COVID-19: What Are Arabic Tweeters Talking About?

Btool Hamoui[1](✉), Abdulaziz Alashaikh[2], and Eisa Alanazi[1]

[1] Center of Innovation and Development in Artificial Intelligence,
Umm Al-Qura University, Makkah, Saudi Arabia
s43680523@st.uqu.edu.sa, eaanazi@uqu.edu.sa

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

Abstract. The new coronavirus outbreak (COVID-19) has swept the world since December 2019 posing a global threat to all countries and communities on the planet. Information about the outbreak has been rapidly spreading on different social media platforms in unprecedented level. As it continues to spread in different countries, people tend to increasingly share information and stay up-to-date with the latest news. It is crucial to capture the discussions and conversations happening on social media to better understand human behavior during pandemics and alter possible strategies to combat the pandemic. In this work, we analyze the Arabic content of Twitter to capture the main discussed topics among Arabic users. We utilize Non-negative Matrix Factorization (NMF) to discover main issues and topics based on a dataset of Arabic tweets from early January to the end of April, and identify the most frequent unigrams, bigrams, and trigrams of the tweets. Eventually, the discovered topics are then presented and discussed which can be roughly classified into COVID-19 origin topics, prevention measures in different Arabic countries, prayers and supplications, news and reports, and finally topics related to preventing the spread of the disease such as curfew and quarantine. To our best knowledge, this is the first work addressing the issue of detecting COVID-19 related topics from Arabic tweets.

Keywords: COVID-19 · Twitter · Topic discovery · Arabic

1 Introduction

In recent years, social networks have become a remarkable source of information, reflecting societies interest and reactions about a specific topic. Analyzing the content and the diffusion of social networks information has been shown useful and increasingly used in many fields to characterize an event of interest, e.g., political, sports, or medical events. Lately, it was worthwhile to direct this capability

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

S. Chellappan et al. (Eds.): CSoNet 2020, LNCS 12575, pp. 425–436, 2020.
https://doi.org/10.1007/978-3-030-66046-8_35

toward the pandemic spread of corona virus. Consequently, an expedited research effort has been applied on analyzing social networks contents and activities during the pandemic spread to help recognize and characterize the social response [1].

In the meanwhile, with coronavirus infection spreading around the world, Arabic countries have been suffering from the outbreak of COVID-19 as the rest of the world. Nowadays, many individual's activities and conversations related to the pandemic are carried out through social media platforms such as Facebook, Twitter, Instagram, etc. Twitter is one of the most famous social media platforms that has a strong growth in the Arabic region, the number of posts reaches 17 million tweets per day according to the Arab social media report [2].

Due to its overwhelming usage and popularity, tweet content mining can potentially provide valuable information during health crises. Several studies have shown that Twitter can be exploited as a valuable data source for detecting and managing the outbreaks [3,4]. Recently, the rise of coronavirus cases in the Arabic countries has led to an escalating discussions related to the COVID-19 pandemic on social media platforms. Therefore, identifying the main concerns, thoughts, and topics regarding the coronavirus crises might be useful to assist public health professionals and social scientists. The main goal of this paper is employing text mining techniques to get an overview of the most discussed topics by Arabic tweeters during the pandemic. Particularly, we use Non-negative Matrix Factorization (NMF) to identify latent COVID-19 related topics in Arabic tweets.

2 Related Work

There has been a growing body of work aiming at mining content related to the COVID-19 pandemic. A study done by Alshaabi et al. [5] analyzed tweets in the context of COVID-19 by extracting 1,000 unigrams in 24 languages from tweets posted in early 2020 and compared it with the ones used a year ago. The authors observed that the first peak was around January 2020 and the second peak was in early March. Li et al. [6] performed an analysis (on Twitter and Weibo) by tracking the change of topic trends, sentiments, and emotions to understand the public attitude towards coronavirus crisis. The vast majority of the previous studies have been on English Twitter content. Recently, some studies (e.g., [7] and [8]) gave attention to analyzing Arabic Twitter content during the pandemic. A dataset of Arabic tweets, ArCOV-19, collected from January 27^{th}, 2020 to March 31^{st}, 2020 [7]. The sentiment analysis of Arabic tweeters in Saudi Arabia toward the preventive measures to combat COVID-19 were conducted in [8].

