ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks

Fatima Haouari, Maram Hasanain, Reem Suwaileh, Tamer Elsayed
Computer Science and Engineering Department, Qatar University
{200159617, maram.hasanain, rs081123, telsayed}@qu.edu.qa

Abstract

In this paper, we present ArCOV-19, an Arabic COVID-19 Twitter dataset that spans one year, covering the period from 27^{th} of January 2020 till 31^{st} of January 2021. ArCOV-19 is the *first* publicly-available Arabic Twitter dataset covering COVID-19 pandemic that includes about 2.7M tweets alongside the *propagation networks* of the most-popular subset of them (i.e., most-retweeted and -liked). The propagation networks include both retweets and conversational threads (i.e., threads of replies). ArCOV-19 is designed to enable research under several domains including natural language processing, information retrieval, and social computing. Preliminary analysis shows that ArCOV-19 captures rising discussions associated with the first reported cases of the disease as they appeared in the Arab world. In addition to the source tweets and propagation networks, we also release the search queries and language-independent crawler used to collect the tweets to encourage the curation of similar datasets.

1 Introduction

Twitter streams hundreds of millions of tweets daily. In addition to being a medium for the spread and consumption of news, it has been shown to capture the dynamics of real-world events including the spread of diseases such as the seasonal influenza (Kagashe et al., 2017) or more severe epidemics like Zika (Vijaykumar et al., 2018), Ebola (Roy et al., 2020), and H1N1 (McNeill et al., 2016). Moreover, collective conversations on Twitter about an event can have a great influence on the event's outcomes, e.g., US 2016 presidential elections (Grover et al., 2019). Analyzing tweets about an event, as it evolves, offers a great opportunity to understanding its structure and characteristics, informing decisions based on its development, and anticipating its outcomes as represented in Twitter and, more importantly, in the real world.

Since the first reported case of Novel Coronavirus (COVID-19) in China, in November 2019, the COVID-19 topic has drawn the interest of many Arab users over Twitter. Their interest, reflected in the Arabic content on the platform, has reached a peak after two months when the first case was reported in the United Arab Emirates late in January 2020. This ongoing pandemic has, unsurprisingly, spiked discussions on Twitter covering a wide range of topics such as general information about the disease, preventive measures, procedures and newly-enforced decisions by governments, up-to-date statistics of the spread in the world, and even the change in our daily habits and work styles.

In this work, we aim to support future research on social media during this historical period of our time by curating an Arabic dataset (ArCOV-19) that exclusively covers tweets about COVID-19. We limit the dataset to Arabic since it is among the most dominant languages in Twitter (Alshaabi et al., 2020a), yet under-studied in general.

ArCOV-19 is the *first* Arabic Twitter dataset designed to capture tweets discussing COVID-19 starting from January 27^{th} 2020 till end of January 2021[1], constituting about 2.7M tweets, alongside the propagation networks of the most-popular subset of them. To our knowledge, there is no publicly-available Arabic Twitter dataset for COVID-19 that includes the propagation networks of a good subset of its tweets. Some existing efforts have already started to curate COVID-19 datasets including Arabic tweets, but Arabic is severely under-represented (e.g., (Chen et al., 2020a; Singh et al., 2020)) or represented by a random sample that is not specifically focused on COVID-19 (e.g., (Alshaabi et al., 2020b)), or the dataset is limited in the period it covers and does not include the propagation networks (e.g., (Alqurashi et al., 2020)).

[1]The dataset will be continuously augmented with new tweets over the coming months.

The contribution of this paper is three-fold:

- We release ArCOV-19,[2] the first Arabic Twitter dataset about COVID-19 that comprises about 2.7M tweets collected via Twitter search API. It covers a full year allowing for capturing discussions on the topic since it started to be popular in the Arab world. In addition to the tweets, ArCOV-19 includes propagation networks of the most popular subset, search queries, and documented implementation of our language-independent tweets crawler.
- We present a preliminary analysis on ArCOV-19, which reveals insights from the dataset concerning temporal, geographical, and topical aspects.
- We suggest several use cases to enable research on Arabic tweets in different research areas including, but not limited to, emergency management, misinformation detection, and social analytics.

The reminder of the paper is organized as follows. We present the crawling process followed to acquire ArCOV-19 in Section 3. We then thoroughly discuss the analysis that we conducted on the dataset in Section 4. We suggest some use cases to enable research on Arabic tweets in Section 5. We finally conclude in Section 6.

