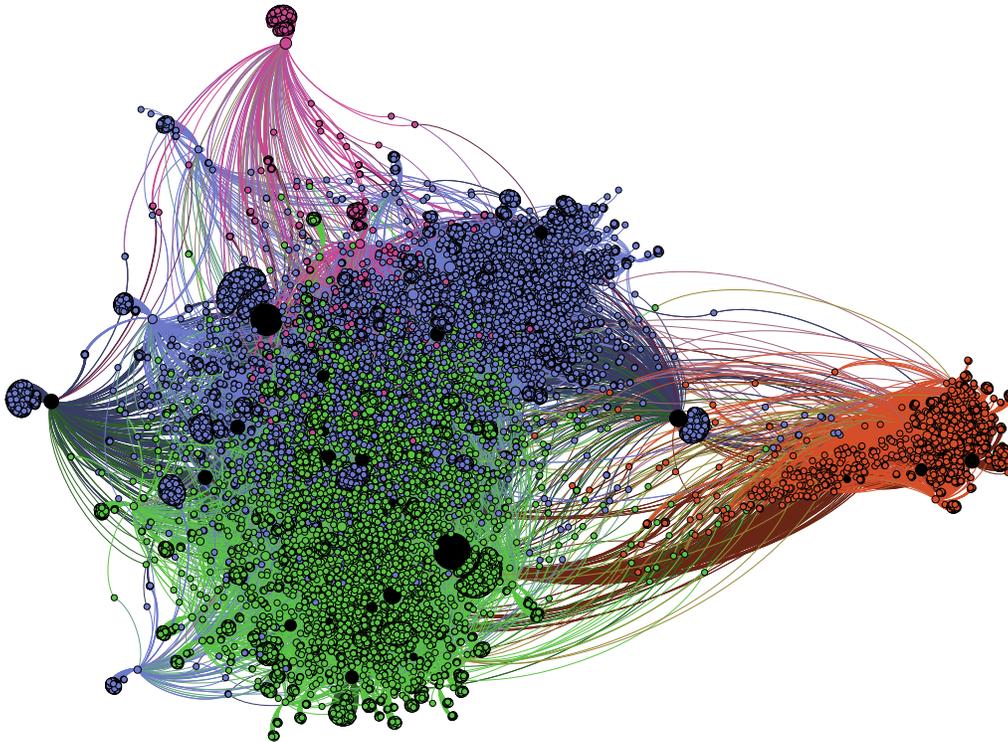


Figure 4: The subgraph for the top 100 users in HIT and PageRank.

<i>Account Class</i>	<i>Verified</i>	<i>Followers Count</i>	<i>Location</i>
Media/news account	False	12.8M	Saudi Arabia
Government account	True	3.6M	Saudi Arabia
Medianews account	False	3.9M	Saudi Arabia
Citizens account	True	3.5M	Saudi Arabia
Medianews account	True	1M	Saudi Arabia
Medianews account	False	7.1M	Saudi Arabia
Medianews account	True	3.9M	Saudi Arabia
Citizens account	False	4M	Saudi Arabia
Journalists account	True	5.6M	Qatar
Medianews account	True	8.8M	Egypt
Journalists account	False	99.5K	Unknown
Medianews account	True	1.6M	Kuwait
Medianews account	True	3M	Egypt
Medianews account	True	19M	UAE
Government account	True	2M	Saudi Arabia
Medianews account	True	13M	Saudi Arabia
Medianews account	True	590.1K	Unknown
Celebrity account	True	8.2M	Egypt
Journalists account	True	435.8K	Sweden
Celebrity account	True	6.7M	Unknown

Table 3: Page-rank Top 20 Twitter Accounts

As shown by the result, most of the influential accounts were news accounts followed by government accounts for ministries in Saudi Arabia. Three of these accounts belong to the Ministry of Health in Saudi Arabia, followed by journalists, celebrities, and citizens. The news and government accounts that belong to the Ministry of Health in Saudi Arabia were at the top of the influential list for both algorithms in the Arabic Twitter in March. This is due to the emergence of confirmed cases of COVID-19 in March in the Middle East, especially in the gulf area. Therefore, people turned to verified news sources such as the Health Ministry accounts on Twitter for information about COVID-19 which indicates that Twitter became a significant source of news and information about the disease dynamics. However, not all of the news accounts are reliable sources. 50% of the influential accounts were located in Saudi