Various works have been done to extract topics from Twitter with varieties of algorithms. Prier et al. [9] explored tobacco-related tweets from health-related tweets by applying (LDA) algorithm. Besides, (LDA) algorithm employed by Sokolova et al. [10] to identify election-related events tweets. Alternatively, NMF can be employed to extract topics from text. It has been frequently used to analyze tweets' text. Geo-tagged tweets were analyzed by detecting trending topics

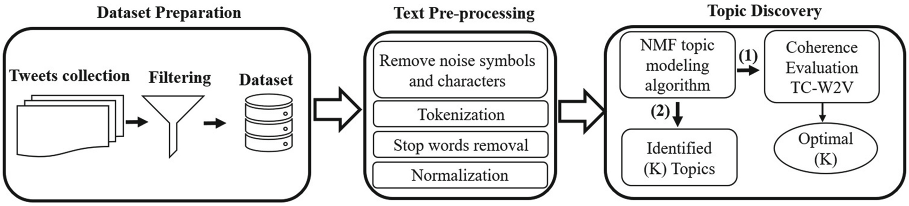

Fig. 1. Methodology workflow

using NMF for urban monitoring in certain areas in Indonesia [11]. Moreover, Klinczak and Kaestner [12] showed that the result of NMF on Twitter data yeilded better performance over the other two clustering algorithms, k-means, and k-medoids.

3 Methodology

This section describes the workflow of the methodology adapted for this study and explains the main steps. The workflow is depicted in Fig. 1 and is composed of the following steps: 1) Dataset preparation, 2) Text pre-processing, and 3) Topics discovery and themes identification, which involves: NMF for topic modelling, Topic model coherence evaluation employing word2vec, and Exploratory topic discovery.

3.1 Dataset Preparation

We use the dataset of the Arabic Twitter COVID-19 collection[1] [13], which contains 3,934,610 Arabic tweets related to COVID-19. The original dataset was collected through Twitter's streaming API and covers the time span from January 1, 2020 to April 30, 2020. Figure 2 illustrates the COVID-19 collected tweets frequency per day. To build a better-quality potential dataset for the experiment, certain filtration and cleaning are applied on the tweets collection to remove noise from the data:

- Filtering non-Arabic tweets: many tweets founded were multilingual tweets, since the Arab users may post tweets written in different languages besides Arabic. Therefore, we opted to filter out the multilingual tweets. The non-Arabic tweets were identified using the language field in the tweets metadata.
- Filtering out the retweets: the retweets were removed from the dataset to eliminate the duplicated content tweets.
- Filtering out short tweets: the tweets with one or two words usually could be ambiguous, hence, this will not provide meaningful information. Therefore, the tweets with less than three words were filtered out.

Applying the previous filtering steps, we ended up with 2,426,850 tweets.

[1] Available at: https://github.com/SarahAlqurashi/COVID-19-Arabic-Tweets-Dataset.