2 Related Work

Social media platforms and Twitter specifically showed to be an indispensable medium for sharing information and discussions about the COVID-19 pandemic since December 2019. A considerable body of raw social media datasets were released to facilitate analyzing these discussions including Twitter datasets (Banda et al., 2020; Lopez et al., 2020; Chen et al., 2020b; Qazi et al., 2020; Gao et al., 2020; Alqurashi et al., 2020; Dashtian and Murthy, 2021), Instagram (Zarei et al., 2020), or the Chinese social media platform Weibo (Hu et al., 2020; Gao et al., 2020).

Banda et al. (2020) released a multilingual dataset that includes over 800M tweets, the data is collected since 1^{st} of January 2020 using Twitter streaming API. Dashtian and Murthy (2021), collected multilingual tweets posted between March and July 2020. The most dominant languages covered by their dataset are English and Spanish constituting 65% and 12% of the tweets respectively. Chen et al. (2020b) are continously collecting tweets since 21^{st} of January 2020, where at the time of writing, around 66% and 11% of the tweets are in English and Spanish respectively. Another multilingual ongoing collection is released by Lopez et al. (2020). Qazi et al. (2020) enriched their large-scale multilingual Twitter dataset with geolocation information.

The only raw Arabic Twitter dataset available is the one released by Alqurashi et al. (2020), however it covers the period from 1^{st} of January to 15^{th} April 2020 only. Moreover, it does not include the propagation networks of tweets.

Compared to existing datasets, we release an only-Arabic collection of tweets, where we exclude retweets to avoid having copies of the same tweet. Moreover, we also release the most-popular subset of them (i.e., most-retweeted and -liked and spam free tweets). Furthermore, we collect the propagation networks including both retweets and conversational threads (i.e., threads of replies) for the most-popular subset.

3 Data Collection

ArCOV-19 includes two major components: the **source tweets** (i.e., tweets collected via Twitter search API every day) and the **propagation networks** (i.e., retweets and conversational threads of a subset of the source tweets). In this section, we present how we collected each in Sections 3.1, and 3.2 respectively, and summarize the released data in Section 3.3.

3.1 Tweets Collection

To collect the source tweets, we used our tweet crawler[3] that uses Twitter *search* API[4]. The crawler takes a set of manually-crafted queries, comprising a target topic, as input. At the end of each day, the crawler issues a search request for each of those queries. Queries can be keywords (e.g., "Corona"), phrases (e.g., "the killing virus"), or hashtags (e.g., "#the_new_coronavirus"). Twitter returns a maximum of 3,200 tweets per query. We customized the search requests to return only *Arabic* tweets,[5] and to exclude all retweets to avoid having copies

[2] https://gitlab.com/bigirqu/ArCOV-19/

[3] https://gitlab.com/bigirqu/ArCOV-19/-/tree/master/code/crawler

[4] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets

[5] Language of tweets to return is a configurable parameter.

of the same tweet. Additionally, duplicate tweets (returned by different queries) are removed in each day. Finally, tweets are sorted chronologically.

We started collecting our data since 27^{th} of January 2020 using a set of queries that we manually-updated based on our daily tracking of trending keywords and hashtags. For example, starting from 16^{th} of December 2020, we extended our queries to cover tweets related to COVID-19 vaccines as that sub-topic was gaining interest on Arabic social media. The full list of queries used in each day is released alongside our dataset. We denote all collected tweets as ***source tweets***.

Due to technical reasons, we missed collecting tweets for a few days. Since Twitter search API limits the search results to the past 7 days, we missed the old tweets. To overcome that, we used GetOldTweets3[6] python library to download the search results using the same trending keywords and hashtags we selected around those days.

3.2 Propagation Networks Collection

In addition to the source tweets, we also collected the ***propagation networks*** (i.e., retweets and conversational threads) of the top 1000 most popular (i.e., top-retweeted & -liked) tweets each day. To our knowledge, this is the *first* Arabic tweet dataset to include such propagation networks.

Before getting the most popular tweets on any day, we started from source tweets collected in that day and applied a qualification pipeline. We first excluded tweets containing any inappropriate word (using a list of inappropriate words we constructed). Next, tweets with more than two URLs or four hashtags, or shorter than four tokens (all are potentially spam) were dropped. Additionally, duplicate tweets that have exact textual content are also dropped to avoid redundancy; only the most popular of them (according to our scoring criterion below) is kept. Qualified tweets are then scored by popularity defined by the sum of the tweet's retweet and favorite counts. We finally sort the qualified tweets by their scores and select the most popular 1,000 tweets. We denote the set of all such tweets over all days as the **top subset**.

