data

MDPI

Data Descriptor

Arabic Twitter Conversation Dataset about the COVID-19 Vaccine

Huda Alhazmi

Department of Computer Science, Umm Al-Qura University, Makkah 24236, Saudi Arabia; hnhazmi@uqu.edu.sa

Abstract: The development and rollout of COVID-19 vaccination around the world offers hope for controlling the pandemic. People turned to social media such as Twitter seeking information or to voice their opinion. Therefore, mining such conversation can provide a rich source of data for different applications related to the COVID-19 vaccine. In this data article, we developed an Arabic Twitter dataset of 1.1 M Arabic posts regarding the COVID-19 vaccine. The dataset was streamed over one year, covering the period from January to December 2021. We considered a set of crawling keywords in the Arabic language related to the conversation about the vaccine. The dataset consists of seven databases that can be analyzed separately or merged for further analysis. The initial analysis depicts the embedded features within the posts, including hashtags, media, and the dynamic of replies and retweets. Further, the textual analysis reveals the most frequent words that can capture the trends of the discussions. The dataset was designed to facilitate research across different fields, such as social network analysis, information retrieval, health informatics, and social science.

Dataset: https://data.mendeley.com/datasets/zmwfnsms9n

Dataset License: CC BY 4.0

Keywords: COVID-19; pandemic; vaccine; Twitter; dataset; Arabic

check for updates

Citation: Alhazmi, H. Arabic Twitter Conversation Dataset about the COVID-19 Vaccine. *Data* **2022**, *7*, 152. https://doi.org/10.3390/data7110152

Received: 16 October 2022
Accepted: 28 October 2022
Published: 4 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Summary

Social media platforms such as Twitter, Facebook, YouTube, and Instagram can be a powerful source of data [1]. In recent years, Twitter has been considered a popular source for news broadcasting, marketing, advertising, emerging technologies, global events, and politics [2]. Millions of users use Twitter to interact and exchange news and information. Recently, Twitter had 217 million daily active users who post 500 million tweets a day [3]. Twitter is not only a platform for users to socially interact and maintain social ties, but it has become a communication channel between organizations and society. Leaders, government organizations, and institutions communicate with society through their posts [4,5]. Moreover, conversations on Twitter regarding an evolving topic can offer a great opportunity for investigating people's opinions and understanding their behaviors. Such understanding can help governments and organizations in decision- and policy-making.

The role of social networks, and in particular Twitter, during crises or pandemics has provoked interest among researchers and experts. The data content can provide important insights into the management and analysis of crises, such as Ebola [6] and the seasonal influenza epidemics [7]. Since the outbreak of COVID-19, there has been a significant increase in the number of posts on Twitter related to this pandemic. The World Health Organization (WHO) used Twitter as a communication channel to inform people about the coronavirus and vaccination, to prevent false or fake information [8]. Further, the conversation about the COVID-19 pandemic has drawn people's attention over the world, and people have turned to Twitter to share their opinion and look for information [9]. The development and production of vaccines [10] offered a potential solution to controlling the pandemic. Most vaccine distribution campaigns started on December 2020 [11]. The spread of vaccine information that can impact vaccine uptake can be significantly increased with the wide usage of social media [12]. Therefore, studying and understanding the content of social media around vaccinations can shape public opinion.

Mining the content of social media to develop datasets, particularly tweeted conversations, has gained considerable attention from the scientific community in different fields, such as data mining, machine learning, natural language processing, big data analysis, and social network analysis. In public health, using Twitter to carry out research has increased significantly. Hence, developing tweet datasets to be used by researchers will enrich the research in this field. Recent work developed data on drug safety [13], inflammatory bowel disease [14], and personal health information [15]. As the surge of COVID-19 started, considerable efforts have been made to develop a multilingual Twitter dataset related to coronavirus [16–18]. Arabic posts were also collected to develop Arabic datasets [19,20] regarding COVID-19. Although there have been several works on Twitter datasets of COVID-19 vaccination, they are confined to some aspects covering vaccine misinformation detection [21,22], vaccine stance [23], or sentiment analysis [24,25].

Despite Arabic being one of the most dominant languages on Twitter [26], few studies have developed a COVID-19 vaccine tweet dataset. Also, to the best of our knowledge prior works on developing an Arabic dataset related to the COVID-19 vaccine focus on developing labeled datasets for a specific purpose such as detecting vaccine misinformation or sentiment analysis. To address this limitation, we developed a Twitter conversation dataset related to the COVID-19 vaccine with a focus on Arabic tweets only. The aim of this work is to explore and analyze the dynamics of the conversations on Twitter regarding the vaccine to develop a dataset that could be used in the investigation of different research topics. The contribution of this research is threefold:

1. We build an Arabic Twitter dataset of 1.1 M Arabic posts that was streamed over one year, covering the period from January to December 2021. The data collection started when most countries around the world started the COVID-19 vaccination campaigns. Thus, the dataset covers the initial dynamic conversation on vaccine distribution.
2. We performed a preliminary analysis on the raw data which revealed topical insights and resulting in seven database tables. Further analysis can be done among multiple database tables.
3. We release the dataset to be freely available to the research community in the Mendeley data repository https://data.mendeley.com/datasets/zmwfnsms9n (accessed on 31 October 2022). The dataset can be useful for researchers in different fields to analyze people's activity following the first announcement of the vaccine distribution or to perform comparative analysis. Moreover, scientific communities, public health agencies, and analysts might be interested in this dataset to obtain insights, make decisions, or design strategies that might help in some potential situations

The rest of this article is organized as follows. Section 2 presents an overview of recent works. Section 3 provides the description of the dataset. A detailed description of the developing method of the dataset is presented in Section 4. Section 5 contains the results and analysis. A brief description of the potential research application of the dataset is presented in Section 6. The conclusion is presented in Section 7.

