Analysis of Twitter Data Using Evolutionary Clustering during the COVID-19 Pandemic

Ibrahim Arpaci[1], Shadi Alshehabi[2], Mostafa Al-Emran[3, *], Mahmoud Khasawneh[4], Ibrahim Mahariq[4], Thabet Abdeljawad[5, 6, 7] and Aboul Ella Hassanien[8, 9]

Abstract: People started posting textual tweets on Twitter as soon as the novel coronavirus (COVID-19) emerged. Analyzing these tweets can assist institutions in better decision-making and prioritizing their tasks. Therefore, this study aimed to analyze 43 million tweets collected between March 22 and March 30, 2020 and describe the trend of public attention given to the topics related to the COVID-19 epidemic using evolutionary clustering analysis. The results indicated that unigram terms were trended more frequently than bigram and trigram terms. A large number of tweets about the COVID-19 were disseminated and received widespread public attention during the epidemic. The high-frequency words such as "death", "test", "spread", and "lockdown" suggest that people fear of being infected, and those who got infection are afraid of death. The results also showed that people agreed to stay at home due to the fear of the spread, and they were calling for social distancing since they become aware of the COVID-19. It can be suggested that social media posts may affect human psychology and behavior. These results may help governments and health organizations to better understand the psychology of the public, and thereby, better communicate with them to prevent and manage the panic.

Keywords: Twitter, social media, evolutionary clustering, COVID-19, coronavirus.

1 Introduction

[1] Department of Computer Education and Instructional Technology, Tokat Gaziosmanpasa University, Tokat, Turkey.

[2] Department of Computer Engineering, University of Turkish Aeronautical Association, Ankara, Turkey.

[3] Applied Computational Civil and Structural Engineering Research Group, Faculty of Civil Engineering, Ton Duc Thang University, Ho Chi Minh, Vietnam.

[4] College of Engineering and Technology, American University of the Middle East, Kuwait, Kuwait.

[5] Department of Mathematics and General Sciences, Prince Sultan University, Riyadh, 66833, Saudi Arabia.

[6] Department of Medical Research, China Medical University, Taichung, 40402, Taiwan.

[7] Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan.

[8] Faculty of Computers and AI, Cairo University, Giza, Egypt.

[9] Scientific Research Group in Egypt, Cairo, Egypt.

[*] Corresponding Author: Mostafa Al-Emran. Email: al.emran@tdtu.edu.vn.

Received: 11 May 2020; Accepted: 05 June 2020.

CMC. doi:10.32604/cmc.2020.011489 www.techscience.com/journal/cmc

Social media is considered as an important channel for promoting risk communication during the previous epidemics such as Zika and Ebola outbreak [Househ (2016)]. Likewise, in the COVID-19 epidemic, people tend to use social media effectively to acquire information about their health [Li, Zhang, Wang et al. (2020)]. Besides, people have used social media platforms to exchange their opinions, and the number of posting tweets related to the COVID-19 has exceeded 468+ million during the outbreak.

With the increasing development of social media networks, these networks have become the carrier of social relationships and message propagation [Gu, Wang and Yin (2019)]. As communication channels, the previous literature categorized these platforms as complex networks due to the massive amount of information [Zhou, Tan, Yu et al. (2019)]. Social media can be used to disseminate useful and reliable information, whereas it may have a negative role in increasing the panic and fear among people [Nayar, Sadasivan, Shaffi et al. (2020)]. Social media platforms such as Instagram, Twitter, and YouTube allow the dissemination of an enormous amount of disinformation and rumors that manipulate human behaviors [Burnap, Williams, Sloan et al. (2014)]. Accordingly, “infodemic” term has been suggested by WHO to point out the severity of disinformation about the COVID-19 outbreak [Zarocostas (2020)].

The dissemination of infodemic about COVID-19 on social media negatively affects individuals’ psychology and behaviors [Sharot and Sunstein (2020)]. The social media panic, which travels faster than the COVID-19 outbreak, has been increasingly threatening humanity [Depoux, Martin, Karafillakis et al. (2020)]. For example, misinformation about a national-lockdown in the USA has disrupted the supply-chain by panicking them in buying stationeries and groceries [Spencer (2020)]. This has negatively affected the public health nutrition and resulted in food insecurities among low socio-economic status individuals [Tasnim, Hossain and Mazumder (2020)].