For those 1K top tweets per day, we then collected all retweets and conversational threads (i.e., direct and indirect replies).[7] We collected the retweets using Pickaw,[8] a platform for organizing contests on social media, and the replies using PHEME (Zubiaga et al., 2016) Twitter conversation collection script.[9]

3.3 Data Release

In summary, we release the following resources as ArCOV-19 dataset, taking into consideration Twitter content redistribution policy:[10]

- **Source Tweets**: IDs of the tweets crawled in each day.
- **Search Queries**: the list of search queries, including keywords, phrases, and hashtags, used in each period to collect our source tweets.
- **Top Subset**: IDs of the top 1,000 most popular tweets for each day.
- **Propagation Networks**: the propagation networks for the top subset which include for each tweet in the top subset:
  - **Retweets**: IDs of the full retweet set.
  - **Conversational Threads**: tweet IDs of the full reply thread (including direct and indirect replies).

Along with the dataset, we provide some pointers to publicly-available crawlers that users can easily use to crawl the tweets given their IDs.

4 ArCOV-19 in Numbers

In this section, we present a statistical summary and conduct an analysis on ArCOV-19 to shed some light on its major characteristics.

4.1 Tweets and Users Distribution

Table 1 presents an overall summary of the tweets and users statistics in ArCOV-19. It indicates that the total number of source tweets in the dataset is about 2.7M posted by over 690k unique users. We note that 18.66% of the tweets were posted by verified users (who constitute only 0.81% of the unique users). This is a relatively large percentage, showing that a good portion of the source tweets are popular. The average numbers of followers, friends, and statuses of users are also relatively large, showing that users observed in ArCOV-19 are also popular (and possibly more influential).

[6] https://github.com/Mottl/GetOldTweets3

[7] At the time of writing, we have collected the replies for tweets until end of April 2020 and we are still collecting the rest; we will make all available in the near future.

[8] https://pickaw.com/en/twrench-becomes-pickaw

[9] https://github.com/azubiaga/pheme-twitter-conversation-collection

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

The table also indicates that 25.40% of the tweets include URLs. We anticipate that this is due to the extensive spreading of news (linked from tweets) during this period. The numbers of tweets that are geotagged and geolocated are also indicated in the table; however, we defer the discussion of such types of tweets to Section 4.3.

Figure 1 illustrates the monthly distribution of tweets in ArCOV-19. The volume of tweets drastically increased in March when the virus started to spread in several Arab countries. Then, tweets number started to decrease monthly. We speculate this is due to the fact that the topic was not heavily discussed as in the first days. In December, we expanded our queries list to include hashtags and keyword related to more current sub-topics such as vaccines and the new variant of the virus, which resulted in an increase in the volume of tweets.

Figure 2 presents the top 25 tweeting users in ArCOV-19. We notice that several of them are news sources, and 20 are verified Twitter accounts.

Tweets Statistics		
Source Tweets	2,675,049	
Geotagged	2,078	(0.08%)
Geolocated	60,873	(2.28%)
Posted by verified users	499,351	(18.66%)
Include URL	679,482	(25.40%)
Include Media	973,952	(36.41%)
Top Subset	370,132	(13.84%)
Retweets of Top Subset	7,925,821	
Replies of Top Subset	1,476,950	
Users Statistics		
Unique	690,339	
Verified	5,575	(0.81%)
Average followers count	4,630	
Average friends count	918	
Average statuses count	9,054	

Table 1: Tweets and users statistics of ArCOV-19. Stats for replies are for tweets until end of April 2020.

4.2 Tweets Content & Topics

It is important to demonstrate that the tweets in ArCOV-19 constitute a good representative sample of the Arabic tweets posted during the target period on COVID-19, and that they cover the prevalent topics discussed over Twitter during that period. To help examine this hypothesis, we analyzed the textual content of the tweets; in particular, we identified the most-frequent words, hashtags, and Arab

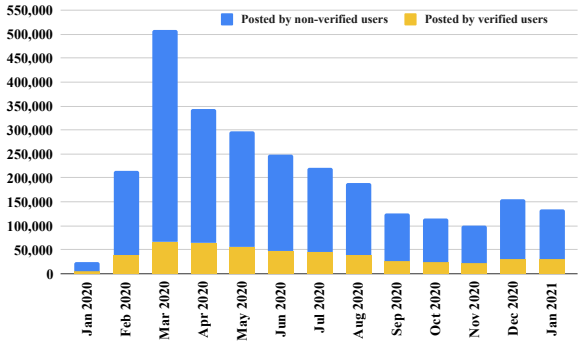

Figure 1: Monthly distribution of source tweets of ArCOV-19.

