Navigating Healthcare Through Challenging Times
D. Hayn et al. (Eds.)

doi:10.3233/SHTI210084

Topic Discovery on Farsi, English, French, and Arabic Tweets Related to COVID-19 Using Text Mining Techniques

Hamoon JAFARIAN[a,1], Mahin MOHAMMADI[b] , Alireza JAVAHERI[c], Makram SUKARIEH[d], Mohsen YOOSEFI NEJAD[e], Abbas SHEIKHTAHERI[f], Mehdi HOSSEINZADEH[g], Elaheh MOMENI-ORTNER[h] and Reza RAWASSIZADEH[d,2]

[a] *Independent Researcher, Vaughan, Canada*
[b] *School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran*
[c] *Independent Researcher, Rouen, France*
[d] *Department of Computer Science, Metropolitan College, Boston University, MA, USA*
[e] *Department of Computer Engineering and Information Technology, Payame Noor University, Tehran, Iran*
[f] *Health Management and Economics Research Center, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran*
[g] *Mental Health Research Center, Psychosocial Health Research Institute, Iran University of Medical Sciences, Tehran, Iran*
[h] *University of Vienna, Vienna, Austria*

Abstract. Background: Social networks are a good source for monitoring public health during the outbreak of COVID-19, these networks play an important role in identifying useful information. Objectives: This study aims to draw a comparison of the public's reaction in Twitter among the countries of West Asia (a.k.a Middle East) and North Africa in order to make an understanding of their response regarding the same global threat. Methods: 766,630 tweets in four languages (Arabic, English French, and Farsi) tweeted in March 2020, were investigated. Results: The results indicate that the only common theme among all languages is “government responsibilities (political)” which indicates the importance of this subject for all nations. Conclusion: Although nations react similarly in some aspects, they respond differently in others and therefore, policy localization is a vital step in confronting problems such as COVID-19 pandemic.

Keywords. COVID-19, Natural Language Processing, Social Networking, Epidemics

[1] Contributed equally

[2] Corresponding Author: Reza Rawassizadehi, Department of Computer Science, Metropolitan College, Boston University, MA, US, E-Mail: rezar@bu.edu

1. Introduction

The global spread of the COVID-19 pandemic, an infectious disease caused by the pathogen severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has already unleashed an unprecedented impact on public health, economy, and human society worldwide[1]. Social media platforms (such as Twitter, Facebook, Reddit, Tumblr, Pinterest and Instagram) have seen unprecedented growth in the era of big data. For example, Twitter, one of the most popular social network websites, which has been growing at a very fast pace. It has 284 million monthly active users, and 500 million tweets are sent per day [2]. Twitter has been used as an early warning notifier, emergency communication channel, public perception monitor, and proxy public health surveillance data source in a variety of disaster and disease outbreaks from hurricanes[3]. Millions of people are talking about the coronavirus on social media, particularly on Twitter, where there are massive conversations around a variety of topics related to COVID-19 [1].

On the other hand, public contribution is the key to bringing under control this pandemic. Researchers are making every effort to anticipate the pandemic's trajectory [4, 5]. Our work is a step forward toward better understanding of public opinions and concerns about this pandemic. In particular, this paper aims to answer the following two questions: RQ1) Which aspects of the COVID-19 pandemic has attracted the most public attention in West Asia and North Africa? and RQ2) What are the differences and similarities of the public response among the countries of West Asia and North Africa?

To address these questions, we investigated Twitter (https://twitter.com/) as a public opinion platform. Although generalization from tweets might lead to a certain degree of inaccuracy, Twitter is a repository of billions of attitude and expressions, and serves as a practical source for topic modelling. Investigating a public opinion about a wide spectrum of topics in Twitter has been made in plenty of studies [6-13]. The advantages of social media mining have been pointed out in a number of papers [14, 15] and it has been demonstrated that the data extracted from the social media platforms is comparable with that provided otherwise (e.g. questionnaires, etc.). Furthermore, tweets are largely clear from the errors inherent in traditional means of information gathering (e.g. polls and questionnaires) [16-18], where participants' opinions are dependent upon the context of the questions, their format, wording, and ordering.

Several studies have been carried out aiming to detect the topics related to COVID-19 [3, 19-26]. However, a mere of inadequate researches have focused on several languages. Nor adequate research has been conducted to address people's reaction to COVID-19 issue particularly in West Asia and North Africa. The MENA (Middle East and North Africa) region consist of Algeria, Bahrain, Egypt, Iran, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Syria, Tunisia, United Arab Emirates and Yemen. This region enjoys a broad diversity of languages. Nevertheless, most significant spoken language of the region is Arabic, spoken in Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Qatar, Saudi Arabia, Syria, Tunisia, United Arab Emirates and Yemen. Persian or Farsi, is another dominant language of this region and is spoken mainly in Iran (and Afghanstan and Tajikstan but they are outside of MENA). Although English and French are not among the official languages of this region, they are widely used throughout these countries for education, diplomacy, and business.