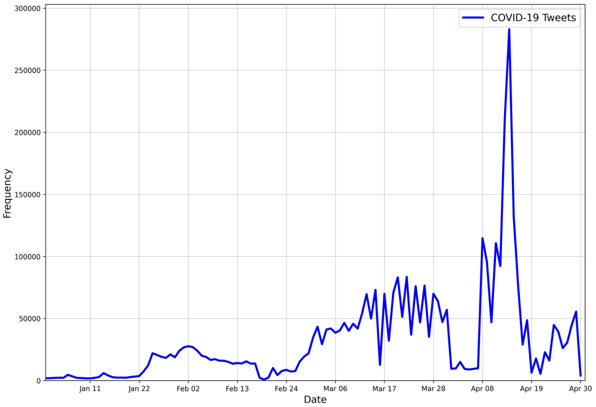

Fig. 2. COVID-19 Tweets Frequency per day

3.2 Text Pre-processing

The pre-processing involved applying several steps to the entire dataset with the aim of reducing the amount of trivial noise to clean the data. The following text pre-processing techniques were applied: First, the cleaning step is performed to remove noise data, in which we remove mentions, URLs links, emojis, punctuation, Non-alphabetic characters, and non-Arabic words. The hashtags were also removed yet maintaining its content in tweets. Furthermore, the Arabic vowel diacritics, 'Tashkeel' تشكيل: : 'Tashkeel [14] are diacritical marks appeared above or below each letter, used in the Arabic to affect the way of Arab pronunciation were removed; to unify the shape of tweeted words' format. Second, we performed tokenization on every tweet where each word was tokenized. Third, stop words removal was applied. For example the Arabic prepositions {من, الى, على, في, ...etc.}, along with other common words in Arabic that have no polarity significance in tweets were deleted. Lastly, we applied normalization to convert multiple forms of a letter into one uniform letter. To unify the form of 'Alef' and the form of 'Taa Marbotah', we replaced {أ, إ, آ} with {ا} and replace {ة} with {ه}. We also applied an extra normalization to the word virus, as it is pronounced in two different ways, as "Fairus" or "Firus". The word virus in Arabic was converted from فايروس to فيروس.

3.3 Topic Discovery and Themes Identifying

NMF for Topic Modelling. Non-negative Matrix Factorization (NMF) is an unsupervised technique for reducing the dimensionality of non-negative matrices [15]. It has been successfully applied in the field of text mining to identify topics [16]. Our study utilizes (NMF) according to its ability to give semantically meaningful results. A study done by O'callaghan et al. [17] found that

NMF produces more coherent topics than other popular topic modelling technique such as the latent Dirichlet al.location (LDA) model. To apply NMF, the pre-processed tweets are transformed to log-based Term Frequency-Inverse Document Frequency (TF-IDF) vectors, where each row corresponds to a term and each column to a document [18]. NMF based on (TF-IDF) values approves its usefulness since it can account for the importance of a word to a document within a collection of texts [17,19].

Topic Model Coherence Evaluation Employing Word2vec. According to the difficulty of defining the similarity measure in high-dimensional sparse vector space, we incorporated the potential of word embedding techniques to determine the number of topics. We opted to use Topic Coherence-Word2Vec (TC-W2V) metric, presented in [17], that measures the coherence between words assigned to a topic via Word2Vec. Word2Vec basically consists of a model to represent words as vectors. It is one of the most promising techniques in NLP that captures the meaning of the words [20].

We employed word2vec by training our model based on the 2,426,850 collected tweets using the Skipgram algorithm with a dimension of 200. To build the model[2], we used a small window of size 3 since the maximum length of a tweet is 280 characters, and we set the minimum word count to 10. The word vectors were produced using the Gensim package in Python. Given the trained word2vec model, we explored 11 words that have arisen and used frequently during COVID-19 pandemic such as {covid, mask, Wuhan, quarantine}, and visualize it's relevant words (top most 20 words similar) that have the same meaning, as shown in Fig. 3. As observed, Fig. 3 shows that the words {علاج, “treatment”}, {عقار, “drug”}, {ترياق, “antidote”}, {مصل, “serum”} are the words closest vector of “Vaccine” as an example. Figure 3 illustrates that the model was able to capture the similarity of the meaning of words.

After the word2vec model has been constructed, we trained the NMF model for different values of k, the number of topics. Then, we calculated the average TC-W2V for each model across all topics by extracting the similarity between all top-n words pairs in each topic from the word2vec model. The final NMF model trained with the highest average TC-W2V. As shown in Fig. 4, the highest average value was 0.3504 with $k = 11$. Hence, we trained the NMF model with the optimal number of topics using the scikitlearn implementation of NMF (including NNDSVD initialization) with k equal to 11.

4 Result and Analysis

4.1 Unigrams, Bigrams and Trigrams Frequency over Time Exploration

Basic unigrams, bigram and trigram frequency analysis over time will reflect the change of Arabic tweeters trends and concerns during the pandemic. After

[2] Available at: https://github.com/BatoolHamawi/COVID19Word2Vec.