<i>Account Class</i>	<i>Verified</i>	<i>Followers Count</i>	<i>Location</i>
Government account	True	3.6M	Saudi Arabia
Medianews account	False	12.8M	Saudi Arabia
Citizens account	True	3.5M	Saudi Arabia
Government account	True	226.8K	Saudi Arabia
Medianews account	True	1M	Saudi Arabia
Government account	True	3.8M	Saudi Arabia
Government account	True	2M	Saudi Arabia
Medianews account	False	7.1M	Saudi Arabia
Charity account	True	27.3K	Saudi Arabia
Government account	True	2.8M	Saudi Arabia
Business account	False	6.1K	Saudi Arabia
Government account	True	173.8K	Unknown
Charity account	False	4.4K	Saudi Arabia
Medianews account	True	3.9M	Saudi Arabia
Citizens account	False	4M	Saudi Arabia
Medianews account	True	13M	Saudi Arabia
Government account	True	2.5M	Saudi Arabia
Government account	True	1M	Saudi Arabia
Medianews account	True	19M	UAE
Government account	True	977.6K	Saudi Arab

Table 4: HITS Top 20 Authorites Twitter Accounts

<i>Account Class</i>	<i>Verified</i>	<i>Followers Count</i>	<i>Location</i>
Citizens account	False	1K	Unknown
Citizens account	False	348	Unknown
Citizens account	False	10.6K	Saudi Arabia
Citizens account	False	778	Saudi Arabia
Citizens account	False	1.5K	Unknown
Citizens account	False	246	Unknown
Citizens account	False	16	Unknown
Citizens account	False	218	Saudi Arabia
Citizens account	False	382	Unknown
Citizens account	False	2.9K	Saudi Arabia
Citizens account	False	64	Saudi Arabia
Citizens account	False	13	Saudi Arabia
Citizens account	False	1.4K	Saudi Arabia
Citizens account	False	260	Saudi Arabia
Citizens account	False	5K	Unknown
Citizens account	False	280	Unknown
Citizens account	False	61	Saudi Arabia
Citizens account	False	38	Unknown
Citizens account	False	213	Unknown
Government account	True	226.8K	Saudi Arabia

Table 5: HITS Top 20 Hub Twitter Accounts

Arabia. This can be explained as Saudi Arabia is one of the largest countries in the world that actively uses Twitter. Statistics show that in October 2019 Twitter had 10.9 million active users from Saudi Arabia [32] and 41% of Internet users in Saudi Arabia use Twitter. Moreover, most of the accounts that have been ranked as influencers are verified well-known accounts on Twitter and have a large number of followers from all over the Arab world which confirms the role of the social media in providing greater opportunities to influence others directly in a way that was previously not possible. Therefore, individuals play an important role in the spread of information through Twitter, regardless of the information quality and credibility.

In addition to identifying influencers, ranking them based on their influence power is also important to use their influence positively. The results show that both of the ranking algorithms ranked diverse groups with different levels of influences on Twitter. It is also noted that both of the algorithms classified four accounts of the Saudi Ministry of Health as influencers accounts. In the Page Rank list of influencers, the Saudi Ministry of Health account was in the second place, and in the fourth place was the account of the Saudi Minister of Health, Tawfiq Al-Rabiah. While in the HITS algorithm, four health-related accounts were ranked, at the top of the list was the Saudi Ministry's account, then in the third-place was the account of the Saudi Minister of Health Tawfiq Al-Rabiah. The Ministry of Health 937 account ranked as the fourth and the account of the spokesman of the Ministry of Health ranked as a twelfth in the list.

This indicates the important role social media, especially Twitter, played in educating individuals and society in Saudi Arabia. However, it is possible scenario that influencers in some groups that have been identified, such as media/news, journalists, and celebrities, may contribute to disseminating false information about COVID-19. It is our hope that information superspreaders use their influence on social platform to the greater good of the society.

6 Conclusions

Identifying information superspreaders in social media can be an effective strategy for increasing awareness during pandemics. Using reliable and trustworthy superspreaders could help in reducing the spread of misinformation about diseases and promoting better health practices to a larger audience. This study identified influencers on Twitter in the Arabic content of Twitter during the month of March 2020. However, more research is needed in this regard. One potential work is analyzing the role of influencers in disseminating misinformation and whether they utilize their influence for the greater good of the society. Another line for research is analyzing information superspreaders in other social platforms such as Facebook and Instagram. In this regards, the used methods in this study can be adopted to other social platforms albeit with different interactions.

Acknowledgements

This work was supported by King Abdulaziz City for Science and Technology. Grant Number:5-20-01-007-0033

