 doi:10.20944/preprints202007.0172.v1

WHAT ARE COVID-19 ARABIC TWEETERS TALKING ABOUT?

A PREPRINT

Btool Hamoui*
Center of Innovation and Development in Artificial Intelligence
Umm Al-Qura University
Makkah, Saudi Arabia
s43680523@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
eaanazi@uqu.edu.sa

July 8, 2020

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. The final 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 for 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,

*Corresponding author

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

it was worthwhile to direct this capability 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 data for detecting the outbreaks of a pandemic such as the case with H1N1 virus [3], and the out-breaks of influenza [4]. Moreover, it plays an important role in understanding the public behaviors and impressions towards health crises, by analysing the tweet text to identify the main concerns about the Zika virus [5], and filtering the topics related to Ebola [6].

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. It will provide an instantaneous snapshot of the Arabic social opinions and behavioural responses to understand issues more properly. To this aim, getting an overview of the most discussed topics leveraging Arabic content tweets posted by Arabic tweeters by employing text mining techniques is the main goal of this paper. This study presents the first step toward extracting the main topics discussed by Arabic Twitter users regarding the COVID-19 pandemic. This paper used Non-negative Matrix Factorization (NMF), a topic modelling method, to identify latent topics.

2 Methodology

In this section, we describe the workflow of the methodology we adapted for this study, and explain the main steps we are following in detail. The workflow is depicted in Figure 1 and is composed of the following steps:

- Dataset preparation.
- Text pre-processing.
- Topics discovery and themes identification:
  - NMF for topic modelling.
  - Topic model coherence evaluation employing word2vec.
  - Exploratory topic discovery.

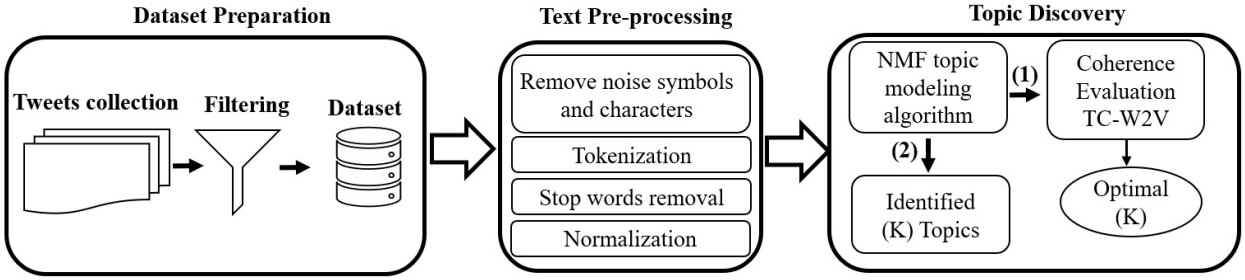

Figure 1: Methodology Workflow

2.1 Dataset preparation

We use the dataset of the Arabic Twitter COVID-19 collection[2] [7], 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. 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:

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

- 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 [8]. The non-Arabic tweets identified using the language field in the tweets metadata [9].
- 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.

2.2 Text pre-processing

To prepare the tweets for text mining, it is necessary to represent the tweets in more appropriate form that can be analyzed. 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:

- Noise symbols and characters removal: We processed the dataset first by deleting all mentions and URLs links from the tweet. Then, deleting all emojis, Arabic and English punctuation, all numbers, and all non-alphabet characters. We also removed non-Arabic letters, to keep only the Arabic alphabets. Finally, we removed leading and trailing spaces in addition to line breaks.
- Removing the Arabic vowel diacritics, 'Tashkeel' تشكيل: 'Tashkeel' [10] are diacritical marks appeared above or below each letter, used to affect the way of Arab pronunciation in accordance to some syntax and grammatical rules. Hence, the words with diacritical marks result in different shapes for words of the same origin. To unify the shape of similar words formats, we removed it from the tweets.
- Tokenization: we tokenized each word in tweets, which resulted in a group of raw tokens as an array of strings.
- Stopwords removal: the Arabic stop words are certain common words such as من, الى, على, في, ...etc.. Stop words that do not influence the topic modeling were removed from tweets.
- Normalization: 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}. In Arabic language, we have two ways to spell the word virus, which pronounced as "Fairus" or "Firus". Therefore, we also applied an extra normalization to unify virus word in Arabic, we converted from فايروس to فيروس.