2. Literature Review

During the pandemic, social media platforms played an important role in informing the public and spreading information. These channels provide timely and reliable data that can be valuable for mining and investigating the public's stance. Consequently, several studies have developed Twitter post datasets related to the coronavirus or vaccination in different languages. Singh et al. [16] collected 2.79 M tweets from 16 January to 15 March 2020. The tweets include multilingual conversations related to COVID-19. They used two datasets, the first one made up of tweets with location mentions and the second comprising geotagged tweets. They performed a cross-correlation analysis between the two datasets and data from the WHO. They found that the conversations were highly correlated with the confirmed cases of COVID-19. They suggested that the tweeted conversations may be a leading guide to the cases. Then, they analyzed the content of the English tweets to identify themes and the predominant myths.

Another study by Chen et al. [17] collected and analyzed over 72 M tweets from 21 January to 21 March 2020. The dataset includes multilingual COVID-19 Twitter posts. They performed an initial analysis including hashtags, languages, and verified users. They found that hashtag usage increased when COVID-19 was declared a global public health emergency. Their analysis showed that the most active accounts are news and political accounts. Recently, Aguilar et al. [18] gathered 8.98 M Twitter posts over 23 days that reflect the early discussion about coronavirus. The dataset is multilingual with a focus on English, Portuguese, and Spanish languages. They developed different databases including, hashtags, links and media, and retweets. Also, they classified tweets into different types based on whether they were original tweets, retweets, with mention, or without. Their analysis showed that the most retweeted tweets belong to accounts such as news media, politicians, actors, official institutes, and activists.

A multilingual COVID-19 Twitter dataset covering 268 countries was presented by Abdul-Mageed et al. [27]. The billion-scale dataset was classified as pandemic-relevant tweets or misinformation. For classification, they used two predictive models. They trained a COVID-relevant classifier using a sample of multilingual tweets developed by [14] and considered them as the positive class. To train the misinformation detection model, they used two publicly available datasets as a positive class.

This pandemic has drawn the interest of users who are writing in the Arabic language to become involved in discussions that cover a range of topics related to the coronavirus. Haouari et al. [19] collected about 2.7 M Arabic posts regarding COVID-19 for a year. The data covers the topic in Arab countries, and it includes the tweets and propagation networks of the most 1000 popular tweets. They performed a preliminary analysis of the tweets and user distribution that revealed temporal information and geographical aspects. Further analysis of trending topics discovered a considerable relationship between the frequency peak and the first reported case of the disease. An annotated tweet dataset was developed by Elhadad et al. [20] and contains Arabic and English tweets to detect misleading tweets related to COVID-19. The tweets were annotated using automatically different machine learning techniques and several feature extraction techniques.

Another study by Haouari et al. [28] covered 9.4 K labeled tweets and their propagation networks related to the detection of misinformation and rumors about COVID-19. They used an Arabic BERT-based model to classify tweets based on two levels of misinformation detection. Alam et al. in [29] released a 16-K tweet dataset in multiple languages that focused on COVID-19. The dataset was manually annotated for disinformation analysis. Through annotation, the authors determined if the tweets contained accurate claims and their potential for causing harm. They used a training model for each language, and then multilingual training was performed for the data in all languages. An annotated dataset of 10 K Arabic and English tweets was developed by Yang et al. [30] for sentiment analysis. They categorized the tweets into 10 classes using multi-label classifiers based on deep neural network models. The analysis showed that the positive sentiment increased over time. Alsudias et al. [31] collected a dataset of 1 M Arabic tweets about COVID-19. Then, they performed an annotation process for rumor detection on a random sample of the tweets. They applied three machine learning techniques with two sets of features to classify the tweets into false, correct, or unrelated tweets. Further, they performed an analysis to predict the source of the tweets regarding COVID-19.

More recently, several studies analyzed and provided tweet datasets about the COVID-19 vaccine. A dataset developed by Zhou et al. [32] called ReCOVery includes about 2000 news articles with 140 K tweets related to this news to predict reliable and unreliable news. For the prediction task, they used a neural network model. They found that 60 percent of the news was identified as extremely high or low in credibility. Then they tracked the news on Twitter through their URLs. Their dataset includes multimodal information from news articles. Malagoli et al. [25] collected about 12 M tweets over two months during the early stage of vaccination. They performed sentiment and psycholinguistic analysis to investigate user engagement. For sentiment analysis, they identified the strength of positive and negative opinions. Further, they studied the psycholinguistic property of the tweets to identify users' communication using the Word Count lexicon technique.

