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

---

¹https://www.internetworldstats.com/stats7.htm
²https://www.crimsonhexagon.com/
shows the number of tweets, the number of retweets, 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 number of replies, and the number of likes for each account.
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, 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 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.

---

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

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