2.3 Topic discovery and themes identifying

2.3.1 NMF for topic modelling

Non-negative Matrix Factorization (NMF) is an unsupervised technique for reducing the dimensionality of non-negative matrices [11]. It has been successfully applied in the field of text mining to identify topics [12, 13]. Our study utilized (NMF) according to its ability to give semantically meaningful results. A study done by O'callaghan et al. [14] founded that NMF produced more coherent topics than other popular topic modelling technique such as the latent Dirichlet allocation (LDA) model. To apply NMF, the pre-processed tweets were 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 [15]. 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 [14, 16].

2.3.2 Topic model coherence evaluation employing word2vec

According to the difficulty of defining the similarity measure in high-dimensional sparse vector space, we incorporate the potential of word embedding techniques to determine the number of topics. We opted to use the presented measure, Topic Coherence-Word2Vec (TC-W2V) metric in [14] 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 [17]. We employed word2vec by training our model based on the 2,426,850 tweets using the Skipgram algorithm with a dimension of 50. The word vectors were produced using the Gensim package in Python.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

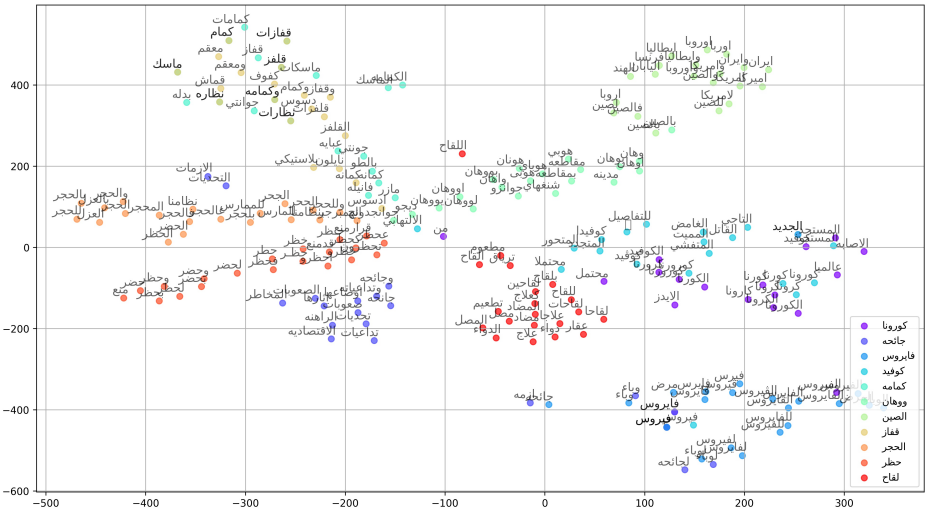

Figure 2: Word2vec Arabic Covid-19 model

Given the trained word2vec model, we explored the top 20 Arabic words related to COVID-19 pandemic illustrated in Figure 2. By observation, the model was able to capture the similarity meaning of words such as:

- كمامة : ماسك، كمامات
- حظر: منع ، حضر
- قفاز: قلفز، قفازات
- ووهان: اووهان، يووهان
- لقاح: مصل، عقار
- جائحة : ازمة ، اجتياح

After the word2vec model constructed, we trained the NMF model for different values of k, the number of topics. Then, we calculate 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 Figure 3, the highest average value was 0.3504 with k=11. Based on the result obtained from (TC-W2V) metric, 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.