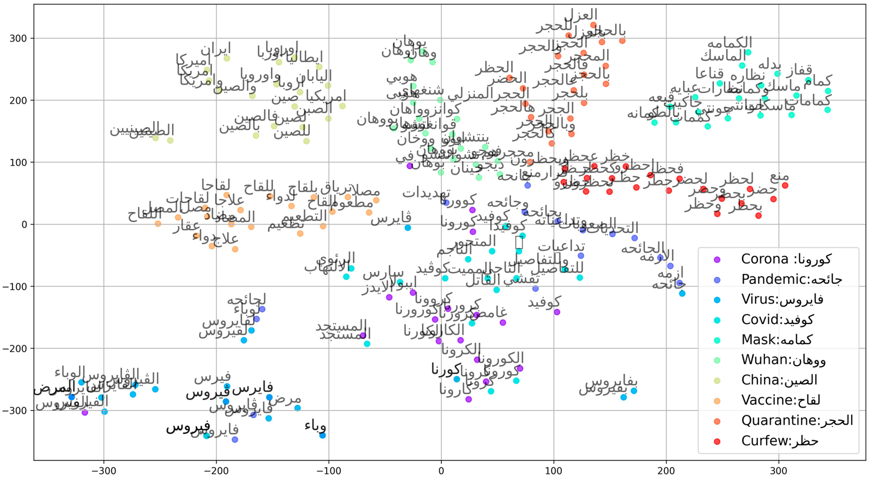

Fig. 3. t-SNE reduced visualization for 20 words closest to the chosen words that are related to the pandemic in the trained Word2vec Arabic Covid-19 model.

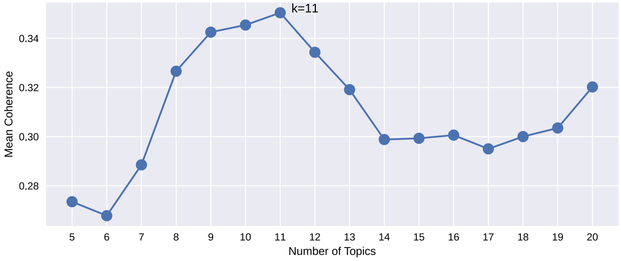

Fig. 4. Average TC-W2V for k from 5 to 20

applying the pre-processing steps, we constructed unigrams, bigrams and trigrams frequency table for the entire pre-proccessed dataset that resulted with 2,426,850 tweets. Then, we analyzed the frequency of each gram over the whole dataset, and explored the topmost unigrams, bigrams and trigrams over weeks. Figures 5 and 6 are the plots of grams frequency per day. In the plotting analysis, the series of grams counts smoothed by moving average to clearly presented the n-grams frequency over days and weeks. For unigram frequency, we investigated the volume of specific words appeared in the Arabic tweet content in January and associated with COVID-19 pandemic; كورونا، وباء، ووهان, which stand for "corona", "epidemic", and "Wuhan", respectively. Figure 5(A) plots the number of occurrences of these words. An increase is clearly noticed over the last two weeks of the January and reached the highest occurrences on the 25^{th} January.

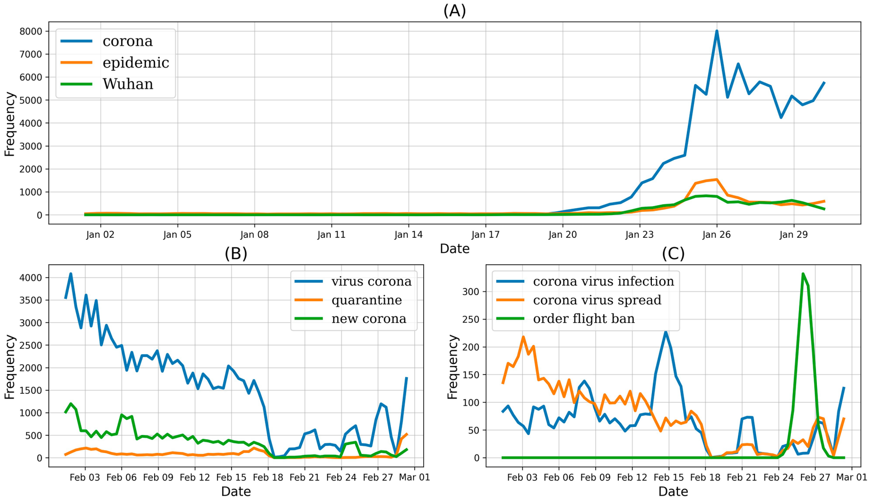

Fig. 5. COVID-19 n-grams frequency for January and February. (A) COVID-19 related words in January, (B) bi-grams frequency in February, and (C) trigrams frequency in February

Table 1. Top 10 bigrams and trigrams.