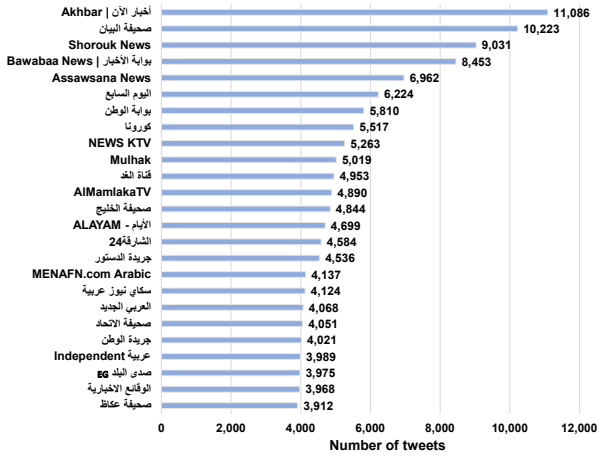

Figure 2: Number of tweets posted by the top 25 tweeters in ArCOV-19 for the period of 27^{th} Jan 2020 to 31^{st} Jan 2021.

country names for the tweets posted in the period of 27^{th} Jan to 31^{st} Mar 2020. We focus on this critical period since the virus started to spread in the Arab world during that time. We then tracked their frequency over time. Figure 3 shows the time series (over days) of the three types of keywords.[11]

The 10 most-frequent words shown in Figure 3(a) indicate two different types of words: those that are directly related to COVID-19 (e.g., "health") and those that are not but related to prayers and supplications (e.g., "Allah/God")). It is interesting to see the word "Allah" is very frequent early on when the news about the virus started to spread (probably over discussions around whether the pandemic is a punishment from God or not, we believe), then declines over time only to become frequent again when the virus started to widely spread in the Arab world.

[11] When identifying the most frequent words and hashtags, we excluded the ones we used in our search queries since they are expected to be very frequent by definition.