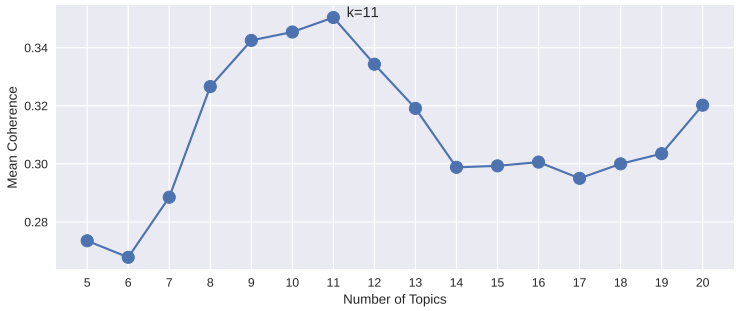

Figure 3: Average TC-W2V for k from 5 to 20

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

3 RESULTS & DISCUSSION/ANALYSIS

3.1 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 applying the pre-processing steps, we constructed unigrams, bigrams and trigrams frequency table for the entire pre-proccessed dataset. Then, we analyzed the frequency of each gram over the whole dataset, and explored the topmost unigrams, bigrams and trigrams over weeks.

For unigram frequency, we generated the word cloud of the top 10,000 unigrams in the dataset. Figure 4 illustrated that the word "كورونا" which translates to "corona" was the most frequent word. Also 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 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 25th of January to reach 13,938, 3,717, and 2,013 for corona, epidemic, and Wuhan, respectively.

Figure 4: Unigram word cloud

With respect to bigrams and trigrams, Table 1 shows the top 10 bigrams and trigrams from the dataset. From the constructed bigrams and trigrams table, we created a list of bigrams and trigrams. Then, we tracked each of them by combining each bigrams and trigrams with its corresponding grams that have the same meaning. The day with highest frequency for each month represented in Table 2. The month that has zero or very low frequency counts for a bigram, or a trigram, was omitted from the table.

In February 2020, the news about coronavirus started to disseminate over Arabic countries. The bigrams "corona virus", فيروس كورونا, and "corona covid", كورونا كوفيد, appeared mostly at the first week of the month, while the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

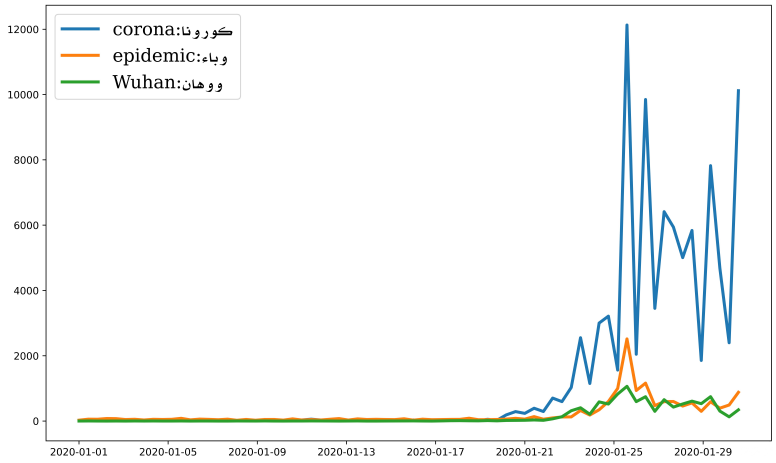

Figure 5: COVID-19 related words frequency in January

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	وزاره صحه	60712	world health organization	منظمه صحه عالميه	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	جديده فيروس كورونا	17148
Virus spread	انتشار فيروس	34908	new virus infection	اصابه فيروس كورونا	13222
World health	صحه عالميه	32951	facing virus corona	مواجهه فيروس كورونا	12188
Health quarantine	حجر صحي	32667	new virus infection	اصابه جديده فيروس	10052

Table 1: Top 10 bigrams and trigrams.

bigram “quarantine”, الحجر الصحي, started to increase over the last days in February as shown in Figure 6. Similarly, Figure 7 shows the evolution of the top three trigrams, the trigram “corona virus infection”, اصابه بفيروس كورونا, had the highest occurrences in the second week. 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 track the bigrams and trigrams for both March and April 2020 as done previously.