Research by Muric et. al. [33] presented a Twitter dataset in English on anti-vaccine sentiments. They collected the historical tweets of accounts that engaged in antivaccination conversations. They performed an initial descriptive analysis, such as hashtag frequency and detecting the source of the news. In addition, they produced a geographical analysis of the tweets. Similarly, a Twitter dataset in English was presented by DeVerna et al. [34]. They performed descriptive analysis for one week of extracted Twitter posts about the COVID-19 vaccine. The analysis includes hashtag clusters and geographical distribution of the tweets. Moreover, the authors present a visualization of the statistical data through a data dashboard. Hu et al. [24] collected over 300 K tweets in the US related to COVID-19 vaccines. Using spatiotemporal patterns, they investigated the sentiments and emotions of the public over time. For sentiment analysis, they used a rule-based model, whereas for emotion analysis the lexicon base was used. Further analysis was performed using topic modeling and word cloud mapping.

Another work by Memon et al. [22] proposed a misinformation tweets dataset related to the COVID-19 vaccine. The dataset distinguishes the users who post truthful information and the others who spread misinformation. They used several machine learning and deep learning algorithms. They performed two labeling steps, manual annotation and then validation by a medical expert. The ArCovidVac dataset introduced by Mubarak et al. [23] annotated a 10-K Arabic tweet dataset. Informativeness analysis and stance towards vaccination across an Arab country were presented. The tweets are annotated based on their stance on the vaccination process into a pro-, against vaccination, or neutral stance. Then, they explore the content of the annotated tweets for topics, hashtags, and source of the tweet. Further analysis was performed to investigate the public stance over time. Table 1 summarizes the recent works on COVID-19 and vaccine datasets.

Table 1. Recent research works regarding COVID-19 and vaccine datasets.

Study	Available Online	Period	Dataset	Language	Application
[16]	No	January 2020–March 2020	COVID-19 tweet conversation	Multilingual	Content analysis and topic and prevalent myths detection
[17]	Yes	January 2020–March 2020	COVID-19 Twitter posts	Multilingual	Initial content analysis
[16] [18]	Yes	January 2020–February 2020	COVID-19 Twitter posts	Multilingual	Statical and content analysis
[19]	Yes	January 2020–January 2021	COVID-19 Twitter posts	Arabic	Statical and content analysis
[20]	Yes	February 2020–March 2020	COVID-19 Tweets	English and Arabic	Misinformation detection
[22]	Yes	December 2020–July 2021	COVID-19 vaccine annotated tweets	English	Misinformation detection
[23]	Yes	January 2021–February 2021	COVID-19 vaccine annotated tweet dataset	Arabic	Vaccination stance detection and content analysis
[24]	No	March 2020–February 2021	COVID-19 vaccines tweets in US.	English	Sentiment analysis and emotion analysis Topic modeling and word cloud mapping.
[25]	Yes	December 2020–January 2021	COVID-19 vaccines tweets	English	Sentiment and psycholinguistic analysis
[27]	Yes	January 2020–July 2020	COVID-19 tweets	Multilingual	Analysis and classification

Table 1. *Cont.*

Study	Available Online	Period	Dataset	Language	Application
[28]	Yes	January 2020–January 2021	COVID-19 tweets	Arabic	Misinformation detection
[29]	Yes	January 2020–March 2021	COVID-19 tweets	Multilingual	Disinformation analysis
[30]	Yes	March 2020–May 2020	COVID-19 tweets	English and Arabic	Sentiment analysis
[31]	No	December 2020–April 2020	COVID-19 tweets	Arabic	Rumor detection
[32]	Yes	January 2020–May 2020	COVID-19 vaccine news articles and related tweets	English	Reliable and unreliable news prediction
[33]	Yes	October 2020–December 2020	Twitter dataset in anti-vaccine.	English	Antivaccination descriptive analysis
[34]	Yes	December 2020–January 2021	COVID-19 vaccines Twitter posts.	English	Descriptive analysis and statistics visualization

All the mentioned developed datasets related to COVID-19 vaccination mainly focused on rumor detection, misinformation detection, vaccine hesitancy, or sentiment analysis. Moreover, there is a limitation on publicly available Arabic datasets regarding COVID-19 vaccine. Therefore, we develop an Arabic twitter posts dataset targeting the vaccination discussion regarding the COVID-19 vaccine. We presented a basic analysis that shows the dynamic of vaccination-related conversation.

3. Data Description

This paper presents a collection of 1,101,349 Arabic posts from Twitter. These Arabic tweets reflect the discussion about the COVID-19 vaccine. The tweets were streamed for about twelve months from January 2021 to December 2021. This period was selected because most countries around the world started the COVID-19 vaccination campaigns in December 2020. Figure 1 illustrates the monthly distribution of the collected tweets. We noticed that the volume of tweets increased considerably in February and continued rising until May. Then, the tweets' number started to drop as the topic becomes out of date. However, it started to increase again in July 2021 as many people expressed their opinion after receiving the first dose of the vaccine.

