doi:10.20944/preprints202004.0263.v1

Coronavirus: Public Arabic Twitter Dataset

Aseel Addawood
Information Science Department, Imam Mohammad Bin Saud University
Department of Informatics, University Illinois at Urbana-Champaign
aasdawood@imamu.edu.sa

Abstract

The COVID-19 pandemic spread of the coronavirus across the globe has affected our lives on many different levels. The world we knew before the spread of the virus has become another one. Every country has taken preventive measures, including social distancing, travel restrictions, and curfew, to control the spread of the disease. With these measures implemented, people have shifted to social media platforms in the online sphere, such as Twitter, to maintain connections. In this paper, we describe a coronavirus data set of Arabic tweets collected from January 1, 2020, primarily from hashtags populated from Saudi Arabia. This data set is available to the research community to glean a better understanding of the societal, economical, and political effects of the outbreak and to help policy makers make better decisions for fighting this epidemic.

1 Introduction

After the wide spread of the coronavirus (COVID-19) began, the World Health Organization declared it a pandemic on January 30, 2020 (World Health Organization and others, 2020). The first case was reported as originating in the city of Wuhan, China, where the government had to quarantine the whole city to overcome the quick spread of the disease. However, with globalization and the way the modern world functions, the pandemic has affected 213 countries, with more than one and a half million confirmed cases to date (World Health Organization and others, 2020).

This spread has led governments around the globe to start implementing crisis management plans and pandemic control strategies. Although governments and public health authorities may implement prevention measures and control policies, the public plays a vital role in following these measures to contain the spread of the disease.

The most important measures used to combat the spread of the virus are limiting physical contact between people and reducing the time people spend next to one another. People now rely more on the internet and online platforms to continue their social interactions. One of the most widely used social media platforms is Twitter, popular for its accessibility and ease of information sharing.

In this work, we focus on Arabic online conversation because Arabic is ranked fourth among the top 10 languages used on the web [1]. The main focus in the data set collected was on hashtags used in Saudi Arabia, although they might be used in Arabic-speaking countries outside of Saudi Arabia. Saudi Arabia is among the countries with the highest number of Twitter users among its online population (Clement, 2020; Puri-Mirza, 2019). Moreover, Saudi Arabia produces 40% of all tweets in the Arab world (Mourtada and Salem, 2014).

The data set shared is divided into conversations discussing the precautionary measures governments have applied, conversations showing social solidarity, and conversations supporting decisions governments have taken. The data set also contains data from three Saudi official accounts. The total number of tweets collected so far is 3.8 million.

Policy and decision makers can use the described data set to understand people's engagement in social media and to track the spread of misinformation.

In the following sections we describe data collection, data set statistics, and information about how to access the data set.

2 Data Collection and Description

The data collection started by identifying a list of trending hashtags and key words mostly used by the public. We used *Crimson Hexagon*, [2] , which is a social media analytic platform that provides paid data stream access. This tool allowed us to obtain tweets and retweets discussing the epidemic in Arabic. We collected data starting from January 1, 2020, until April 10, 2020, collecting 3.8 million tweets until that date. More data will be collected as the project continues.

To capture conversations related to the epidemic and people's reactions toward it, we continuously observed trending topics and hashtags. Around 70 keywords and hashtags were selected; these were later categorized based on how they oriented the conversations because this is the main purpose of hashtags. Table 1 lists hashtags that mainly discuss precautionary measures governments have applied. These include discussions of curfew, business closures, and travel restrictions.

Table 2 lists hashtags that show some kind of social solidarity within the community after applying such prevention

[1] https://www.internetworldstats.com/stats7.htm

[2] https://www.crimsonhexagon.com/

Table 1: Hashtags discussing precautionary measures applied by governments

Hashtags	English translation	Number of Tweets
إيقاف_صلاة_الجماعة	Stopping congregational prayer	7502.00
إغلاق_الحدائق	Gardens closure	180.00
صلوا_في_رحالكم	Pray in your travel	37997.00
ايقاف_الصلاة_بالمسجد	Stop praying in the mosque	2953.00
ايقاف_صلاة_الجمعة_والجماعة	Stopping Friday and group prayers	79.00
إغلاق_محلات_الحلاقة	barber shops closure	686.00
اغلاق_المقاهي	Cafes closure	30.00
اغلاق_الصالونات	Salons closure	13466.00
ايقاف_الدوري	Stopping football league	18923.00
تعليق_النشاط_الرياضي	Sports suspension	2824.00
تعليق_الرحلات_الدوليه	International flights suspended	102.00
تعليق_الرحلات_الداخليه	Internal flights suspended	94.00
تعليق_العمل	Work suspension	35170.00
تعليق_الدراسة	School Suspension	127064.00
اغلاق_النوادي_الرياضية	Gym closure	1625.00
___	Close the malls in Saudi Arabia	7614.00
إغلاق_المجمعات_التجارية	Close the commercial complexes	239.00
تعليق_القطاع_الخاص	private sector suspension	18962.00
منع_التجول	Curfew	199925.00
منع_التنقل_بين_المناطق	Prevent movement between regions	16908.00

Table 2: Hashtags showing social solidarity

Hashtags	English translation	Number of Tweets
المسافة_ما_تفرقنا	distance does not separates us	35,498
نتفرق_لصحتنا	Lets Separate for our health	27,612
احنا_قدها	We can do it	8,889
سحابة_وتعدي	A cloud and it will pass	18,800
يارب_ارفع_عنا_البلاء	Oh Lord, raise us from calamity	72,401
مبادرة_أنتم_أبطال	You are a champion initiative	1,440
تجارنا_فيهم_الخير	Our merchants are good	27,714
ابطال_الصحة_بكم_نفخر	Health champions we are proud of you	12,720
البيض_متوفر	Eggs_available	14,694

measures as social distancing. This category includes such hashtags as "distance does not separate us."