The bigrams list was separated into two lists: bigrams related to coronavirus, and bigrams included the Health ministry bigram and four bigrams about prevention measures as shown in Figure 8 and, Figure 9. In terms of bigrams frequency related to coronavirus, Figure 8 showed that in March there was stability in the pattern of bigrams in comparing with April.

Regarding the second list, the bigrams quarantine, الحجر الصحي , and curfew, حظر التجول, appeared as the topmost frequent bigrams from the second week of March to the end of the fourth week. However, these bigrams were used during April albeit less frequently as shown in Figure 9. Moreover, the bigrams “washing hand”, غسل اليدين, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

English bigram	Lists of bigrams and related	Day of highest frequency	English trigram	Lists of trigrams and related	Day of highest frequency
Virus corona	فيروس كورونا	01/02/2020 13/03/2020 13/04/2020	Corona virus infection	اصابه فيروس كورونا، اصابه جديده فيروس	14/02/2020 10/03/2020 14/04/2020
New virus corona	كورونا مستجد، كورونا الجديد	01/02/2020 20/03/2020 14/04/2020	Corona virus spread	انتشار فيروس كورونا، تفشي فيروس كورونا	03/02/2020 22/03/2020 14/04/2020
Corona covid	كورونا كوفيد	16/03/2020 14/04/2020	order flight ban	نطالب وقف طيران	26/02/2020
Quarantine	حجر منزلي، حجر الصحي	29/02/2020 16/03/2020 14/04/2020	Home quarantine activity	فعاليات حجر منزلي	15/03/2020 03/04/2020
Home isolation	عزل منزلي	16/03/2020	Supplication	اللهم اكشف بلاء	24/03/2020 13/04/2020
Ministry of health	وزاره صحه	14/04/2020	Ramadan Supplication	اللهم بلغنا رمضان	24/03/2020 14/04/2020
Curfew	حظر تجول، حضر تجول	20/03/2020 14/04/2020	Please stay home	تكفون اقعدو بيوتكم	20/03/2020
Hand washing	غسل اليدين	26/03/2020			

Table 2: The list of bigrams and trigrams with day of highest frequency.

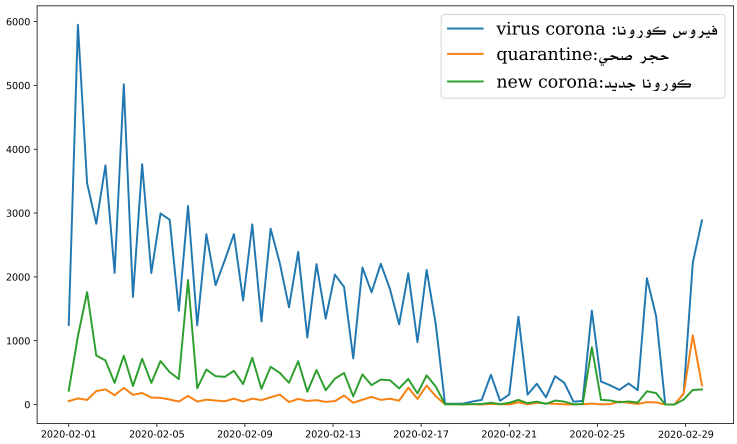

Figure 6: COVID-19 top bigrams frequency in February

"Home isolation", العزل المنزلي, were mostly used in March, while its frequency going down to have very low frequency during April. The topmost frequent bigram in April was "Ministry of Health", وزاره الصحه, and it has a 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. 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