On the other hand, social media has a significant role in the dissemination of useful information. For example, social media can be employed to predict the number of cases which would help health organizations to identify the potential or high-risk outbreak locations, and thereby, governments can prepare themselves for the intervention or prevention earlier [Qin, Sun, Wang et al. (2020)]. Further, governments can use social media platforms to promote a wide-spread acceptance of “social distancing” and “stay at home” [Kayes, Islam, Watters et al. (2020)].

The COVID-19 pandemic has destructive effects on the economic, psychological, and social well-being of individuals [Duan and Zhu (2020); Wang, Cheng, Yue et al. (2020)]. Not only individuals but also public and non-governmental organizations have been extensively used social media to disseminate information to each other during the COVID-19 pandemic. However, a limited number of research studies focused on the content of social media posts that were shared during the pandemic. Analyzing the social media content may help governments, healthcare organizations, and decision-makers to understand the requirements of the people, and thereby, address that needs properly. Therefore, this study aimed to analyze tweets in a time window during the COVID-19 outbreak and describe the trend of public attention given to the topics related to the COVID-19 epidemic by conducting an evolutionary clustering analysis.

2 Evolutionary clustering

As Twitter receives a massive amount of data generated by people on a daily basis, such data streams need to be analyzed and summarized. The traditional methods of data summarization can be used; however, these methods consume a lot of time and efforts. Artificial intelligence (AI) methods can be employed to recognize unexpected patterns [Jiang, Coffee, Bari et al. (2020)]. Evolutionary clustering is one of such methods, which can be used for data analysis and summarization. Evolutionary clustering is the technique of processing data streams over different time spans to produce a sequence of clustering. In each clustering, the sequence should be similar to the clustering at the previous time span and should clearly represent the data generated during that time span [Chakrabarti, Kumar and Tomkins (2006)]. In evolutionary clustering, the data stream at any given time span should be integrated into the previous clustering at the previous time span, and the clustering will be modified according to the current data stream.

In this research, we applied the evolutionary clustering by studying the evolution of social media terms in clusters rather than studying each term separately. Therefore, each cluster will represent a degree of importance that can be used to detect outliers (i.e., the extreme values—highest or lowest frequencies). To address the issue of event detection, we propose the following algorithm that deals with data streams of terms. In this algorithm, events are modeled as a list of clusters of trending entities over time, benefiting from the clustering methods abilities to learn the distribution of the data objects and defining their interest centers. The best number of clusters is determined from the beginning by the Elbow method. The method will be tried from two clusters and keep increasing it in each step by one until reaching the maximum number of clusters. For each given number of clusters, the sum of intra-cluster distances is computed. The variance of the intra-cluster distances between two consecutive numbers of clusters is computed. Then, the Elbow method looks at the percentage of variance explained as a function of the number of clusters. One should choose a number of clusters so that adding another cluster does not give much better modeling of the data.

```
Algorithm: Evolutionary K-means over time
Input: Data stream $D=\{D_{t_i} \mid D_{t_i}$ is the data stream at time $t_i$, $i=1, \ldots, n\}$, Times $T=\{t_1, t_2, \ldots, t_n\}$
        $D_{t_i} = \{d_{t_i} \mid d_{t_i}\ is\ an\ object\ of\ the\ the\ data\ stream\ D_{t_i}\}$
Output: Set of clustering $C = \{C_{t_i} \mid C_{t_i}$ is a clustering at time $t_i\}$
Apply K-means to $D_{t_1}$ at $t_1$ with an incremental number of clusters
Best number of clusters ($K$) is determined by Elbow algorithm
$C_{t_1} = \{c_1^{t_1}, c_2^{t_1}, \ldots, c_K^{t_1}\}$; $c_j^{t_1}$ is the output cluster $j$ at time $t_1$
foreach $t_i \in T \setminus \{t_1\}$
    Initialize K centroids $c_j^{t_i} = c_j^{t_{i-1}}$; j=1, …, K
    Assign each object to the closest centroid
    Recompute the new cluster centroids $c_j^{t_i}$
```

3 Research methodology

3.1 COVID-19 Twitter dataset characteristics

The COVID-19 Twitter dataset used in this research was launched in the early of 2020 and published by [Banda and Ramya (2020)]. The dataset consists of 43M+(43.845.712 tweets). Those tweets had no retweets were 7.479.940 unique tweets. It contains the top 1000 frequent terms, the top 1000 bigrams, and the top 1000 trigrams. The tweets have identifiers with date and time added.