Table 3 lists hashtags that show support for the decisions and prevention measures governments have taken. This group includes hashtags encouraging people to stay home, exercise at home, or enjoy their time while in quarantine.

Table 5 lists hashtags populated by Saudi governmental Twitter accounts. These hashtags urge the community to be responsible about decreasing the number of cases by following prevention measures, reassure the community about the availability of products, and answer common questions about COVID-19. The table shows the list of hashtags accompanied by the governmental account from where they started.

One of the main ways to overcome misinformation is to take information from known, reputable sources. In social media, the most reliable sources are governmental sources. Table 4 below lists the Saudi Arabia Ministry of Health accounts and the number of tweets collected for each account.

Preliminary statistics are given in Table 6. The table shows the number of tweets, the number of retweets, the

Table 3: Hashtags supporting decisions taken by the government

Hashtags	English translation	Number of Tweets
اسبوعين_فقط	Only two weeks	29,355
تمرينك_ببيتك	Exercise at home	2,538
بيتك_ناديك	Your home is your gym	5,425
بتمرن_بالبيت	Exercise at home	40,196
اجلس_بالبيت	Sit at home	69,065
قهوتك_في_بيتك	Your coffee in your home	696
نبيها_صفر	We want it zero	94,627
المملكة_تستاهل_اكثر	The kingdom deserves more	51,914
كلنا_مسؤول_عن_الوطن	We are all responsible for the country	5,479
لبيه_ياوطن	To him, my homeland	17,572
اكثر_شي_سويته_بالحجر_المنزلي	The most thing you did in the Quarantine	30,911
الحجر_المنزلي_واجب_وطني	Quarantine is a national duty	55,942
حظر_كامل	Curfew	82,349
كلنا_بالبيت_لاجل_السعوديه	We are all home for Saudi Arabia	34,340
حظر_التجول_الكامل	Curfew	37,862
حجر_كامل	Curfew	15,579
قاعد_بالبيت	Staying home	64,355
خلك_بالبيت	Stay home	177,619
خلك_في_البيت	Stay home	220,759
انثبروا_ببيوتكم	Stay home	30,683
حظر_التجول_في_السعودية	Curfew in Saudi Arabia	36,292
وضعنا_مع_الحجر	Our situation with Quarantine	21,059
استراحتي_في_بيتي	My rest in my home	5,479
نشاطي_في_منزلي	My activity in my home	22,750
فعاليات_الحجر_الصحي	Quarantine activities	24,239
الحجر_المنزلي	Quarantine	362,132
كلنا_في_البيت_لاجل_السعوديه	We are all at home for Saudi Arabia	77,021
الزم_بيتك_حمايه_لك_ولمجتمعك	stay home for you and your community safety	31,584
الحجر_الصحي	Quarantine	141,311
اعزل_نفسك	Isolate yourself	25,344

Table 4: Saudi Arabia ministry of health accounts

Account name	Number of Tweets
@spokesman_moh	66
LiveWellMOH	637
@SaudiMOH	897

Table 5: Hashtags populated by Saudi governmental accounts

Hashtags	English translation	Account	Number of Tweets
الوقاية_من_كورونا	Corona Prevention	@SaudiMOH	293,123
كلنا_مسؤول	We are all responsible	@spokesman_moh	596,288
عش_بصحة	Live healthily	LiveWellMOH	28,541
أسئلة_كورونا	Corona's Questions	@SaudiMOH	15,455
أبطال_الصحة	Health Heroes	KSAMOFA	155,770
أبطال_المجتمع	Community Heroes	@SaudiMOH	16,641
المنتجات_متوفرة	Products available	@MCgovSA	22,992
الخدمات_مستمرة	Services continuous	@MCgovSA	4,501
متر_ونص	One and a half meters	@SaudiMOH	8,616
شكراً_أبطال_التعليم	Thanks Education heroes	@moe_gov_sa	18,227
الدراسة_مستمرة	Schools are continuing	@moe_gov_sa	152,901

Table 6: Descriptive statistics of the dataset.

Number of original tweets	707,829
Number of retweets	3,093,026
Number of unique users	1,000,244
Number of unique users who wrote original tweets	323,876

number of unique users, and the number of unique users who wrote original tweets.

3 Dataset Access

The data set is accessible through GitHub at this address: https://github.com/aseelad/Coronavirus-Public-Arabic-Twitter-Data-Set/

To comply with Twitter's Terms Conditions[3], we are unable to distribute the text of the collected data set. For that, only tweet IDs can be released and then used to retrieve the full tweet object. To do so, some tools have been developed to make the process easier; Hydrator[4] is one of these options.

Acknowledgements

The author gratefully thanks Afnan abdullah Aloqaily and Djedaini abdelhak for their help in data collection.

References

Clement, J. 2020. Countries with most twitter users 2020.

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

Puri-Mirza, A. 2019. Saudi arabia: number of internet users 2023.

World Health Organization, W., et al. 2020. Coronavirus disease 2019 (covid-19): situation report, 83.

[3] https://developer.twitter.com/en/ developer-terms/agreement-and-policy

[4] https://github.com/DocNow/hydrator