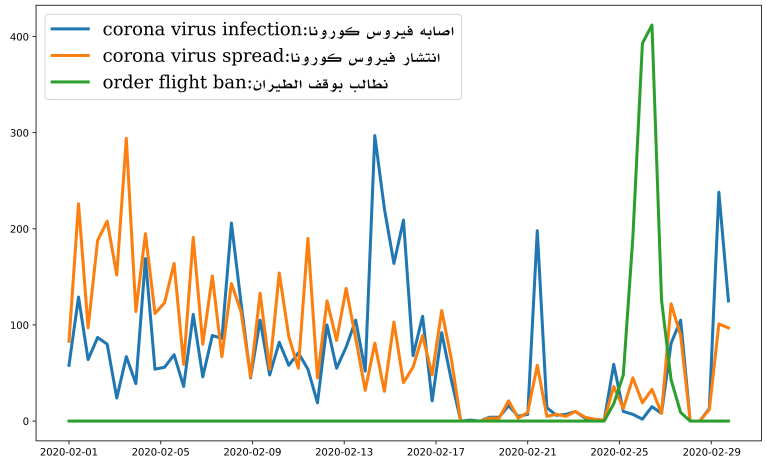

Figure 7: COVID-19 top trigrams frequency in February

more frequently in March than April. However, the trigram "corona virus infection", اصابه بفيروس كورونا was the highest frequent trigram in April, and it appeared in higher occurrences comparing with 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, was appeared in March only. Figure 10 showed the top trigrams frequency for both March and April.

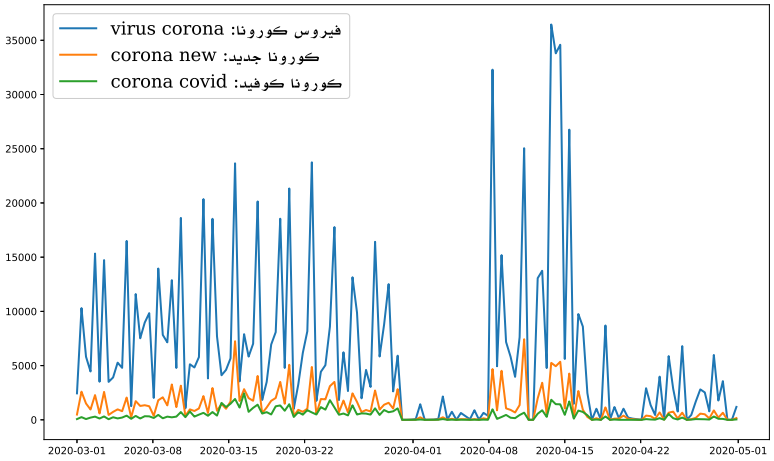

Figure 8: Top bigrams related to coronavirus terms frequency in March and April

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 July 2020 doi:10.20944/preprints202007.0172.v1

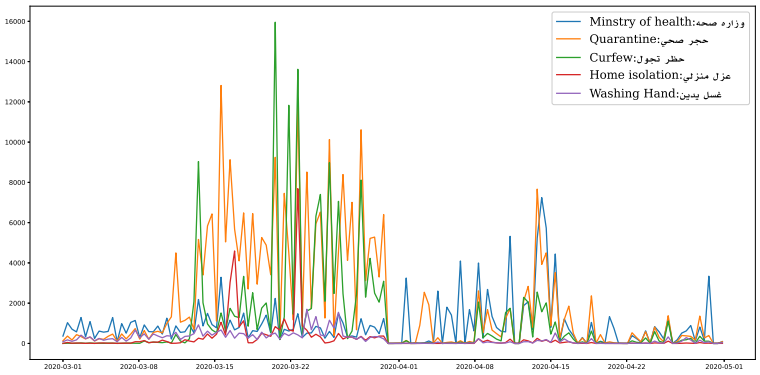

Figure 9: Top bigrams include Ministry of health and prevention measures frequency in March and April