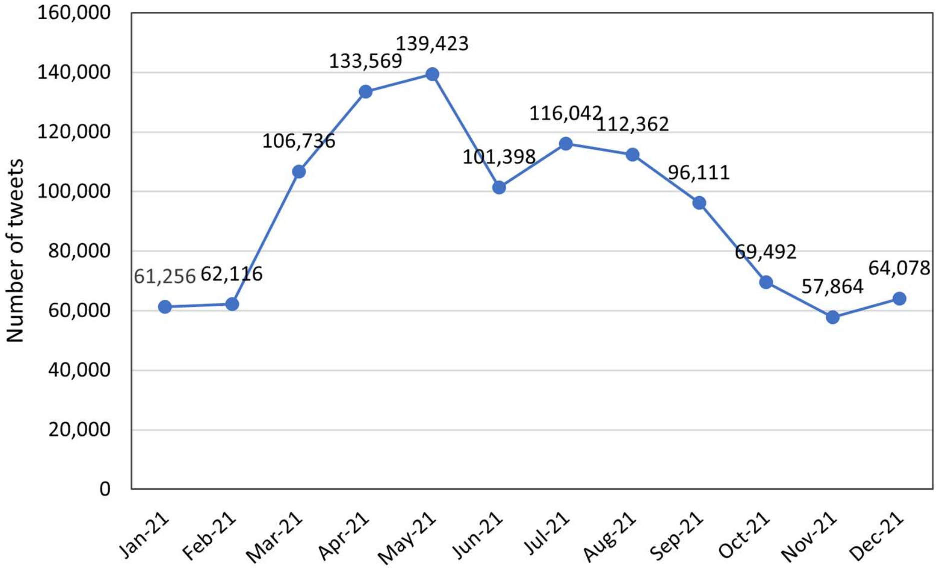

Figure 1. The distribution of Twitter posts regarding COVID-19 vaccine over twelve months.

The raw data was filtered and analyzed to create different databases which are available in a Mendeley dataset. The dataset is published in compliance with Twitter's terms and conditions, which do not allow the publication of the text of a tweet [35]. Therefore, we released Tweet IDs in all the database files which is a unique identifier that can be used to retrieve tweet's object using Twitter's API. Table 2 shows a brief description of the databases and their fields. Furthermore, all the database files contain the filed tweet-id which can be used to join them for further analysis.

Table 2. Databases in Mendeley dataset.

Database	Description	Fields
D1.General	Collection of tweets regarding the COVID-19 vaccine. Estimated size: 58.64 MB.	tweet_id: unique id for each post. datetime: the date and time of creation of the tweet. keyword: term used to extract the tweets.
D2.Media	Collection of tweets with at least one media. Estimated size: 24.25 MB.	tweet_id: unique id for each post. media_type: type of the media (photo, gif, or video) media_url: complete URL of the media
D3.Hashtag	Collection of hashtags in each tweet. Estimated size: 27.51 MB.	tweet_id: unique id for each post. datetime: the date and time of creation of the tweet. hashtag: terms used as hashtag within the tweet.
D4.Reply	Collection of tweets that had at least one reply and the count of all the replies to the tweet. Estimated size: 20.12 MB.	tweet_id: unique id for each post. datetime: the date and time of creation of the tweet. twreply_count: number of replies to each tweet
D5.Retweet	Collection of tweets that had at least one retweet and the count of all the retweets for the tweet. Estimated size: 6.012 MB.	tweet_id: unique id for each post. datetime: the date and time of creation of the tweet. retweet_count: number of retweets for each tweet
D6.Vaccine_type	Collection of tweets about different types of vaccine Estimated size: 26.64 MB.	tweet_id: unique id for each post. datetime: the date and time of creation of the tweet. vac_type: type of the vaccine
D7.Users	Collection of nodes of unique users. Estimated size: 5.684 MB.	user_id: user's id account

In the database "D1.General", each tweet is associated with the keyword that was used to retrieve it. The database "D6.Vaccine_type" includes a variable that represents the type of vaccine and the collection of posts related to each type. Figure 2 illustrates the proportion of the vaccine type; we noticed that most of the conversation was related to the vaccine type "Pfizer". Unique users were collected in the database "D7.Users" which can be used to construct social interaction networks.