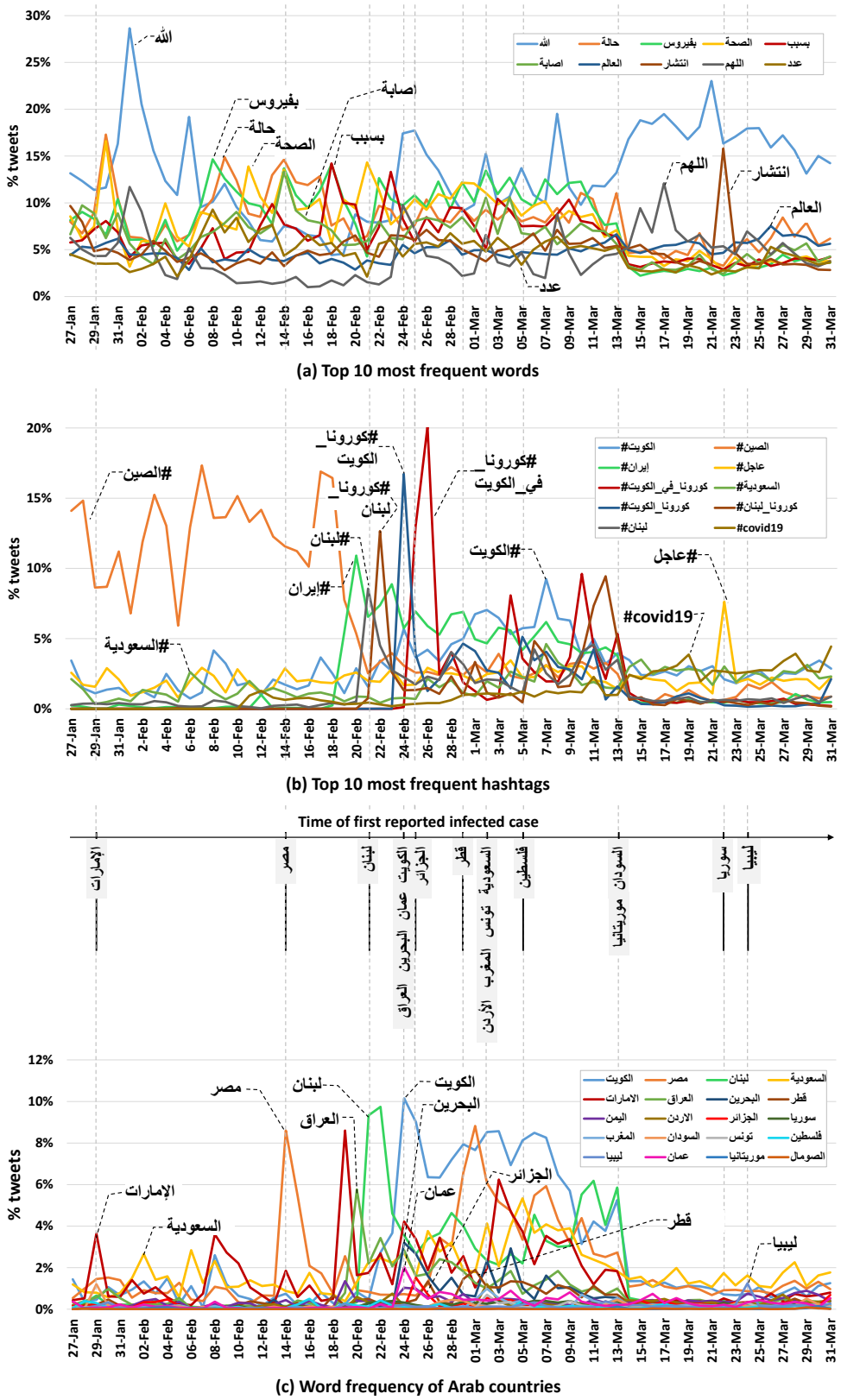

Figure 3: Time series of the frequency (in percentage of tweets) of top general keywords, hashtags, and Arab country words for the period of 27^{th} Jan to 31^{st} Mar. The time of first reported cases in Arab countries is also indicated and aligned with the time series.

Figure 3(b) demonstrates how "#China" hashtag was the most trending one from 27 of January until 20 of February, as COVID-19 was prevalent only in China and still not widely spread (at that time) in other countries. We can see how it started to be less trending as the number of cases started to decline in China by that date.[12] On the other hand, when COVID-19 started to spread in the Arab world, other hashtags started to become viral. Furthermore, Figure 3(b) shows that frequency spikes of "#Iran", "#Lebanon", and "#Kuwait" exactly match the confirmation dates of first reported cases in Iran,[13] Lebanon,[14] and Kuwait,[15] respectively.

To further analyze trending topics, Figure 3 features the timeline of the first reported cases in the Arab countries and aligns them with the time series throughout the figure. Aligning the timeline with the series in Figure 3(c) reveals a significant match between the frequency peaks of several country names and the corresponding dates of first reported cases in those countries, most notably in UAE, Egypt, Lebanon, Kuwait, Oman, Bahrain, Algeria, and Libya.

Figure 3 also demonstrates the power of ArCOV-19 in capturing further controversial and trending topics. Table 2 shows a timeline covering dates of specific topics of discussion trending on social media around times of peaks in tweeting frequency in Figure 3(c).

Country	Date	Topic	Related News
UAE	Feb 19	Yemeni Foreign Affairs Minister thanks UAE for evacuating Yemeni students from China	http://tiny.cc/71qtmz
Iraq	Feb 20	Iraq announces closure of borders with Iran	http://tiny.cc/yzrtmz
Egypt	Mar 2	Egyptian health minister announces a visit to China	http://tiny.cc/7iqtmz
	Mar 7	Kuwait bans travels with Egypt including entry of Egyptian residents	http://tiny.cc/lqstmz
KSA	Mar 4-5	Closure of The Grand Mosque in Mecca	http://tiny.cc/iqrtmz

Table 2: Examples of trending topics in social media matched by spikes in tweeting frequency in ArCOV-19

We further explore the topics discussed in ArCOV-19 by considering the domains of URLs shared in the tweets. We focused on the top tweets subset of ArCOV-19 (constructed as detailed in Section 3.2) and identified the URLs posted in those tweets. We then expanded them (since Twitter URLs are shortened) and dropped URLs of images and videos uploaded to Twitter, to focus only on URLs referring to external sources. We extracted 110,220 URLs from 2,617 unique domains. Figure 4 shows that URLs from news websites are dominantly the most-commonly shared. We observe that these news websites mainly originate from three Arab countries, namely, Egypt, Saudi Arabia, and United Arab Emirates. Interestingly, videos from YouTube are among the most commonly shared media. Coupled with the fact that 36% of the tweets in ArCOV-19 include embedded images and videos (Table 1), we believe this enables ArCOV-19 to be a potential dataset to further support the evaluation of multi-modal retrieval and classification systems.