3.2 Procedure and technique

This study used MATLAB and employed clustering algorithms to analyze the most frequent terms posted on Twitter for a period of nine days, from March 22nd to March 30th, 2020. The main concern is to focus on three types of terms, including unigram, bigram, and trigram. The data were clustered using evolutionary clustering that relies on k-means clustering [Chakrabarti, Kumar and Tomkins (2006)]. Initially, for the first day of the data (i.e., March 22nd), the optimal number of clusters k was determined by applying the Elbow method. Therefore, six clusters were generated for each type of term on that day. Then, these formed clusters were used as initial centroids for clustering the second-day data (i.e., March 23rd). The data is normalized using Min-Max normalization, and the optimal number of clusters k was determined by applying the Elbow method. The method has been tried from two clusters to a maximum number of clusters which is set to 100, and the fraction of variance explained should be greater than 98%. Therefore, six clusters were generated for each type of term on that day. Then, these formed clusters were used as initial centroids for clustering the second-day data (i.e., March 23rd). The k-means method was applied for all remaining days. The generated clusters of the previous day were used as initial centroids for clustering the data of the next day. Each cluster represents the frequency level of terms. All clusters are arranged in a descending order from clusters 1 to 6. For example, cluster 1 represents the highest level of the frequency, while cluster 6 represents the lowest frequency level. The cluster centroid represents the frequency average of terms assigned to that cluster.

4 Results

It is depicted from the results shown in Figs. 1 to 3 that unigram terms were trended more frequently than bigram and trigram terms. For example, cluster 1 of unigram terms was trending on Twitter up to 3 times and 70 times more than that of bigram and trigram terms, respectively. For unigram terms shown in Fig. 1, clusters 1 and 2 have the highest frequent terms, and they increase over time. On the other hand, clusters 3-5 are almost trending equally over time. Figs. 1 to 3 show the averages of term frequencies in clusters, which in turn represent the centroids of clusters.