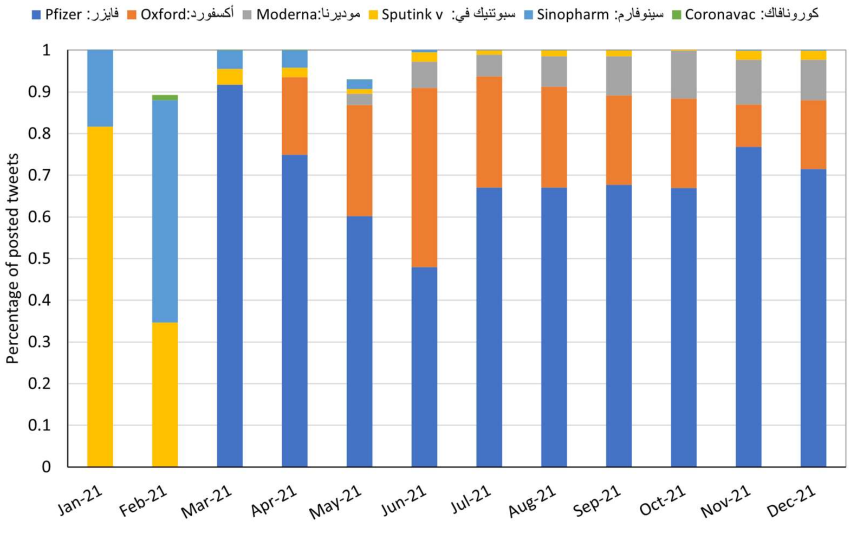

Figure 2. The percentage of Twitter posts by the type of vaccine per month.

4. Results and Analysis

We present an in-depth analysis of the dataset, underlining the most trending hashtags, the tweets containing media, and the unique users. Moreover, we analyzed the retweet and replay interactions that reflect the dynamic of the conversations in the dataset. Further, we investigated the text of the tweets to explore the predominant terms related to the COVID-19 vaccination conversation.

4.1. Hashtag

Figure 3 shows a word cloud of the most frequent hashtags in the dataset. This visualization shows they are highly related to COVID-19 vaccination. Table 3 illustrates the top 50 hashtags used in the dataset because they were the most frequently occurring. We can observe that these hashtags include names of different Arabic countries, which suggests these countries might be the source of the hashtags. Interestingly, some hashtags show a positive attitude toward the vaccination, such as (يدا_بيد_نتعافى), and some against, such as (لا_للتطعيم_الاجباري). Moreover, they can give us an indication of the most popular hashtags in the Arab region. Exploring the hashtag data can provide researchers with some insights into the topics or trends analysis.

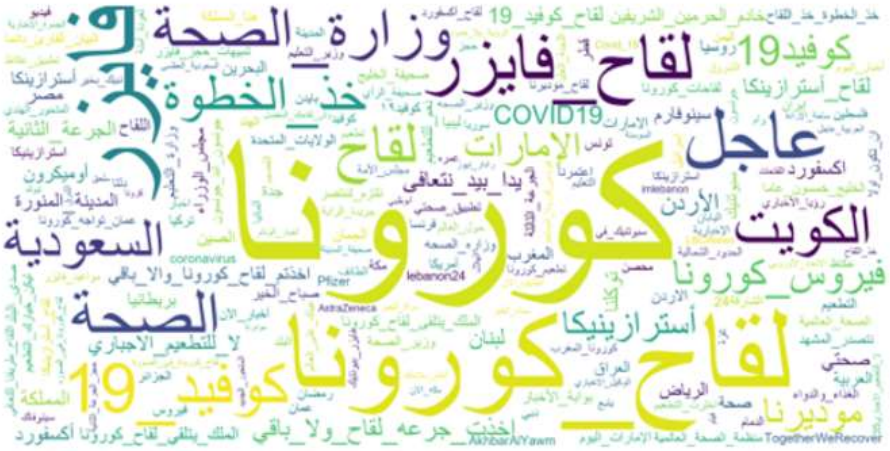

Figure 3. Word cloud of the hashtags related to the COVID-19 vaccine conversation on Twitter.

Table 3. The top 50 hashtags related to the COVID-19 vaccine conversation on Twitter.

Hashtag	English Translation	Counts	Hashtag	English Translation	Counts
كورونا	Corona	65,130	توكلنا	Twakklna (App used in SA)	2709
لقاح_كورونا	Corona vaccine	38,778	أكسفورد	Oxford	2630
فايزر	Pfizer	32,770	اكسفورد	Oxford	2585
عاجل	Urgent	13,683	صحتي	Sehaty (App used in SA)	2542
لقاح_فايزر	Pfizer vaccine	11,907	COVID19	COVID19	2540
الصحة	Health	9433	مصر	Egypt	2481
كوفيد_19	COVID-19	9157	المدينة_المنورة	Madinah	2427
وزارة_الصحة	Ministry of health	8786	أوميكرون	Omicron	2368
لقاح	Vaccine	8612	سينوفارم	Sinopharm	2239
الكويت	Kuwait	8419	اخذتم_لقاح_كورونا_والا_باقي	Did you take the vaccine or not yet	2204

Table 3. *Cont.*

Hashtag	English Translation	Counts	Hashtag	English Translation	Counts
السعودية	Saudi	7399	أسترازينكا	Astrazeneca	2137
خذ_الخطوة	Take the step	6470	خادم_الحرمين_الشريفين	Custodian of the two holy mosques	2131
فيروس_كورونا	Corona virus	5347	الرياض	Riyadh	2095
كوفيد19	COVID-19	5290	يدا_بيد_نتعافى	Hand by hand recovering	2035
الإمارات	Emirates	5263	بريطانيا	United Kingdom	1895
موديرنا	Moderna	5217	المغرب	Morocco	1870
أسترازينيكا	Astrazeneca	4221	الملك_يتلقي_لقاح_كورونا	The king got the corona vaccine	1852
اخذت_جرعه_لقاح_ولا_باقي	Did you take the dose or not yet	4204	صحة	Health	1785
الأردن	Jordan	4192	الصين	China	1770
لقاح_أسترازينك	Astrazeneca vaccine	3370	اللقاح	Vaccine	1742
لبنان	Lebanon	3200	روسيا	Russia	1708
الجرعة_الثانية	Second dose	2845	البحرين	Bahrain	1650
لا_للتطعيم_الاجباري	No to compulsory vaccination	2780	العربية	Arabia	1647