Bigram	Bigram (Ar)	Frequency	Trigram	Trigram (Ar)	Frequency
virus corona	فيروس كورونا	461443	new corona virus	فيروس كورونا مستجد	31068
home quarantine	حجر منزلي	136004	corona virus spread	انتشار فيروس كورونا	27324
curfew	حظر تجول	101077	corona virus new	فيروس كورونا جديد	23543
Ministry of health	وزارة صحة	60712	world health organization	منظمة صحة عالمية	22787
New corona	كورونا جديد	49870	virus corona outbreak	تفشي فيروس كورونا	18920
Corona epidemic	وباء كورونا	49252	home quarantine activity	فعاليات حجر منزلي	18651
new corona	كورونا مستجد	37985	new corona virus	جديدة فيروس كورونا	17148
Virus spread	انتشار فيروس	34908	new virus infection	اصابة فيروس كورونا	13222
World health	صحة عالمية	32951	facing virus corona	مواجهة فيروس كورونا	12188
Health quarantine	حجر صحي	32667	new virus infection	اصابة جديدة فيروس	10052

With respect to bigrams and trigrams, Table 1 presents the top 10 most co-occurrences bigrams and trigrams identified from the overall tweets in the pre-processed dataset. From the constructed bigrams and trigrams table, we manually crafted a list of bigrams and trigrams. Then, we tracked daily frequency for each of them by combining the identified bigrams and trigrams with its corresponding grams that have the same meaning. In February 2020, the news about coronavirus started to disseminate over Arabic countries. The bi-grams “corona virus”, فيروس كورونا, and “corona covid”, كورونا كوفيد , appeared mostly at the first week of the month, while the bi-gram “quarantine”, الحجر الصحي, started to increase over the last days in February as shown in Fig. 5(B). Similarly,

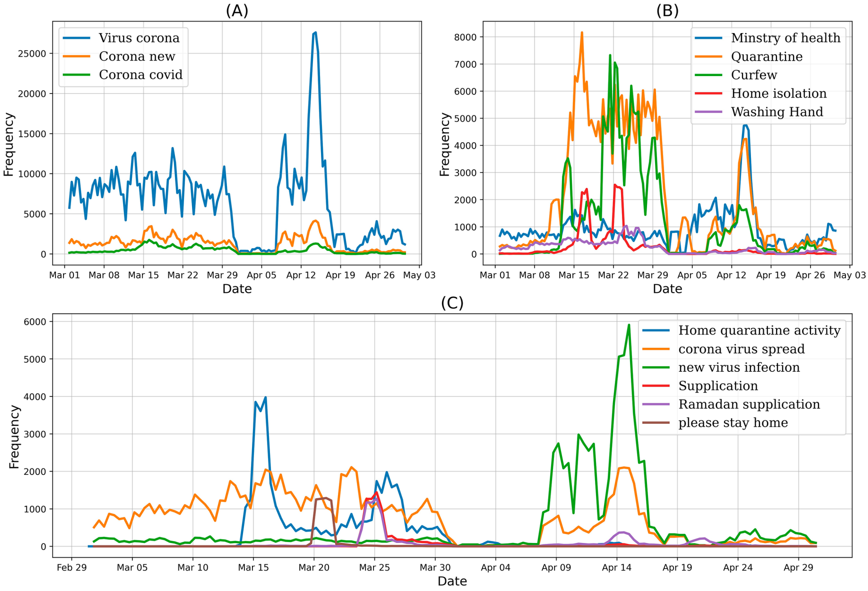

Fig. 6. COVID-19 top bi-grams and trigrams frequency for both March and April. (A) bi-grams related to coronavirus, (B) bi-grams of Health ministry, and preventive measures, and (C) the top trigrams frequency for both March and April

Fig. 5 (C) shows the evolution of the top three trigrams in February. The trigram "corona virus infection", اصابه بفيروس كورونا, had the highest occurrences in the second week. While corona virus spread, انتشار فيروس كورونا, started with the higher occurrences in the first week of February. We also noticed that the trigram order flight ban, نطالب بوقف الطيران, was the most frequent trigram at the end of February.