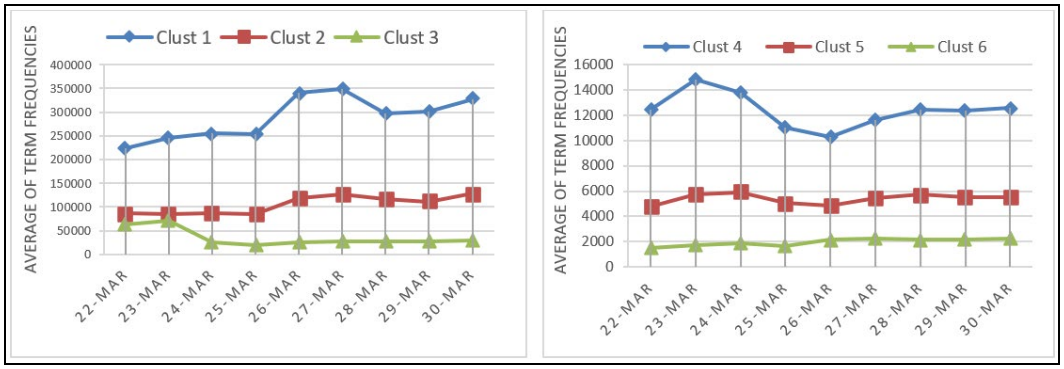

Figure 1: Evolution of clusters for unigram terms

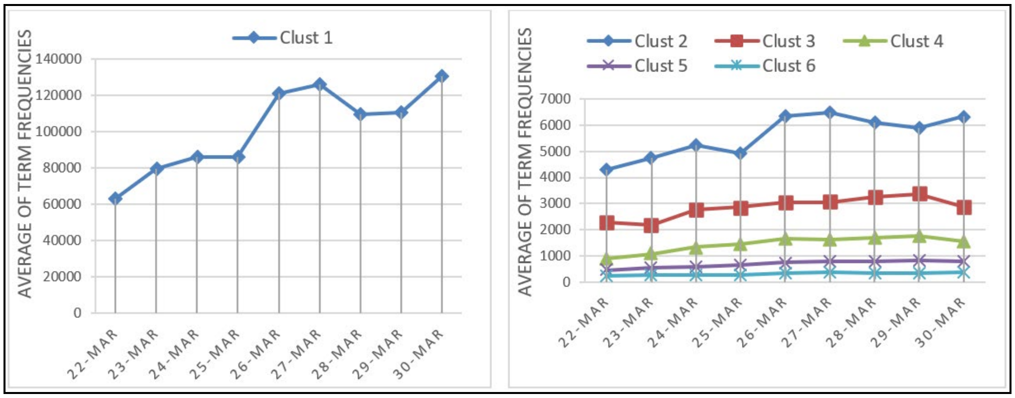

Figure 2: Evolution of clusters for bigram terms

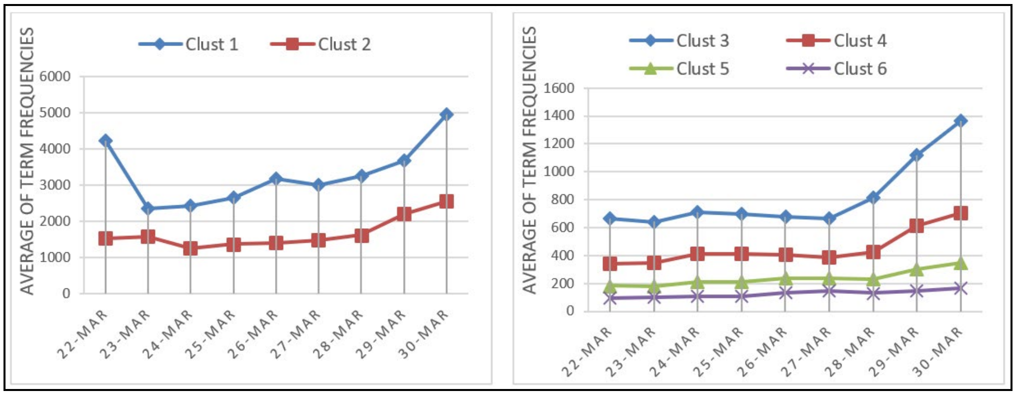

Figure 3: Evolution of clusters for trigram terms

Fig. 2 shows the evolution of clusters for bigram terms. It illustrates that cluster 1 has the highest frequent terms, and it keeps increasing over time, followed by cluster 2, which is, however, 90% less. Other clusters (i.e., clusters 3-6) are approximately unchangeable over time. For trigram terms illustrated in Fig. 3, clusters 1 and 2 are highly trended and keep increasing over time. However, other clusters (i.e., clusters 3-6) are growing slowly over time compared to that in clusters 1 and 2.

Fig. 4 illustrates the evolution of unigram terms in clusters. The "Coronavirus" term remained in the first cluster throughout the whole period, which implies that it was the most trended unigram term on Twitter during that period. The "COVID-19" unigram term was in cluster 3 for the first two days and then jumped to cluster 2 on March 24th, which means it got more importance in people's tweets and became more trending. However, the "COVID-19" term was trending less than "coronavirus" term. Other unigram terms like "virus", "cases", "Pandemic", "test", "spread", "deaths", "crisis", and "lockdown" have fewer tweets from people, and they were fluctuating between clusters 3 and 5 over the time. It is worth mentioning that "test", "spread", "deaths", and "lockdown" unigram terms became more important on March 26th and moved up one cluster.