4.2. Media

We found that 83.51% of the tweets contained photos and 16.49% contained videos. Interestingly, we found that almost 80% of the tweets did not include any media. Figure 4 shows the proportion of tweets that contained media over time. Notably, photos are the most shared media, rather than videos and gifs. Together with the fact that 20% of our dataset includes media links, we think the dataset can be useful for retrieval or classification.

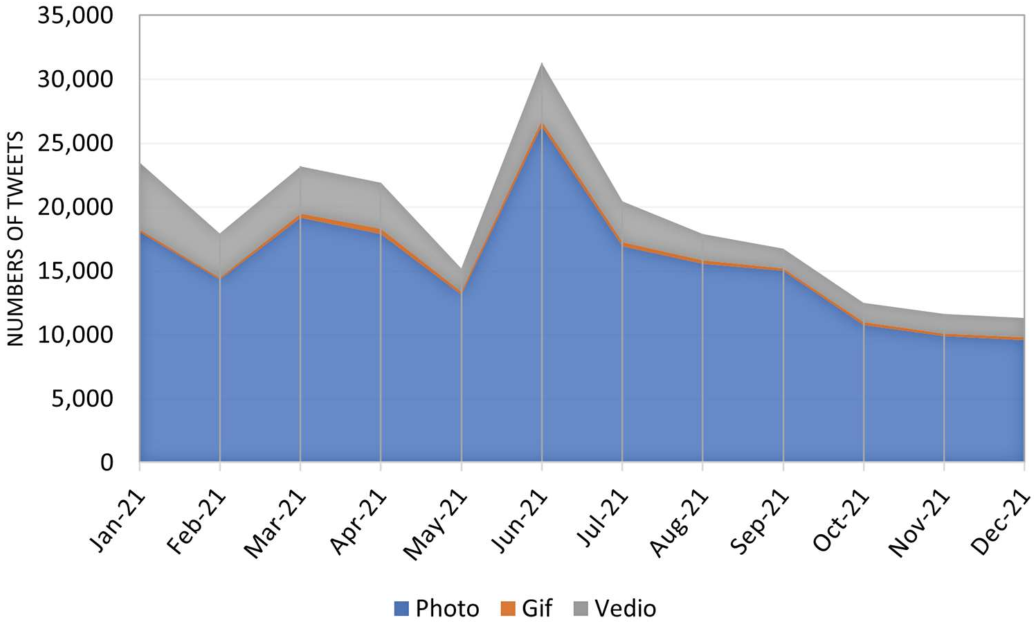

Figure 4. The number of tweets containing media over time.

4.3. Users

We observed that a significant percentage of users (about 60%) posted just one tweet and 30.37% tweeted less than five tweets. Further, we analyzed the unique users to extract the most 20 active users and the number of tweets they had posted. We noted that news organizations dominate the tweet posting among top users. Figure 5 shows the top 20 users with their tweet counts. The highest number of the top users was located in Saudi Arabia with 25%, followed by Egypt with 20%, and Kuwait, as shown in Figure 6.

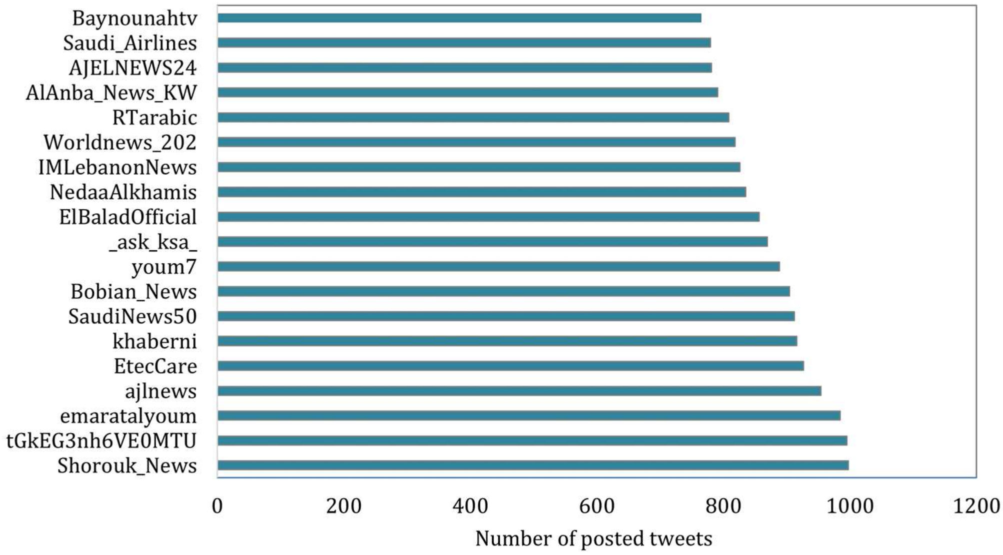

Figure 5. The top 20 users and their posted tweet count.