In March 2020, the number of infections with Corona virus was increasing rapidly in Arabic countries, and so the tweets about the virus. We tracked the bi-grams and trigrams for both March and April 2020 as done previously. The bigrams list was separated into two lists: bi-grams related to coronavirus, and bi-grams that include the "Ministry of Health" bi-gram and four bigrams about prevention measures as shown in Fig. 6 (A) and (B). In terms of bigrams frequency related to coronavirus, Fig. 6 (A) shows that there was stability in the pattern of bigrams in March comparing to April. Regarding the second list, the bi-grams quarantine, الحجر الصحي, and curfew, حظر التجول, they appeared as the topmost frequent bi-grams from the second week of March to the end of the fourth week. However, these bi-grams appeared during April albeit less frequently as shown in Fig. 6 (B). Moreover, the bigrams "washing hand", غسل اليدين, and "Home isolation", العزل المنزلي, were mostly used in March, while their frequency went down during April. The topmost frequent bigram in April was "Ministry of

Health”, وزارة الصحة, and it has higher frequency compared with March. In terms of trigrams frequency in March, the trigram “home quarantine activities”, فعاليات الحجر المنزلي, was the most frequent trigram in March as shown in Fig. 6 (C). Although this trigram was the sixth top frequent trigram in the entire dataset as listed in Table 1, it appeared only a few times over April. The trigram “corona virus spread “انتشار فيروس كورونا” was used more frequently in March than April. However, the trigram “corona virus infection”, اصابة بفيروس كورونا was the highest frequent trigram in April, and it appeared in higher occurrences comparing to March. The rest trigrams which include supplications “oh God, remove the affliction”, اللهم اكشف البلاء, “Allah, let our lives be extended so that we live to see the holy month of Ramadan”, اللهم بلغنا رمضان reached the highest in March, and continued to appear over April with lower frequency. Moreover, the trigram “please stay at home”, تكفون اقعدوا بيوتكم, appeared in March only.

Overall, the presented analysis demonstrates how the COVID-19 pandemic has dominated the Arab conversations over months with different phases; awareness phase, taking the action phase, and evaluation phase. In response to the news about the outbreak of coronavirus in China, some words related to coronavirus were mentioned in the posted tweets in the last ten days of January. The awareness phase about the virus continued to increase in February, the “coronavirus” and “corona COVID” co-occurrences expressed people awareness regarding the new virus alongside fears from virus spread by asking to stop the flight from and to China. During March, the lockdown measures were implemented in most Arabic countries [21]. The partial curfew was announced and imposed by authorities in different Arabic countries, such as Saudi Arabia, Kuwait, Jordan, and Egypt. Consequently, a shift towards discussing the precautionary measurements such as “quarantine”, “curfew”, and “home isolation” were observed and reached its peak in March. Besides, Arab tweeters encouraged each other to stay at home and increase prayers to God, which reflects the Arab attitude to take actions and combat the pandemic. Although, the lockdown measures implementation was extended to April, and some Arabic countries imposed the 24 h curfew [21], the discussion tendency in tweets content change to another phase. Notably, Arab users were more attentive to monitor and evaluate the number of infections, the situation of virus spread, and following up the impact of the precautionary measures with the Ministry of Health.