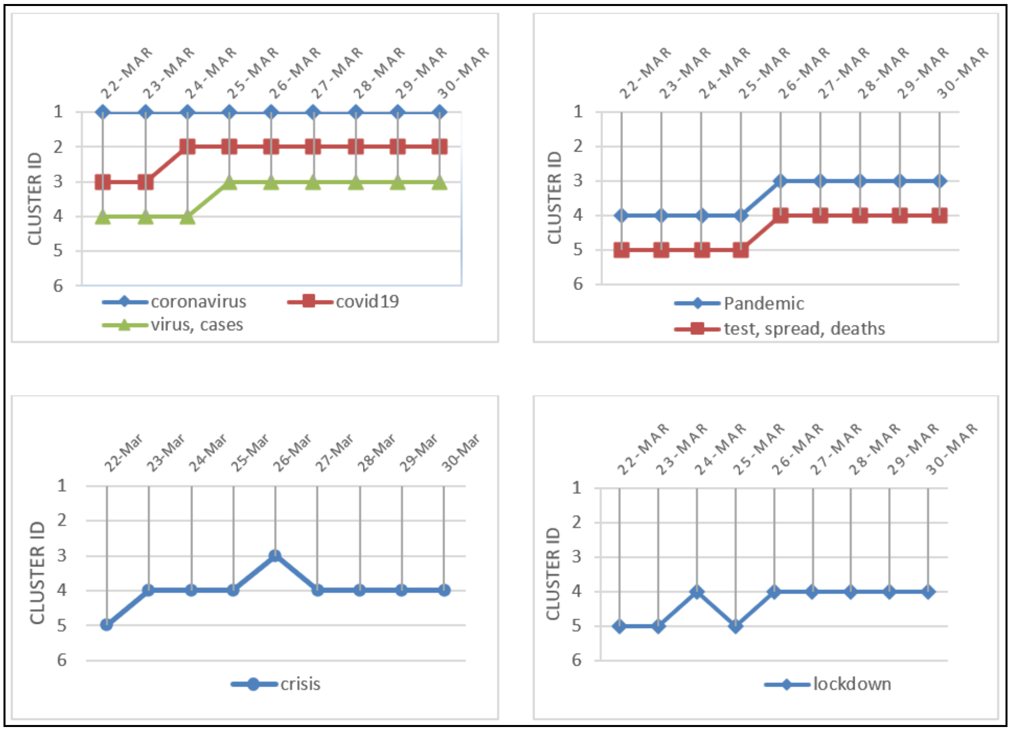

Figure 4: Evolution of unigram terms in clusters

The evolution of bigram terms in clusters for the specified period is shown in Fig. 5. The bigram term "COVID-19" remained in the first cluster throughout the whole period, which implies that it was the most trended bigram term on Twitter for that period, followed by "stay home" and "coronavirus pandemic" bigram terms. They had the same importance throughout the whole period and remained in cluster 2. The "social distancing" bigram term has had less importance than the previous bigram terms;

however, it was tweeted more and jumped up one cluster on March 30th. The bigram term "coronavirus cases" has had the same importance as of "stay home" and "coronavirus pandemic" bigram terms for the first four days; however, it became more important and got more tweets for the duration of the next three days (i.e., March 26-29).

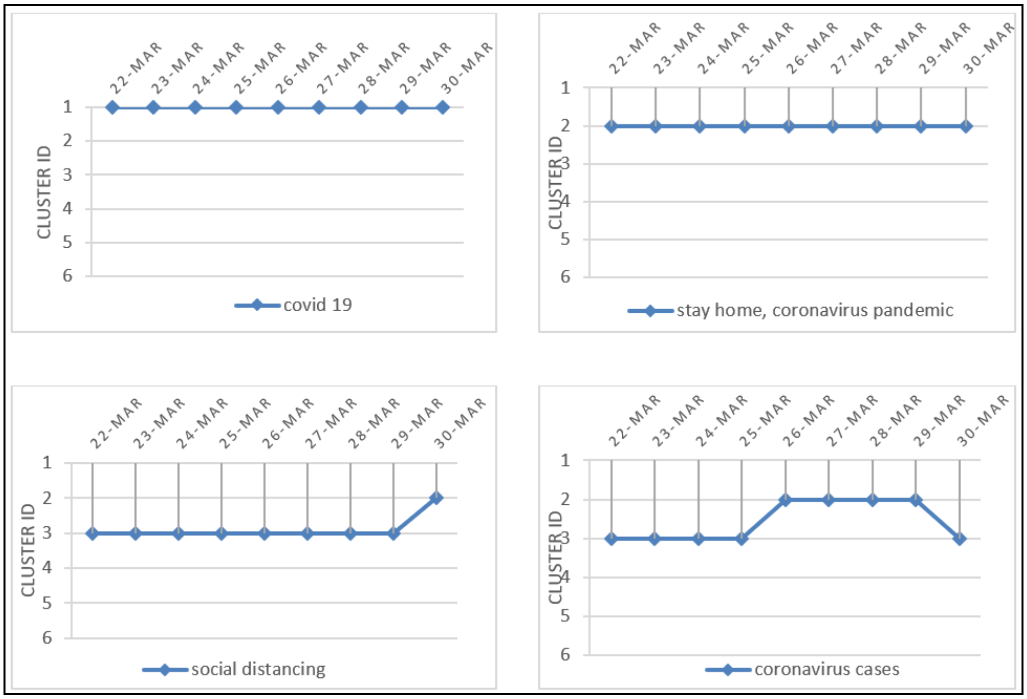

Figure 5: Evolution of bigram terms in clusters

Fig. 6 depicts the evolution of trigram terms in clusters for the aforementioned period. The "COVID-19 pandemic" trigram term was the most trended trigram term on Twitter starting from March 24th. The "COVID-19 cases" trigram term was in cluster 2, which means it got fewer tweets than "COVID-19 pandemic" during the first four days. However, it received more tweets for three consecutive days (March 26-28) and jumped up to cluster 1, so it has had the same importance as the "COVID-19 pandemic" term within the same duration (March 26-28). The trigram term "COVID-19 outbreak" has gained more importance in people's tweets. The results showed that this term was in cluster 2 for seven days during the whole period, which implies that it was one of the most trended trigram terms on Twitter for that period. Other trigram terms like "positive COVID-19" and "stay at home" were fluctuating between the different clusters over time. However, it is worth mentioning that the former was fluctuating in the upper clusters (1-3), while the latter was fluctuating in the lower clusters (3-5). Fig. 7 shows the sum of term frequencies in the clusters.