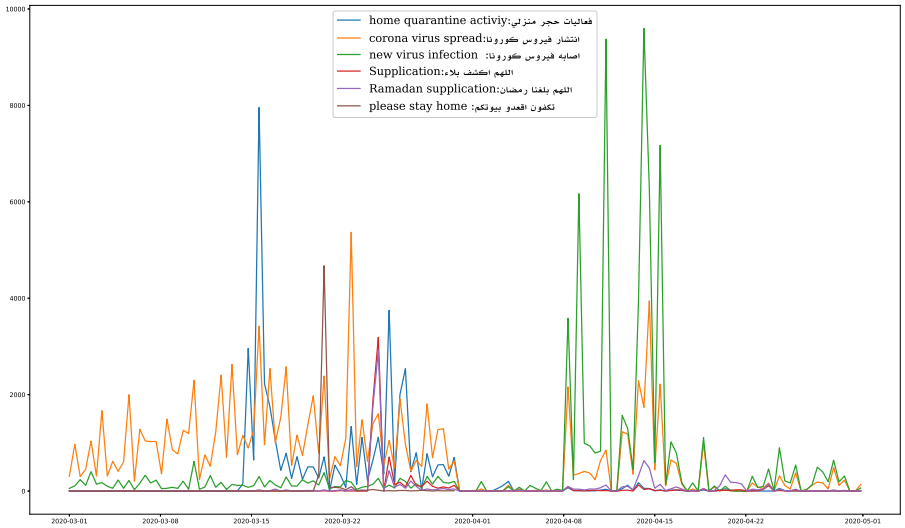

Figure 10: COVID-19 top trigrams frequency in March and April

3.2 Exploratory Topic Discovery

We analyzed the 11 topics extracted from tweets using the NMF described earlier in Section 2.3. The distributions and the top-7 terms associated with each topic are shown in Table 3. To provide an overview of the main discussed topics regarding the coronavirus in Arabic tweets, we inspected a few chosen tweets from each topic along with top bigrams and trigrams, and we observed the following:

- Topic 1: Prevention 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 increase charitable donations.

- Topic 3: Corona is a global epidemic, Stopping schools, 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 mention Kuwait, Saudi Arabia, and Jordan countries in the tweets. Moreover, appeals to sit at home mostly written in Gulf dialect such as please stay home, " تكفون اقعدو بيوتكم" .
- Topic 6: Mainly about coronavirus spreading in Egypt. The most tweets were written in Egyptian dialectal words.
- Topic 7: Supplications, 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 a 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.
- Topic 11: About the World Health Organization, Ministry of Health announcements in different countries, and Health care workers on the front-line (health heroes).

Topic	Topics identified	Keywords	Distribution
1	Prevention measures in different countries	السعوديه ، الكويت ، الاردن ، لبنان ، فيروس ، كورونا ، مصر	17.95 %
2	Quarantine	حجر ، صحي ، ، منزلي ، عزل ، واجب ، بيت ، خليك	6.76 %
3	Corona is a global pandemic	وباء ، عالم ، عالمي ، اخطر ، دول ، مرض ، ناس	5.37 %
4	China	الصين ، ووهان ، وفيات ، ارتفاع ، صينيه ، امريكا ، عالم	15.17 %
5	Curfew	قرار، الكويت، السعوديه ، حظر، تجول ، اجباري، حضر	4.55 %
6	Coronavirus in Egypt	مصر ،كورونا، زمن، عشان، اخطر ، خايف ، علاج	9.31 %
7	Supplications	الله ، يكفينا ، مسلمين ، نسال ، شاء ، حسبي ، كارونا	13.49 %
8	Latest News	تسجيل ، ارتفاع ، حاله ، اصابه ، جديده ، وفاه ، تعلن	5.46 %
9	Ramadan Supplications	اللهم ، رمضان ، شعبان ، يارب ، بلغنا ، مسلمين ، اسقام	3.76 %
10	Coronavirus spread	تفشي ، مواجهه ، وقايه ،فيروس، كورونا، مستجد، انتشار	10.89 %
11	Ministry of Health announcements	صحه ، وزاره ، منظمه ، تعلن ، حالات ، وزير ، عالميه	7.24 %