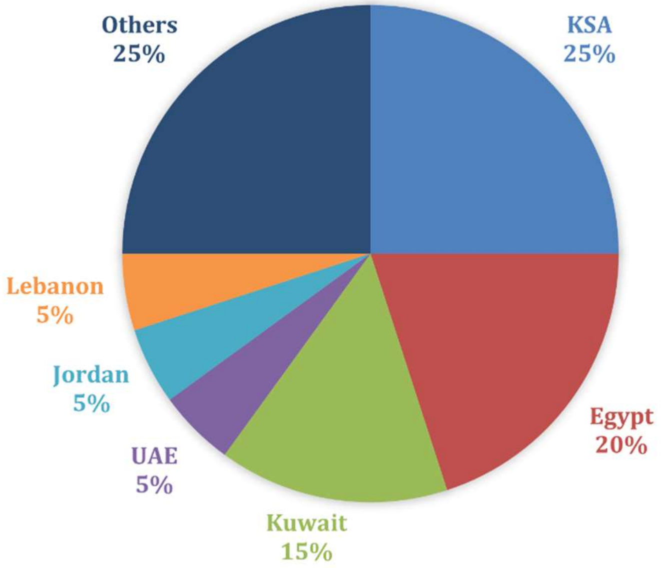

Figure 6. The percentage of top users' tweets by country.

4.4. Retweet and Reply Analysis

We noted that 3.66% of the tweets had been retweeted more than 100 times, and some tweets had more than 1 K or more retweets. Figure 7 shows the counts of tweets that had been retweeted per month as well the counts of tweets that have been replied to per month. We noted the same pattern of retweet and reply through the time. The timeline of retweeting and replying demonstrates the capability of the dataset in capturing the dynamic of conversations regarding the vaccine.

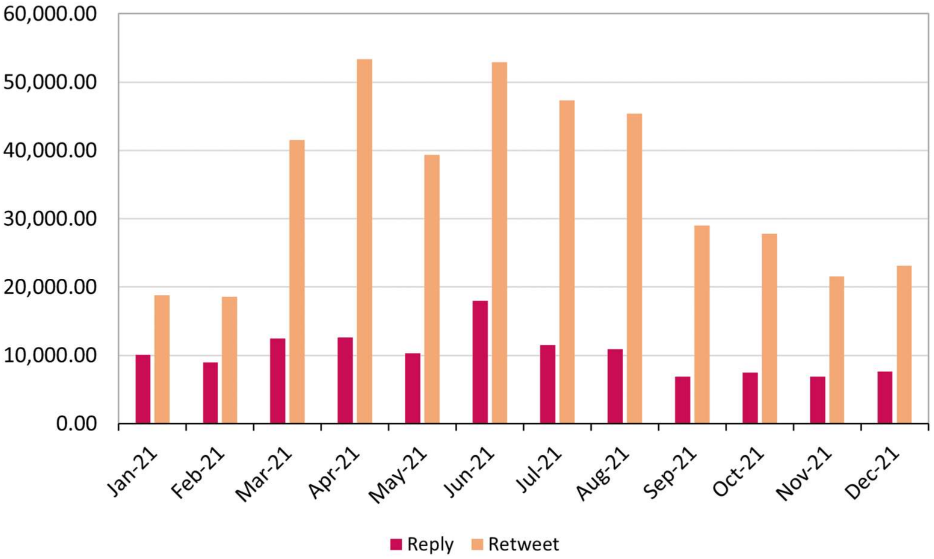

Figure 7. The count of retweets and replies over time.

4.5. Textual Analysis

Table 4 presents the top 20 most frequent words from January to December 2021. We can observe that the terms (كورونا, جرعه) and their corresponding English terms (corona, dose) were the most dominant terms in the conversation. Further, we identify the most frequent words over time. We further look at the prevalence of the words over time as shown in Figure 8. We noted that most of the terms are related to the vaccine, which is not surprising since the conversation is about the coronavirus vaccination.

Table 4. The top 20 most frequent words.