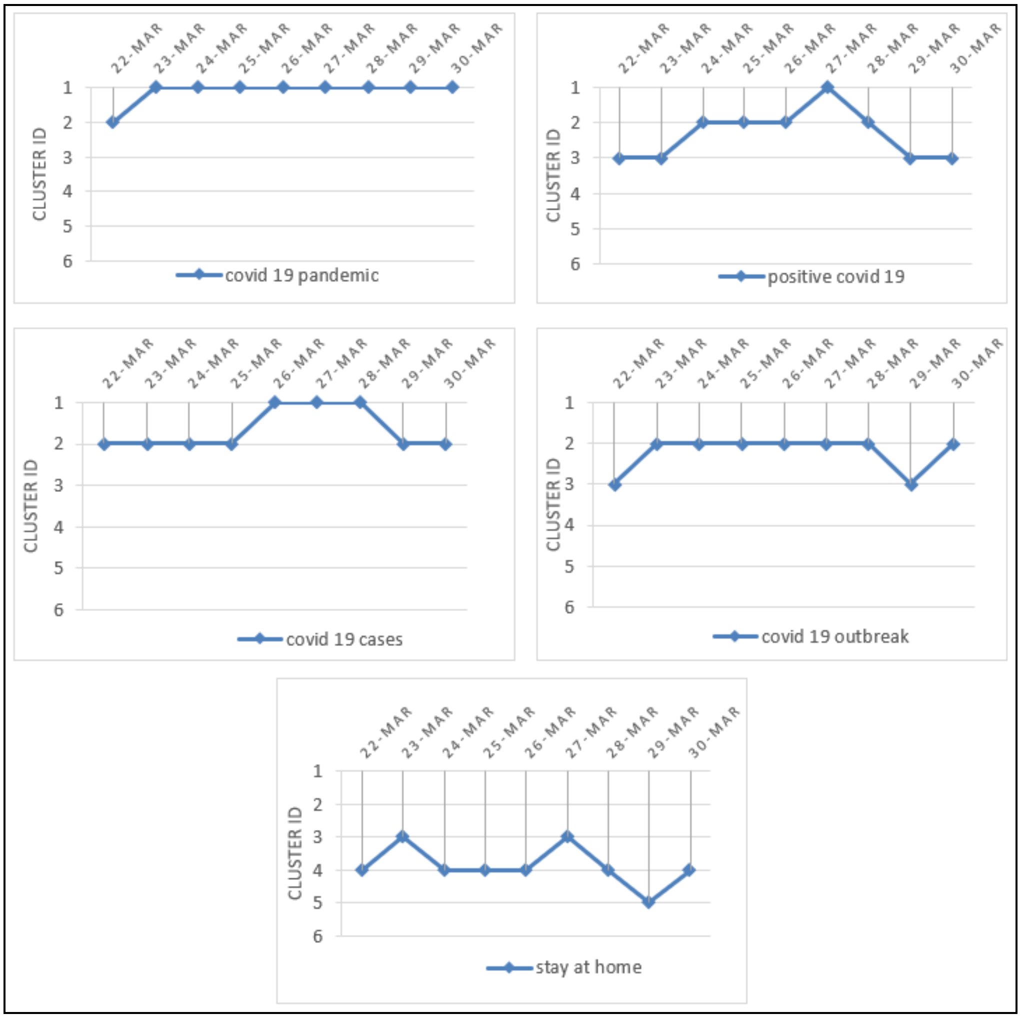

Figure 6: Evolution of trigram terms in clusters

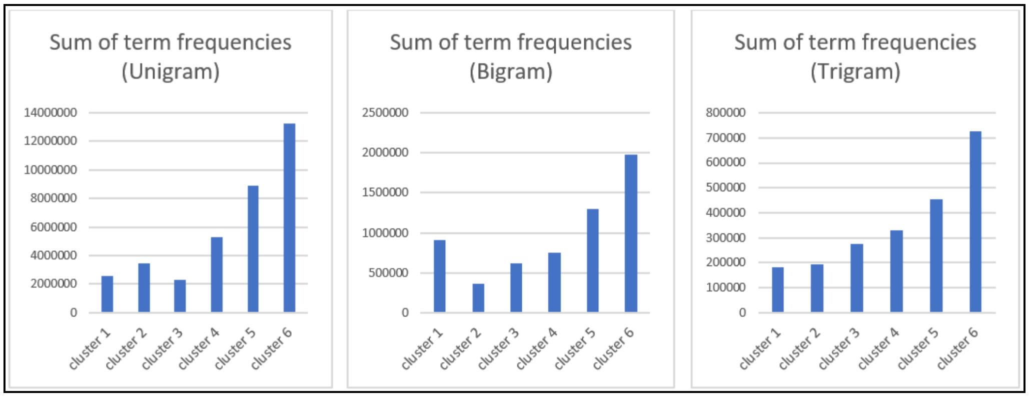

Figure 7: Sum of term frequencies in the clusters

5 Discussion and conclusion

This study conducted an evolutionary clustering analysis on Twitter in a time window during the COVID-19 outbreak. We have analyzed and identified the tweet patterns in three-level n-grams (n=1, 2, and 3), frequent occurrences of single words (e.g., COVID), bigrams (a combination of two words such as "COVID-19", "social distancing", "coronavirus pandemic") and trigram (a combination of three words such as "COVID-19 cases"). Also, we have observed in some experiments that the unigram was trending on Twitter up to 3 times and 70 times more than that of bigram and trigram terms, and it has the highest frequent terms and increases over time. On the other hand, clusters 3-5 are almost trending equally over time. As for the bigram analysis, we observed that the combination of the two words such "coronavirus cases" has had the same importance as of "stay home" and "coronavirus pandemic" bigram terms for the first four days of the study period; however, it became more important and got more tweets for the duration of the next three days. As for the trigram analysis, the terms like "positive COVID-19" and "stay at home" were fluctuating between the different clusters over time. However, it is worth mentioning that the former was fluctuating in the upper clusters, while the latter was fluctuating in the lower clusters.

The results showed that a large number of tweets about the COVID-19 were disseminated and received widespread public attention during the epidemic. The high-frequency words in Fig. 4 (i.e., death, test, spread, and lockdown) suggest that people fear of being infected, and those who got infection are afraid of death. People have a fear of spread, and consequently, they are afraid of a lockdown. The results shown in Fig. 5 revealed a consistency between "Stay home" and "Corona pandemic". The similar increasing level of these terms shows that people agreed to stay at home due to the fear of the spread. The "social distancing" term jumped on March 30th, which implied that people were calling for social distancing since they become aware of the COVID-19.