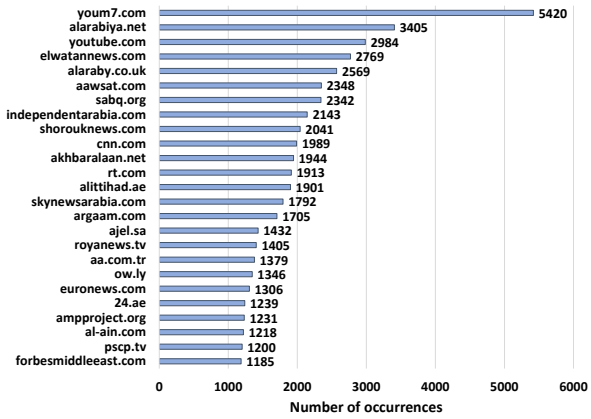

Figure 4: Most frequently-linked domains in the top tweets subset for the period of 27^{th} Jan 2020 to 31^{st} Jan 2021.

4.3 Geographic Distribution of Tweets

Although Twitter provides automatic geo-reference functionally, few users (solely around 1-3% (Murdock, 2011; Jurgens et al., 2015)) opt to enable it due to privacy and safety reasons. Alternatively, to have an insight about the geographical distribution and diversity of the tweets in our dataset, we examined the *place* and *coordinates* attributes of the Tweet object.[16] We note that the *place* attribute is an optional attribute that allows the user to select a

[12] https://tinyurl.com/rfj5khn
[13] https://tinyurl.com/yykgwbox
[14] https://tinyurl.com/yxkg78g5
[15] https://tinyurl.com/y4nlgz5h

[16] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects

location from a list provided by Twitter (therefore, the location might not necessarily show the actual location from where the tweet is posted), whereas the *coordinates* attribute represents the geographic location of the tweet as reported by the user or client application.

Table 1 shows that ArCOV-19 has 60,873 geolocated tweets (i.e., having values in the *place* attribute) and 2,078 geotagged tweets (i.e., having values in the *coordinates* attribute). Those tweets were posted by 24,072 and 256 unique users respectively. The geolocated tweets constitute about 2.28% of the total source tweets, which is consistent with previous studies (Huang and Carley, 2019).

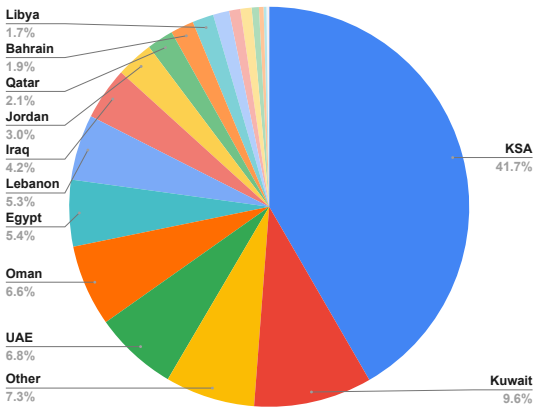

Figure 5: Distribution of geolocated tweets for the period of 27^{th} Jan 2020 to 31^{st} Jan 2021.

The geolocated tweets were indeed posted by users from 102 countries from around the globe. We found that 92.75% of them were posted from the Arab world. The largest contribution of the geolocated content (about 41.7%) comes from Saudi Arabia; this is somewhat expected as Saudi users represent the highest number of active Twitter users in the Arab world.[17] Surprisingly, Kuwait comes second with 9.6% of the geolocated content. We believe the rationale is that it was among the first countries that reported COVID-19 cases in the Middle East. Since then, Kuwait started a series of strict precautions such as a wide lock-down in many vital facilities until the government had imposed a nationwide curfew. We think, after the curfew, the people become more active on Twitter as a platform to break news and discuss developments of the virus. Additionally, we used related phrases and hashtags to "*Kuwait*", "*Lebanon*", and "*UAE*" among the tracking keywords in the period between 22 of February and 13 of March. Furthermore, it is not surprising to see the countries that have a few cases have the smallest portions of contribution to the content. Others have small audience userbases on Twitter (e.g., Tunisia).

[17] https://tinyurl.com/jmkttt3

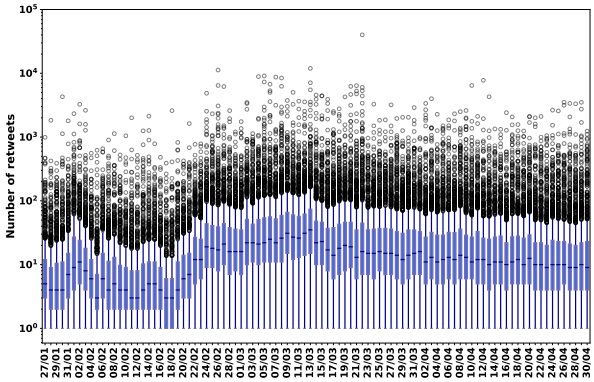

Figure 6: Distribution of retweets per day for the top subset for the period of 27^{th} Jan 2020 to 30^{th} April 2020.