This paper investigates tweets of four languages dominant in West Asia and North Africa: Arabic, English, French, and Farsi. These are among the most spoken languages in this region. Results of our study can help policymakers better recognize the efficiency of their public policies, which is the key to increase the public awareness, and to encourage people to respect as much as possible the restrictive measures and, in turn, shape a better relationship with the public. Furthermore, since there is not a global and unique policy to mitigate the risks of the COVID-19 pandemic, this study provides practical information for the governments to tailor their contamination measures to the local context. In addition, the findings of this research provide helpful information for the sociologists who try to measure the social effects of this pandemic behind the scene.

2. Methods

To categorize the tweets into themes, the process presented in Figure 1 is performed for each language. To collect COVID-19 related tweets, we used Phirehose, an open source, PHP implementation of Twitter Streaming Application Programming Interface (API). Table 1 shows details about the dataset and used filters. Tweets were collected during March 2020 as it was the first huge wave of global spread of the COVID-19.

We used Python 3.7.6 to preprocess the tweets. Preprocessing steps for the four languages are quite the same including: converting letters to lower case, removing URL, mentions, stop-words and emoji, correcting repeated characters, and tokenizing and replacing negations with NOT. Arabic and Farsi language required another step that is normalization. The normalization module is used to unify those words that may be written in different alternative forms.

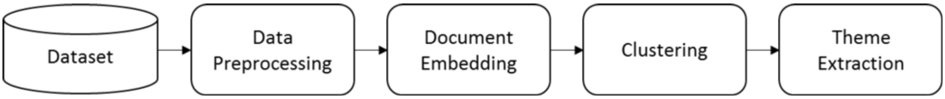


Figure 1. Process pipeline of our method

To find the most proper document embedding and clustering method for topic modeling on short texts, Curiskis et al. [27] investigated different combinations on three datasets on Twitter and Reddit and concluded that Doc2vec embedding along with k-means clustering delivered the best performance and therefore, this combination was employed in the present study. Gensim and Sklearn packages were used for Doc2vec and k-means clustering, respectively.

Table 1. Extracted dataset details

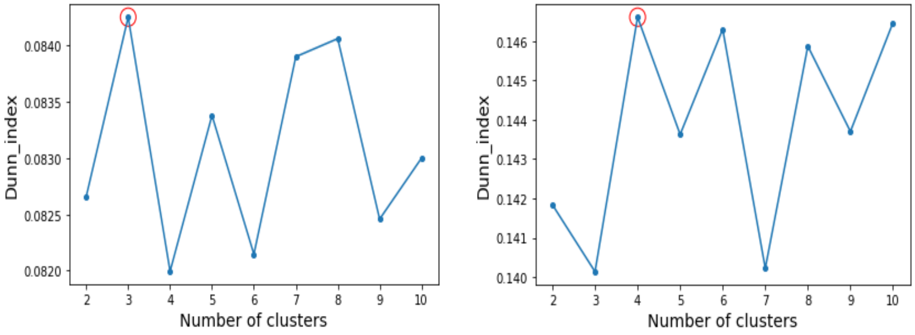
Language	Filtered hashtags	Num. of tweets	RT to Tweet Ratio	Time period
Arabic	#كورونا #كرونا #كورونا_فيروس #كرونا_فيروس #الكورونا	197,794	62.5%	2020-03-03 to 2020-03-12
Persian	#كويد19 #كرونا #ويروس_كرونا #كرونا_ويروس	199,705	57.8%	2020-03-03 to 2020-03-06
English	#corona #coronavirus #covid19 #covid2019 #covid-19	176,370	54.5%	2020-03-03 to 2020-03-09
French		192,761	65.9%	2020-03-03 to 2020-03-18

In developing the Doc2vec model, number of epochs is a case sensitive parameter that considerably affects the performance of topic modeling where it is desired for each cluster to represent the most similar texts, while for clusters to be as distinctive and different as possible. In order to find the fittest number of epochs, cosine similarity criterion [28] was employed as follows. To evaluate the tweets similarity in each cluster, 10 most frequent words in each cluster were converted to vectors using Word2vec and cosine similarity was calculated for each word-pair. The average value is the criterion for each cluster. The final value is the average of the calculated cosine similarity of all clusters. To evaluate the difference among clusters, first, the average vector of 10 most frequent words in each cluster was calculated. Next, pairwise cosine similarity of these vectors were calculated and finally, one minus the average would show how different the topics are. By applying this method, we evaluated performance for different epochs and obtain the optimum number for our Doc2vec model which in our case lead to 10 epochs. Table 2 shows the effect of number of epochs in Doc2vec model on the performance of topic modeling. Based on Table 2, 10 epochs for Doc2vec model was chosen.