The prior findings suggested that social media posts may affect human psychology and behavior [Arpaci, Karataş and Baloğlu (2020)]. For example, social media had been used to direct people by spreading liberal postings during the events of the Arab Spring [White and Borgatti (1994)], the launch of WikiLeaks [Sifry (2011)], and the Gezi Park movement in Turkey [Chrona and Bee (2017)]. In the context of this study, false information has also been a genuine concern among social media platforms during the COVID-19 pandemic. For example, people linked 5G to the spread of COVID-19 by using the "#5G Coronavirus" hashtag as it was trending on Twitter, and that misinformation has led to the burning of 5G towers in the UK [Ahmed, Vidal-Alaball, Downing et al. (2020)]. On the other hand, the effective use of social media can shorten admission times by establishing factual communication channels, and thereby, helping patients in getting early attention during the lockdown [Huang, Xu, Cai et al. (2020)]. Therefore, we can conclude that the understanding of the social dynamics behind social media may help to design a more efficient communication strategy in the time of such a crisis. Further, this may help governments and health organizations to better understand the psychology of the public, and thereby, better communicate with them to prevent and manage the panic. Moreover, it can be useful for practitioners and researchers in developing crisis management information systems and social media-based emergency programs.

As a limitation, the present study aimed to identify the social media posts during a specific time frame in the pandemic. Therefore, a larger dataset with a longer time span could have more impact on drawing further conclusions. Further, employing a sentiment analysis of Twitter feeds could lead to more interesting results. Finally, it is imperative to mention that Twitter users are not representative of the general population, and it is important to emphasize that all conclusions about social behavior found in relevant studies may apply primarily to Twitter users and not necessarily to the general population.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

Ahmed, W.; Vidal-Alaball, J.; Downing, J.; Seguí, F. L. (2020): COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of Medical Internet Research*, vol. 22, no. 5, pp. 1-9.