4.2 Exploratory Topic Discovery

We analyzed the 11 topics extracted from tweets using the NMF described earlier in Sect. 3.3. The distributions and the top-7 terms associated with each topic shown in Table 2. To provide an overview of the main discussed topics regarding the coronavirus in Arabic tweets, we inspected 1,000 chosen tweets from each topic along with top frequent bigrams and trigrams. Then, we manually analyzed with two Arabic native speakers volunteers the sets of the common bigrams, trigram, and overall 11,000 tweets. We observed the following:

- Topic 1: Preventive measures taken against the virus, staying at home, and protection from coronavirus infection. The most frequent countries mentioned in the tweets were Saudi Arabia, Egypt, Lebanon, China, Jordan and Oman.
- Topic 2: About quarantine, its impact on individuals, and quarantine activities. Moreover, appealing to increased charitable donations.
- Topic 3: Corona is a global epidemic, suspension school, and the coronavirus epidemic.
- Topic 4: About China, flight cancellations from and to China, and discussion about spreading the virus in Wuhan city.
- Topic 5: About curfew- tweeters mostly mentioned Kuwait, Saudi Arabia, and Jordan in the tweets. Moreover, appeals to stay at home mostly written in Gulf dialect such as “Please stay home!”, “تكفون اقعدو بيوتكم”.
- Topic 6: It is mainly about coronavirus spread in Egypt. Most of the tweets were written in Egyptian local dialect.
- Topic 7: Supplications. Asking God for relief and protection from illnesses, such as may Allah save us, and protect Muslims. Examples of trigrams founded: “حمانا الله واياكم” and, “حفظ الله الجميع”.
- Topic 8: About the latest News. The tweets that belonged to this topic mainly showed statistics and, number of cases, the number of new cases every day, and the number of deaths caused by coronavirus in different cities and countries.
- Topic 9: Ramadan Supplications, such as “اللهم بلغنا رمضان,” which mean O’ Allah, let our lives be extended so that we live to see the holy month of Ramadan.
- Topic 10: The main topics founded are about: facing the spread of coronavirus, and corona out-breaks.

Table 2. Identified topics and their components

Topic	Topics identified	Keywords	Distribution
1	Prevention measures in different countries	السعوديه ، الكويت ، الاردن ، لبنان ، فيروس ، كورونا ، مصر Saudi Arabia, Kuwait, Jordan, Lebanon, virus, corona, Egypt	17.95%
2	Quarantine	حجر ، صحي ، ، منزلي ، عزل ، واجب ، بيت ، خليك quarantine, healthy, house, isolation, must, home, stay	6.76%
3	Corona is a global pandemic	وباء ، عالم ، عالمي ، اخطر ، دول ، مرض ، ناس epidemic, global, globally, the most dangerous, countries, disease, people	5.37%
4	China	الصين ، ووهان ، وفيات ، ارتفاع ، صينيه ، امريكا ، عالم China, Wuhan, deaths, higher, Chinese, America, world	15.17%
5	Curfew	قرار، الكويت، السعوديه ، حظر، تجول ، اجباري، حضر Curfew, wandering, compulsory, lockdown, decision, Kuwait, Saudi Arabia	4.55%
6	Coronavirus in Egypt	مصر ، كورونا، زمن، عشان، اخطر ، خايف ، علاج Egypt, corona, time, because, danger, afraid, treatment	9.31%
7	Supplications	الله ، يكفينا ، مسلمين ، نسال ، شاء ، حسبي ، كارونا Allah, away from, Muslims, we ask, will, suffices, corona	13.49%
8	Latest News	تسجيل ، ارتفاع ، حاله ، اصابه ، جديده ، وفاه ، تعلن Record, increase, case, infections, new, death, announce.	5.46%
9	Ramadan Supplications	اللهم ، رمضان ، شعبان ، يارب ، بلغنا ، مسلمين ، اسقام O Allah, Ramadan, Shaaban, O Lord, we have reached, ailments	3.76%
10	Coronavirus spread	تفشي ، مواجهة ، وقاية، فيروس، كورونا، مستجد، انتشار Outbreak, confrontation, prevention, virus, corona, novel, spread.	10.89%
11	Ministry of Health announcements	صحة ، وزارة ، منظمة ، تعلن ، حالات ، وزير ، عالمية Health, ministry, organization, announce, cases, minister	7.24%

- Topic 11: About the World Health Organization, Ministry of Health announcements in different countries, and health care workers on the frontline (health heroes).