Table 2. Effects of number of epochs on model performance

	Epochs	Num. topics	Cosine similarity of topics	1-cosine similarity of topics
Farsi	5	8	0.326	0.325
	10	**8**	**0.3**	**0.41**
	15	8	0.262	0,397
English	5	8	0.32	0.40
	10	**8**	**0.29**	**0.48**
	15	8	0.26	0.52
French	5	8	0.248	0.424
	10	**8**	**0.264**	**0.411**
	15	8	0.27	0.40
Arabic	5	8	0.398	0.428
	10	**8**	**0.422**	**0.472**
	15	8	0.398	0.484

To find the optimum number of clusters, Dunn index criteria [29] was utilized. For Farsi, English, French and Arabic tweets, the optimum number of clusters were obtained as 4, 5, 3 and 4, respectively. Figure 2 shows the how Dunn index is used to find the optimum number of clusters for French and Arabic tweets.



(a): French

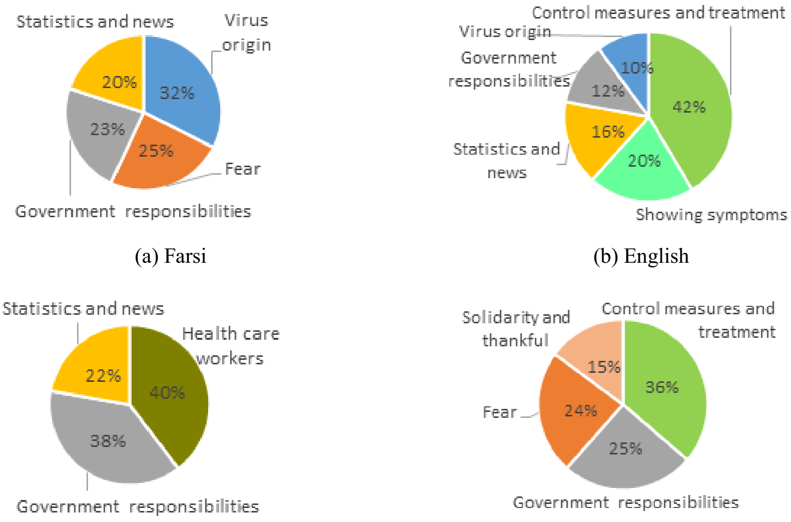
(b): Arabic

Figure 2. Dunn index versus the number of clusters in French and Arabic

After clustering tweets in each language, in order to analyze the content, each cluster should be labeled with a theme manually. To implement this task, 10 most frequent words of each cluster in addition to a random collection of 100 tweets were extracted and presented to three different native speakers of that language along with a list of suggested themes (annotators were encouraged to change the themes or add new theme as pleased and the list was just some examples to clarify the task). Candidates used the information to choose at least one theme for each cluster. In an attempt to measure the inter-rater agreement between the three raters, Fleiss Kappa testing was employed [30].

3. Results

The themes for each language and their distribution are shown in Figure 3. Farsi tweets were categorized into four themes with the Fleiss Kappa of 0.826 where “virus origin” has the largest share. English tweets with seven themes had the most versatile distribution. The Fleiss Kappa was 0.73 and most English tweets were on “control measures and treatment” theme. French and Arabic tweets were divided into three and four themes with Fleiss Kappa of 0.916 and 0.874, respectively. While the majority of French tweets were regarding “government responsibilities”, Arabic tweets were mostly about “control measures and treatment”, similar to English tweets.



(a) Farsi

(b) English

(c) French

(d) Arabic

Figure 3. Farsi, English, French and Arabic tweets analysis results pie chart

To better understand the content of the clusters, as a sample, Table 3 shows the 10 most frequent words and two samples of tweets for the “government responsibilities” theme in English and Farsi languages.