4.4 Propagation Networks

As discussed earlier, we collected the propagation networks of the top subset. Overall, the number of retweets is 7,925,821 for the entire subset, and the number of replies is 1,476,950 [18].

At the time of collecting the retweets, we were not able to get the retweets of some of the tweets either because they were deleted or they were posted from private accounts. For those collected successfully, Figure 6 illustrates the distribution of retweets per day using boxplots. It shows that the average (and median) number of retweets follows a similar pattern to what was shown in Figure 1 (notice that the Y-axis here is in log scale). We also notice that a good number of tweets got more than 100 retweets; some of them got even larger numbers reaching about 10k retweets or more, showing highly propagated content. As shown in Figure 7, we applied a similar analysis to the replies per day, and found similar patterns to the retweets.

5 Enabling Research

With the spread of Novel Coronavirus in several Arab countries and the subsequent procedural measures taken by the local governments, it continued to dominate discussions over social media (and

[18] The stats are for the replies for data until end of April 2020 and we are still collecting the rest; we will make all available in the near future.

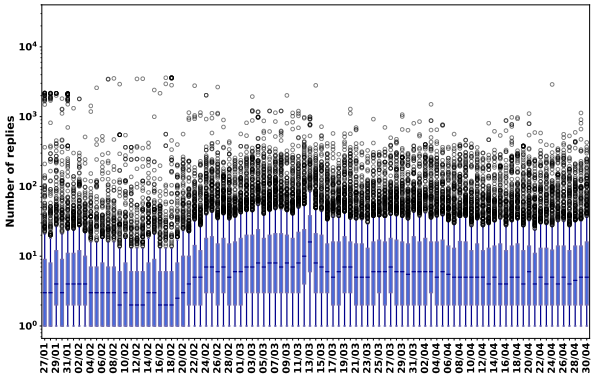

Figure 7: Distribution of replies per day for the top subset for the period of 27^{th} Jan 2020 to 30^{th} April 2020.

Twitter in particular) to the time of this writing. To enable research on different tasks on Arabic Tweets, we make ArCOV-19 publicly-available for the research community. Having quantified the topical, geographical, and community representativeness of ArCOV-19, we envision that it is suitable for diverse natural language processing, information retrieval, and computational social science research tasks including, but not limited to, emergency management, misinformation detection, and social analytics, as we discuss below. Furthermore, we provide propagation networks of daily highly popular subset (the most popular 1K) to ensure the quality of tweets. This sample is drawn after filtering out potential low-quality tweets (e.g., spam) and content duplicates. We anticipate the top subset and the network to support these tasks on the popular Arabic tweets that we assume have the most effect on shaping public opinion and understanding.

Emergency Management: As COVID-19 is an international pandemic, national and international health organizations need to analyze the effects of the outbreak. We believe ArCOV-19 can support several tasks in that domain, such as filtering of informative content, events and sub-events detection, summarization, identification of eyewitnesses, geolocation, and studying information and situational awareness propagation, to name a few.

Social Analytics: In addition to sharing reports and awareness during the outbreak, people tend to discuss their opinions (e.g., stance towards the situation and its consequences) and express their emotions (e.g., big changes in lifestyle, social distancing, loss of their beloved ones, etc.). Therefore, analyzing the tweets to detect sentiment, stance, hate speech, and generally, offensive language, among other aspects, is of interest to many stakeholders.

Misinformation Detection: With the sheer amount of information shared about COVID-19, many rumors are disseminated and getting high attention from the community, which causes a fast propagation of misinformation. This hinders the efforts of the health and governmental organizations on fighting the pandemic as such rumors spread panic and may lead to undesirable consequences (e.g., increase of cases and mortality rate, or lack of supplies due to hoarding). ArCOV-19 supports studying information/claims propagation, claims check-worthiness detection and verification tasks on the most popular tweets. Furthermore, the retweet networks and conversational threads provide a valuable resource for early detection of fake news and identification of malicious rumor-spreading accounts. To further facilitate work in this domain of problems, we recently constructed ArCOV19-*Rumors*, an annotated dataset on top of ArCOV-19 to support claim and tweet verification (Haouari et al., 2021).