Table 3: Identified topics and their components

In summary, the identified topics can be grouped into five groups of themes:

1. Epidemic and Pandemic (31.43%): involves three topics (3, 4, 10); Corona is a global pandemic, China, and Coronavirus spread.
2. Specific Country Related Cases and Discussions (27.26 %): this theme includes topics (1, 6); the prevention measurements in different Arabic countries taken against coronavirus and the spreading of coronavirus in Egypt.
3. Prayers (17.25%): this theme is generated from the two supplications topics (7, 9).
4. News and Reports (12.70%): this theme contains two topics (8, 11); latest news and ministry of health announcements.
5. Methods for Decreasing the Spread of coronavirus (11.31 %): which include Curfew and Quarantine topics (2, 5).

4 Conclusions

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.

Acknowledgements

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

References

[1] S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller, and et al., "Leveraging data science to combat covid-19: A comprehensive review," Apr 2020. [Online]. Available: https://www.techrxiv.org/articles/Leveraging_Data_Science_To_Combat_COVID-19_A_Comprehensive_Review/12212516/1

[2] R. Mourtada and F. Salem, "Citizen engagement and public services in the arab world: The potential of social media," Arab Social Media Report series, 6th edition, June, 2014.

[3] E. De Quincey and P. Kostkova, "Early warning and outbreak detection using social networking websites: The potential of twitter," in International Conference on Electronic Healthcare. Springer, 2009, pp. 21–24.

[4] A. Culotta, "Towards detecting influenza epidemics by analyzing twitter messages," in Proceedings of the first workshop on social media analytics, 2010, pp. 115–122.

[5] E. M. Glowacki, A. J. Lazard, G. B. Wilcox, M. Mackert, and J. M. Bernhardt, "Identifying the public's concerns and the centers for disease control and prevention's reactions during a health crisis: An analysis of a zika live twitter chat," American journal of infection control, vol. 44, no. 12, pp. 1709–1711, 2016.

[6] C. Morin, I. Bost, A. Mercier, J.-P. Dozon, and L. Atlani-Duault, "Information circulation in times of ebola: Twitter and the sexual transmission of ebola by survivors," PLoS currents, vol. 10, 2018.

[7] S. Alqurashi, A. Alhindi, and E. Alanazi, "Large arabic twitter dataset on covid-19," arXiv preprint arXiv:2004.04315, 2020.

[8] I. Alsarsour, E. Mohamed, R. Suwaileh, and T. Elsayed, "Dart: A large dataset of dialectal arabic tweets," in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 2018.

[9] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "Arasenti-tweet: A corpus for arabic sentiment analysis of saudi tweets," Procedia Computer Science, vol. 117, pp. 63–72, 2017.

[10] T. Zerrouki and A. Balla, "Tashkeela: Novel corpus of arabic vocalized texts, data for auto-diacritization systems," Data in brief, vol. 11, p. 147, 2017.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[12] S. Arora, R. Ge, and A. Moitra, "Learning topic models–going beyond svd," in 2012 IEEE 53rd annual symposium on foundations of computer science. IEEE, 2012, pp. 1–10.

[13] D. Kuang, S. Yun, and H. Park, "Symnmf: nonnegative low-rank approximation of a similarity matrix for graph clustering," Journal of Global Optimization, vol. 62, no. 3, pp. 545–574, 2015.

[14] D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," Expert Systems with Applications, vol. 42, no. 13, pp. 5645–5657, 2015.

[15] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information processing & management, vol. 24, no. 5, pp. 513–523, 1988.

[16] D. Greene and J. P. Cross, "Exploring the political agenda of the european parliament using a dynamic topic modeling approach," arXiv preprint arXiv:1607.03055, 2016.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.