References

- [1] World Health Organization, “Coronavirus disease (covid-19) pandemic,” <https://www.who.int/emergencies/diseases/novel-coronavirus-2019>, 2020, (accessed: 2020-04-02).
- [2] SaudiCdc, “dailyupdates,” <https://covid19.cdc.gov.sa/daily-updates/>, 2020.
- [3] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone’s an influencer: quantifying influence on twitter,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, 2011, pp. 65–74.
- [4] Statista. Number of monthly active twitter users worldwide from 1st quarter 2010 to 1st quarter 2019. [Online]. Available: <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/>
- [5] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” Stanford InfoLab, Tech. Rep., 1999.
- [6] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.
- [7] Statista, “Leading countries based on number of twitter users as of january 2020,” <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>, 2020, (Accessed: 2020-04-06).
- [8] S. Vosoughi, “Automatic detection and verification of rumors on twitter,” Ph.D. dissertation, Massachusetts Institute of Technology, 2015.
- [9] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?” in *Proceedings of the 19th international conference on World wide web*, 2010, pp. 591–600.
- [10] D. Tunkelang, “Tunrank: A twitter analog to pagerank. ttp,” thenoisychannel.com/2009/01/13/atwitter-analog-to-pagerank, 2009.
- [11] J. Weng, E.-P. Lim, J. Jiang, and Q. He, “Twitterrank: finding topic-sensitive influential twitterers,” in *Proceedings of the third ACM international conference on Web search and data mining*, 2010, pp. 261–270.
- [12] L. Huang and Y. Xiong, “Evaluation of microblog users’ influence based on pagerank and users behavior analysis,” 2013.
- [13] M.-C. Yang, J.-T. Lee, S.-W. Lee, and H.-C. Rim, “Finding interesting posts in twitter based on retweet graph analysis,” in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 1073–1074.
- [14] M. Zhang, C. Sun, and W. Liu, “Identifying influential users of micro-blogging services: A dynamic action-based network approach,” in *Pacific Asia Conference on Information Systems (PACIS) 2011 Proceedings*, 2011, pp. Paper–223.

- [15] S. Kong and L. Feng, "A tweet-centric approach for topic-specific author ranking in micro-blog," in International Conference on Advanced Data Mining and Applications. Springer, 2011, pp. 138–151.
- [16] L. B. Jabeur, L. Tamine, and M. Boughanem, "Active microbloggers: identifying influencers, leaders and discussers in microblogging networks," in International Symposium on String Processing and Information Retrieval. Springer, 2012, pp. 111–117.
- [17] H. Purohit, J. Ajmera, S. Joshi, A. Verma, and A. Sheth, "Finding influential authors in brand-page communities," in Sixth International AAAI Conference on Weblogs and Social Media, 2012.
- [18] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the covid-19 pandemic: infoveillance study," Journal of medical Internet research, vol. 22, no. 4, p. e19016, 2020.
- [19] M. Saad, M. Hassan, and F. Zaffar, "Towards characterizing the covid-19 awareness on twitter," arXiv preprint arXiv:2005.08379, 2020.
- [20] H. Sha, M. A. Hasan, G. Mohler, and P. J. Brantingham, "Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives," arXiv preprint arXiv:2004.11692, 2020.
- [21] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007, pp. 66–75.
- [22] D. Pastor-Escuredo and C. Tarazona, "Characterizing information leaders in twitter during covid-19 crisis," arXiv preprint arXiv:2005.07266, 2020.
- [23] W. Ahmed, J. Vidal-Alaball, J. Downing, and F. L. Seguí, "Covid-19 and the 5g conspiracy theory: social network analysis of twitter data," Journal of Medical Internet Research, vol. 22, no. 5, p. e19458, 2020.
- [24] A. Mourad, A. Srour, H. Harmanani, C. Jenainatiy, and M. Arafah, "Critical impact of social networks infodemic on defeating coronavirus covid-19 pandemic: Twitter-based study and research directions," arXiv preprint arXiv:2005.08820, 2020.
- [25] M. Cinelli, W. Quattrociocchi, A. Galeazzi, C. M. Valensise, E. Brugnoti, A. L. Schmidt, P. Zola, F. Zollo, and A. Scala, "The covid-19 social media infodemic," arXiv preprint arXiv:2003.05004, 2020.
- [26] C. A. Cassa, R. Chunara, K. Mandl, and J. S. Brownstein, "Twitter as a sentinel in emergency situations: lessons from the boston marathon explosions," PLoS currents, vol. 5, 2013.
- [27] F. Riquelme, "Measuring user influence on twitter: A survey," CoRR, vol. abs/1508.07951, 2015. [Online]. Available: <http://arxiv.org/abs/1508.07951>
- [28] L. P. Forbes, "Does social media influence consumer buying behavior? an investigation of recommendations and purchases," Journal of Business & Economics Research (JBER), vol. 11, no. 2, pp. 107–112, 2013.
- [29] S. Stieglitz and L. Dang-Xuan, "Political communication and influence through microblogging—an empirical analysis of sentiment in twitter messages and retweet behavior," in 2012 45th Hawaii International Conference on System Sciences. IEEE, 2012, pp. 3500–3509.
- [30] S. Alqurashi, A. Alhindi, and E. Alanazi, "Large arabic twitter dataset on covid-19," arXiv preprint arXiv:2004.04315, 2020.
- [31] M. A. Smith, L. Rainie, B. Shneiderman, and I. Himelboim, "Mapping twitter topic networks: From polarized crowds to community clusters," Pew Research Center, vol. 20, pp. 1–56, 2014.
- [32] Countries with most twitter users 2019 — statista. [Online]. Available: <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>