5 Conclusion

This paper presents a preliminary analysis and topic extraction of Arabic tweets posted during COVID-19 pandemic from January to April 2020. An analysis of the topmost frequent bi-grams and trigrams showed change in topic over time. The topics were extracted utilizing the Non-negative Matrix Factorization (NMF) methods. Our results demonstrate the power of NMF in detecting meaningful topics that we believe will give great insights to the current discussions and conversations happening on Arabic Twitter. In the near future, we plan to consider the sentiment of the Arabic users to the current pandemic using deep learning techniques.

References

1. Culotta, A.: Towards detecting influenza epidemics by analyzing Twitter messages. In: Proceedings of the First Workshop on Social Media Analytics, pp. 115–122 (2010)
2. Mourtada, R., Salem, F.: Citizen engagement and public services in the Arab world: the potential of social media. In: Arab Social Media Report Series, 6th edn, June 2014
3. de Quincey, E., Kostkova, P.: Early warning and outbreak detection using social networking websites: the potential of Twitter. In: Kostkova, P. (ed.) eHealth 2009. LNICST, vol. 27, pp. 21–24. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-11745-9_4
4. Morin, C., Bost, I., Mercier, A., Dozon, J.-P., Atlani-Duault, L.: Information circulation in times of Ebola: Twitter and the sexual transmission of Ebola by survivors. PLoS Currents **10** (2018)
5. Alshaabi, T., et al.: How the world's collective attention is being paid to a pandemic: Covid-19 related 1-gram time series for 24 languages on Twitter. arXiv preprint arXiv:2003.12614 (2020)
6. Li, X., Zhou, M., Wu, J., Yuan, A., Wu, F., Li, J.: Analyzing COVID-19 on online social media: trends, sentiments and emotions. arXiv preprint arXiv:2005.14464 (2020)
7. Haouari, F., Hasanain, M., Suwaileh, R., Elsayed, T.; ARCOV-19: the first Arabic COVID-19 Twitter dataset with propagation networks. arXiv, arXiv-2004 (2020)
8. Alhajji, M., Al Khalifah, A., Aljubran, M., Alkhalifah, M.: Sentiment analysis of tweets in Saudi Arabia regarding governmental preventive measures to contain COVID-19 (2020)
9. Prier, K.W., Smith, M.S., Giraud-Carrier, C., Hanson, C.L.: Identifying health-related topics on Twitter. In: Salerno, J., Yang, S.J., Nau, D., Chai, S.-K. (eds.) SBP 2011. LNCS, vol. 6589, pp. 18–25. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-19656-0_4
10. Sokolova, M., et al.: Topic modelling and event identification from Twitter textual data. arXiv preprint arXiv:1608.02519 (2016)

11. Sitorus, A.P., Murfi, H., Nurrohmah, S., Akbar, A.: Sensing trending topics in twitter for greater Jakarta area. Int. J. Electr. Comput. Eng. **7**(1), 330 (2017)
12. Klinczak, M.N., Kaestner, C.A.: A study on topics identification on Twitter using clustering algorithms. In: 2015 Latin America Congress on Computational Intelligence (LA-CCI), pp. 1–6. IEEE (2015)
13. Alqurashi, S., Alhindi, A., Alanazi, E.: Large Arabic Twitter dataset on COVID-19. arXiv preprint arXiv:2004.04315 (2020)
14. Zerrouki, T., Balla, A.: Tashkeela: novel corpus of Arabic vocalized texts, data for auto-diacritization systems. Data Brief **11**, 147 (2017)
15. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
16. Kuang, D., Yun, S., Park, H.: SYMNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. J. Global Optim. **62**(3), 545–574 (2015)
17. O'callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Syst. Appl. **42**(13), 5645–5657 (2015)
18. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. **24**(5), 513–523 (1988)
19. Greene, D., Cross, J.P.: Exploring the political agenda of the European parliament using a dynamic topic modeling approach. arXiv preprint arXiv:1607.03055 (2016)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
21. Abbas, N.: These Arab Countries Are Now In Lockdown, 2020. https://www.forbesmiddleeast.com/industry/healthcare/in-numbers-the-global-ventilator-shortage. Accessed 20 Aug 2020