Table 3 Ten most frequent words and tweet samples for "Governmental responsibilities" theme

Language	Ten most frequent words		Tweet samples
English	outbreak	U.S.	Safety first, authorities implant strict screening measures to conduct medical checkups on passengers #dubai India suspends visas from Iran, Italy, South Korea, and Japan
	China	global	
	due	world	
	Italy	country	
	Iran	spread	
Farsi	(Iran)ایران	(US) آمریکا	ناتوانی دولت در مواجهه با ویروس صدای خبرنگار الجزیره را هم درآورد (Inability of government in confronting the virus made the al jazeera journalist to object) روحانی در جلسه هیئت دولت: قول می دهم در کوتاه ترین مدت از بحران عبور می کنیم (President Rouhani in government meeting: I promise we will pass this crisis in shortest time)
	(people) مردم	فرانسه_انگلیس (France-England)	
	(outbreak)شیوع	جمهوری_اسلامی (Islamic Republic)	
	(China)چین	(help)کمک	
	(country)کشور	مقابل (confronting)	

4. Discussion

Social media act as an appropriate source of information in dangerous situations [31]. Although at this time, the actions and reactions of the people and officials of countries are not possible in real space, the social atmosphere these days is in social media such as Twitter, and demands and actions are easily exchanged. In this study, COVID-19 related topics and discussions in English, French, Persian, and Arabic among Twitter users, inside West Asia and North Africa during the first wave of COVID-19, were analyzed and studied.

Subject of control measures and treatment is a common theme among English-speaking, and Arabic-speaking users. As COVID-19 spreads to other countries and governments try to mitigate its impact by implementing counter measures, people have also used social media platforms to express their opinion about the measures themselves, the leaders implementing them, and the ways their lives are changing [32].

The origin of the virus has been a common theme among English-speaking and Persian-speaking users. For example, in Persian language, when it was announced that the first case of the coronavirus in Iran was discovered in Qom province, the mentality of society gradually emerged that Qom province was the source of the virus, and it was also announced that the virus was transmitted by Chinese immigrants or traders. Then it has gone to other cities from Qom province. We got this in one of the Persian language clusters.

One common theme among English, French, Persian and Arabic users is the government political responsibilities, relatively authoritarian public health measures (such as physical distancing or temporary economic shutdowns) depend on societal compliance. People follow these policies when they have a good and reliable relationship with politicians and also these officials have a political economy that allows their people to stay home without suffering from hunger [33].

A common theme among Persian and Arabic users has been "fear". According to US Center for Disease Control and Prevention (CDC), this pandemic results in fear and anxiety about the new disease and also this fear may result from public health actions, and social distancing, because these actions may make people feel isolated and lonely. Therefore, governments should inform the public about the necessity of these actions and support people emotionally and provide necessary services in this regard[34].

The common theme among English-speaking, French-speaking and Persian-speaking users on Twitter is the subject of statistics and news. It shows that these communities care about the news and death statistics and the prevalence of this disease in their countries. Social media have played pivotal roles in the dissemination of information during the COVID-19 pandemic including both the rapid sharing of scientific research as well as various hoaxes and misinformation.[35].

The results of this study can be a help to improve treatment measures, macro decisions, social support, and a better understanding of people's behavior and reactions during an epidemic. For future studies, it is recommended that based on the geographical locations and time, users' opinions for this pandemic be collected and processed.

Acknowledgements

This research is supported by a grant from the Institute for Health System Innovation & Policy (IHSIP) of Boston University.

References

[1] Z. Fang, R. Costas, Tracking the Twitter attention around the research efforts on the COVID-19 pandemic, arXiv preprint arXiv:2006.05783, (2020)

[2] X. Dai, M. Bikdash, B. Meyer, From social media to public health surveillance: Word embedding based clustering method for twitter classification, in: SoutheastCon 2017, IEEE, 2017, pp. 1-7

[3] C. Ordun, S. Purushotham, E. Raff, Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs, arXiv preprint arXiv:2005.03082, (2020)

[4] K.K. Sahu, A.K. Mishra, A. Lal, Trajectory of the COVID-19 pandemic: chasing a moving target, Annals of translational medicine, 8(11) (2020)

[5] H. Loeffler-Wirth, M. Schmidt, H. Binder, Covid-19 Transmission Trajectories—Monitoring the Pandemic in the Worldwide Context, Viruses, 12(7) (2020) 777

[6] A. Javaheri, N. Moghadamnejad, H. Keshavarz, E. Javaheri, C. Dobbins, E. Momeni-Ortner, R. Rawassizadeh, Public vs media opinion on robots and their evolution over recent years, CCF Transactions on Pervasive Computing and Interaction, (2020) 1-17

[7] C. Buntain, J. Golbeck, B. Liu, G. LaFree, Evaluating public response to the Boston Marathon bombing and other acts of terrorism through Twitter, in: Tenth International AAAI Conference on Web and Social Media, 2016,

[8] T. Mitra, S. Counts, J.W. Pennebaker, Understanding anti-vaccination attitudes in social media, in: Tenth International AAAI Conference on Web and Social Media, 2016,