Arpaci, I.; Karataş, K.; Baloğlu, M. (2020): The development and initial tests for the psychometric properties of the COVID-19 Phobia Scale (C19P-S). *Personality and Individual Differences*, vol. 164, pp. 1-6.

Banda, J. M.; Ramya, T. (2020): A Twitter dataset of 40+ million tweets related to COVID-19. https://zenodo.org/record/3723940.

Burnap, P.; Williams, M. L.; Sloan, L.; Rana, O.; Housley, W. et al. (2014): Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. *Social Network Analysis and Mining*, vol. 4, no. 1, pp. 1-14.

Chakrabarti, D.; Kumar, R.; Tomkins, A. (2006): Evolutionary clustering. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 554-560.

Chrona, S.; Bee, C. (2017): Right to public space and right to democracy: the role of social media in Gezi Park. *Research and Policy on Turkey*, vol. 2, no. 1, pp. 49-61.

Depoux, A.; Martin, S.; Karafillakis, E.; Preet, R.; Wilder-Smith, A. et al. (2020): The pandemic of social media panic travels faster than the COVID-19 outbreak. *Journal of Travel Medicine*, vol. 27, no. 3, pp. 1-2.

Duan, L.; Zhu, G. (2020): Psychological interventions for people affected by the COVID-19 epidemic. *The Lancet Psychiatry*, vol. 7, no. 4, pp. 300-302.

Gu, K.; Wang, L. Y.; Yin, B. (2019): Social community detection and message propagation scheme based on personal willingness in social network. *Soft Computing*, vol. 23, no. 15, pp. 6267-6285.

Househ, M. (2016): Communicating Ebola through social media and electronic news media outlets: a cross-sectional study. *Health Informatics Journal*, vol. 22, no. 3, pp. 470-478.

Huang, C.; Xu, X.; Cai, Y.; Ge, Q.; Zeng, G. et al. (2020): Mining the characteristics of COVID-19 patients in China: analysis of social media posts. *Journal of Medical*

Internet Research, vol. 22, no. 5, pp. 1-11.

Jiang, X.; Coffee, M.; Bari, A.; Wang, J.; Jiang, X. et al. (2020): Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials & Continua*, vol. 63, no. 1, pp. 537-551.

Kayes, A. S. M.; Islam, M. S.; Watters, P. A.; Ng, A.; Kayesh, H. (2020): Automated measurement of attitudes towards social distancing using social media: a COVID-19 case study.

Li, L.; Zhang, Q.; Wang, X.; Zhang, J.; Wang, T. et al. (2020): Characterizing the propagation of situational information in social media during COVID-19 epidemic: a case study on Weibo. *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 556-562.

Nayar, K. R.; Sadasivan, L.; Shaffi, M.; Vijayan, B.; Rao, A. P. (2020): Social media messages related to COVID-19: a content analysis.

Qin, L.; Sun, Q.; Wang, Y.; Wu, K. F.; Chen, M. et al. (2020): Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. *International Journal of Environmental Research and Public Health*, vol. 17, no. 7, pp. 1-14.

Sharot, T.; Sunstein, C. R. (2020): How people decide what they want to know. *Nature Human Behaviour*, vol. 4, pp. 14-19.

Sifry, M. L. (2011): *WikiLeaks and the Age of Transparency*. New York: OR Books.

Spencer, S. H. (2020): *False Claims of Nationwide Lockdown for COVID-19.* https://www.factcheck.org/2020/03/false-claims-of-nationwide-lockdown-for-covid-19/.

Tasnim, S.; Hossain, M. M.; Mazumder, H. (2020): Impact of rumors or misinformation on coronavirus disease (COVID-19) in social media. *Journal of Preventive Medicine & Public Health*.

Wang, C.; Cheng, Z.; Yue, X. G.; McAleer, M. (2020): Risk Management of COVID-19 by Universities in China. *Journal of Risk and Financial Management*, vol. 13, pp. 1-6.

White, D. R.; Borgatti, S. P. (1994): Betweenness centrality measures for directed graphs. *Social Networks*, vol. 16, no. 4, pp. 335-346.

Zarocostas, J. (2020): How to fight an infodemic. *The Lancet*, vol. 395, no. 10225, pp. 676-676.

Zhou, L. L.; Tan, F.; Yu, F.; Liu, W. (2019): Cluster synchronization of two-layer nonlinearly coupled multiplex networks with multi-links and time-delays. *Neurocomputing*, vol. 359, pp. 264-275.