6 Conclusion

In this paper, we presented ArCOV-19, the first Arabic Twitter dataset about the Novel Coronavirus (COVID-19) that includes propagation networks of a large subset of tweets. We release all source tweets, top subset, search queries, and the propagation networks. Preliminary analysis showed that ArCOV-19 captured spikes in tweeting frequency for country-specific tweets that are consistent with the first reported cases of COVID-19 in several Arab countries. We also found dominance of news agencies among top tweeting users and among most shared URLs. ArCOV-19 enables research under many domains including natural language processing, information retrieval, and social computing. We plan to continue collecting tweets for the foreseeable future and the dataset will be continuously updated with newly collected tweets and propagation networks.

Acknowledgments

The work of Tamer Elsayed and Maram Hasanain was made possible by NPRP grant# NPRP 11S-1204-170060 from the Qatar National Research Fund (a member of Qatar Foundation). The work of Reem Suwaileh was supported by GSRA grant# GSRA5-1-0527-18082 from the Qatar National Research Fund and the work of Fatima Haouari was supported by GSRA grant# GSRA6-1-0611-19074

from the Qatar National Research Fund. The statements made herein are solely the responsibility of the authors.

References

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large Arabic Twitter Dataset on COVID-19. *arXiv preprint arXiv:2004.04315*.

Thayer Alshaabi, David R Dewhurst, Joshua R Minot, Michael V Arnold, Jane L Adams, Christopher M Danforth, and Peter Sheridan Dodds. 2020a. The growing echo chamber of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *arXiv preprint arXiv:2003.03667*.

Thayer Alshaabi, JR Minot, MV Arnold, Jane Lydia Adams, David Rushing Dewhurst, Andrew J Reagan, Roby Muhamad, Christopher M Danforth, and Peter Sheridan Dodds. 2020b. 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*.

Juan M Banda, Ramya Tekumalla, Guanyu Wang, Jingyuan Yu, Tuo Liu, Yuning Ding, and Gerardo Chowell. 2020. A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688*.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020a. Covid-19: The first public coronavirus twitter dataset. *arXiv preprint arXiv:2003.07372*.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 2020b. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2):e19273.

Hassan Dashtian and Dhiraj Murthy. 2021. Cml-covid: A large-scale covid-19 twitter dataset with latent topics, sentiment and location information. *arXiv preprint arXiv:2101.12202*.

Zhiwei Gao, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2020. Naist covid: Multilingual covid-19 twitter and weibo dataset. *arXiv preprint arXiv:2004.08145*.

Purva Grover, Arpan Kumar Kar, Yogesh K. Dwivedi, and Marijn Janssen. 2019. Polarization and acculturation in us election 2016 outcomes – can twitter analytics predict changes in voting preferences. *Technological Forecasting and Social Change*, 145:438 – 460.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, WANLP 2021.

Yong Hu, He-Yan Huang, Anfan Chen, and Xian-Ling Mao. 2020. Weibo-cov: A large-scale covid-19 social media dataset from weibo. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Binxuan Huang and Kathleen M Carley. 2019. A large-scale empirical study of geotagging behavior on Twitter. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 365–373.

David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media*.

Ireneus Kagashe, Zhijun Yan, and Imran Suheryani. 2017. Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using twitter data. *Journal of medical Internet research*, 19(9):e315.

Christian E Lopez, Malolan Vasu, and Caleb Gallemore. 2020. Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset. *arXiv preprint arXiv:2003.10359*.

Andrew McNeill, Peter R. Harris, and Pam Briggs. 2016. Twitter influence on uk vaccination and antiviral uptake during the 2009 h1n1 pandemic. *Frontiers in Public Health*, 4:26.

Vanessa Murdock. 2011. Your mileage may vary: On the limits of social media. *SIGSPATIAL Special*, 3(2):62–66.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1):6–15.

Melissa Roy, Nicolas Moreau, Cécile Rousseau, Arnaud Mercier, Andrew Wilson, and Laëtitia Atlani-Duault. 2020. Ebola and localized blame on social media: Analysis of twitter and facebook conversations during the 2014–2015 ebola epidemic. *Culture, Medicine, and Psychiatry*, 44(1):56–79.

Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. 2020. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*.

Santosh Vijaykumar, Glen Nowak, Itai Himelboim, and Yan Jin. 2018. Virtual zika transmission after the first us case: who said what and how it spread on twitter. *American journal of infection control*, 46(5):549–557.

Koosha Zarei, Reza Farahbakhsh, Noel Crespi, and Gareth Tyson. 2020. A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226*.

Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3).