[9] K. Starbird, Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter, in: Icwsm, 2017, pp. 230-239

[10] M. Bagdouri, Journalists and Twitter: A Multidimensional Quantitative Description of Usage Patterns, in: ICWSM, 2016, pp. 22-31

[11] P. Missier, C. McClean, J. Carlton, D. Cedrim, L. Silva, A. Garcia, A. Plastino, A. Romanovsky, Recruiting from the network: Discovering twitter users who can help combat zika epidemics, in: International Conference on Web Engineering, Springer, 2017, pp. 437-445

[12] S. Faralli, G. Stilo, P. Velardi, What women like: A gendered analysis of twitter users' interests based on a twixonomy, in: Ninth international aaai conference on web and social media, 2015,

[13] N. Alsaedi, P. Burnap, O.F. Rana, Automatic summarization of real world events using twitter, in, AAAI, 2016,

[14] A. Hassan, V. Qazvinian, D. Radev, What's with the attitude? identifying sentences with attitude in online discussions, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1245-1255

[15] B. O'Connor, R. Balasubramanyan, B.R. Routledge, N.A. Smith, From tweets to polls: Linking text sentiment to public opinion time series, Tepper School of Business, (2010) 559

[16] H. Schuman, S. Presser, Questions and answers in attitude surveys: Experiments on question form, wording, and context, Sage, 1996.

[17] R.P. Abelson, Conviction, American Psychologist, 43(4) (1988) 267

[18] J.F. Dovidio, R.H. Fazio, New technologies for the direct and indirect assessment of attitudes, (1992)

[19] Q. Liu, Z. Zheng, J. Zheng, Q. Chen, G. Liu, S. Chen, B. Chu, H. Zhu, B. Akinwunmi, J. Huang, Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach, Journal of Medical Internet Research, 22(4) (2020) e19118

[20] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, T. Zhu, Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter, PloS one, 15(9) (2020) e0239441

[21] T. Alshaabi, J.R. Minot, M.V. Arnold, J.L. Adams, D.R. Dewhurst, A.J. Reagan, R. Muhamad, C.M. Danforth, P.S. Dodds, 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)

[22] D.R. Dewhurst, T. Alshaabi, M.V. Arnold, J.R. Minot, C.M. Danforth, P.S. Dodds, Divergent modes of online collective attention to the COVID-19 pandemic are associated with future caseload variance, arXiv preprint arXiv:2004.03516, (2020)

[23] M. Thelwall, S. Thelwall, Retweeting for COVID-19: Consensus building, information sharing, dissent, and lockdown life, arXiv preprint arXiv:2004.02793, (2020)

[24] R. Kouzy, J. Abi Jaoude, A. Kraitem, M.B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E.W. Akl, K. Baddour, Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter, Cureus, 12(3) (2020)

[25] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. Vraga, Y. Wang, A first look at COVID-19 information and misinformation sharing on Twitter, arXiv preprint arXiv:2003.13907, (2020)

[26] T.C. Hamamsy, R. Bonneau, Twitter activity about treatments during the COVID-19 pandemic: case studies of remdesivir, hydroxychloroquine, and convalescent plasma, medRxiv, (2020)

[27] S.A. Curiskis, B. Drake, T.R. Osborn, P.J. Kennedy, An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit, Information Processing & Management, 57(2) (2020) 102034

[28] B. Wang, A. Wang, F. Chen, Y. Wang, C.-C.J. Kuo, Evaluating word embedding models: Methods and experimental results, APSIPA transactions on signal and information processing, 8 (2019)

[29] J.C. Dunn, A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters, (1973)

[30] J.L. Fleiss, B. Levin, M.C. Paik, Statistical methods for rates and proportions, john wiley & sons, 2013.

[31] H.W. Park, S. Park, M. Chong, Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea, Journal of Medical Internet Research, 22(5) (2020) e18897

[32] C.E. Lopez, M. Vasu, C. Gallemore, Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset, arXiv preprint arXiv:2003.10359, (2020)

[33] S.L. Greer, E.J. King, E.M. da Fonseca, A. Peralta-Santos, The comparative politics of COVID-19: The need to understand government responses, Global public health, 15(9) (2020) 1413-1416

[34] Pandemics can be stressful, in, 2020

[35] M.R. Jimenez ‐ Sotomayor, C. Gomez ‐ Moreno, E. Soto ‐ Perez ‐ de ‐ Celis, Coronavirus, Ageism, and Twitter: An Evaluation of Tweets about Older Adults and COVID ‐ 19, Journal of the American Geriatrics Society, (2020)