January–April		May–August		September–December	
كورونا	Corona	جرعه	Dose	كورونا	Corona
جرعه	Dose	كورونا	Corona	جرعه	Dose
اولى	First	اخد	Got	اخذ	Got
صحه	Health	ثانيه	Second	جرعتين	Two doses
موعد	Appointment	صحه	Health	يوم	Day
فيروس	Virus	تطعيم	Vaccination	صحه	Health
يتلقى	Get	يوم	Day	جرعات	Doses
حمدلله	Thank God	اولى	First	موجود	Exist
سلام	Peace	سلام	Peace	ثانيه	Second
تطبيق	Application	موعد	Appointment	حضور	Attendance
صحتي	My health	وزاره	Ministry	ثالثه	Third
مليون	Million	مملكه	Kingdom	تم	Done
نفسي	Myself	حمدلله	Thanks God	طلاب	Students
حمايه	Protection	مليون	Million	عام	Year
سجل	Recorded	جرعتين	Two doses	استكمال	Complete
وطني	My country	ضد	Against	اعمار	Ages
شركه	Company	ناس	People	طالبات	Students
نحمى	Protect	صيني	Chines	حصلو	Got
مجتمع	Community	مصر	Egypt	دراسه	Study
يوم	Day	كويت	Kuwait	تنشيطيه	Booster

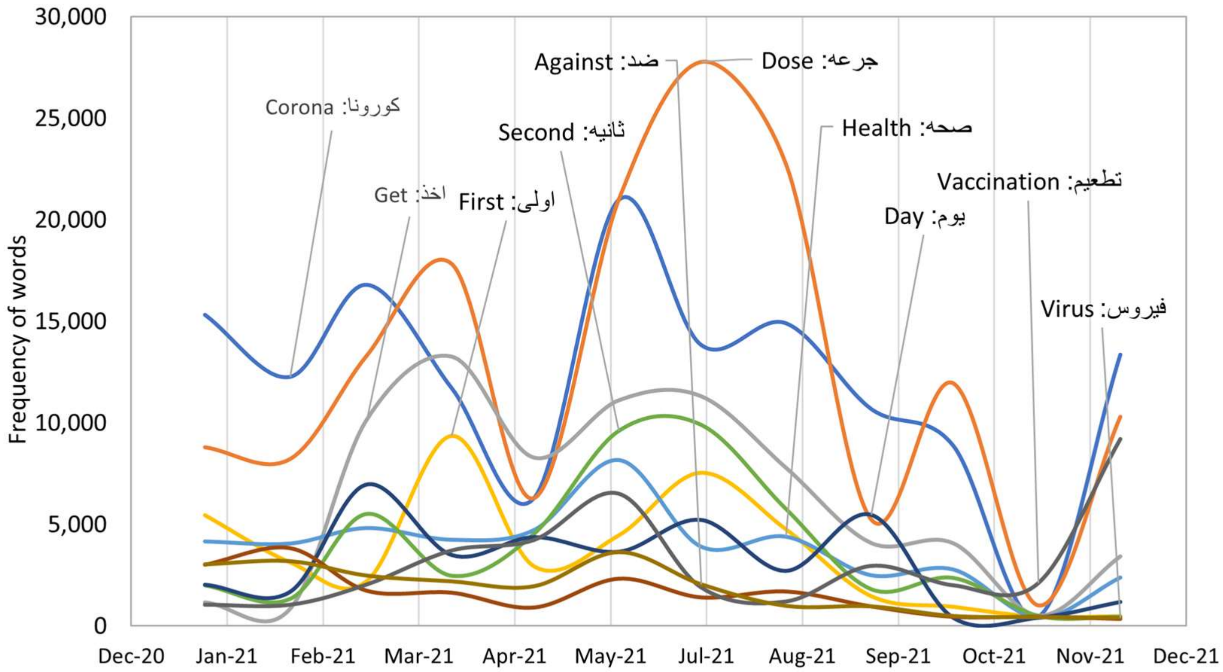

Figure 8. The top 10 most frequent terms over time.

In terms of validity, the data passed through a pipeline of processing and analysis. The initial exploratory analysis, including filtering and cleaning, is considered as a verifying step to fine-tune and filter out irrelevant tweets and keep the tweets that are related to the study. In addition, to avoid a bias in the collected data, we constructed a set of keywords by selecting a seed word and identifying the terms associated with this word, and then the selected keywords were varied by author. On top of that, the drawn result can give an indication of the validity of the analysis. The results and findings as graphs and tables verified that the tweets were a good representative sample of Arabic tweets related to the COVID-19 vaccination.

To demonstrate the importance of our dataset, we compared it with the ArCovidVac dataset [10] which is the only publicly available Arabic tweet dataset related to the COVID-19 vaccine. ArCovidVac used Twitter posts as the source of the data and focused on Arabic tweets as our dataset. However, they offer annotated datasets within different layers of annotation information, fine-grained content, and vaccination stance. That limited the use of the dataset in specific applications such as opinion analysis or misinformation detection. We offer a dataset that includes a collection of data that can be used in different applications. The dataset can be analyzed to explore the sentiments of people or to assess feelings, such as fear and panic. In addition, it can be analyzed in the long term to discover patterns or trends, as well as track vaccine misinformation and rumors. Moreover, it can be used to develop experiments that study changes in people's behaviors regarding the pandemic vaccination. The dataset can be used in different machine learning techniques for clustering and classification analysis.

5. Methods

This section describes the method that was used to develop the dataset. The framework for developing the dataset is shown in Figure 9. The tweets related to the COVID-19 vaccine were first collected and filtered to clean unnecessary tweets. After preprocessing and filtering, a preliminary analysis was performed. The dataset contains a collection of 1.1 M Arabic Twitter posts related to the COVID-19 vaccine.

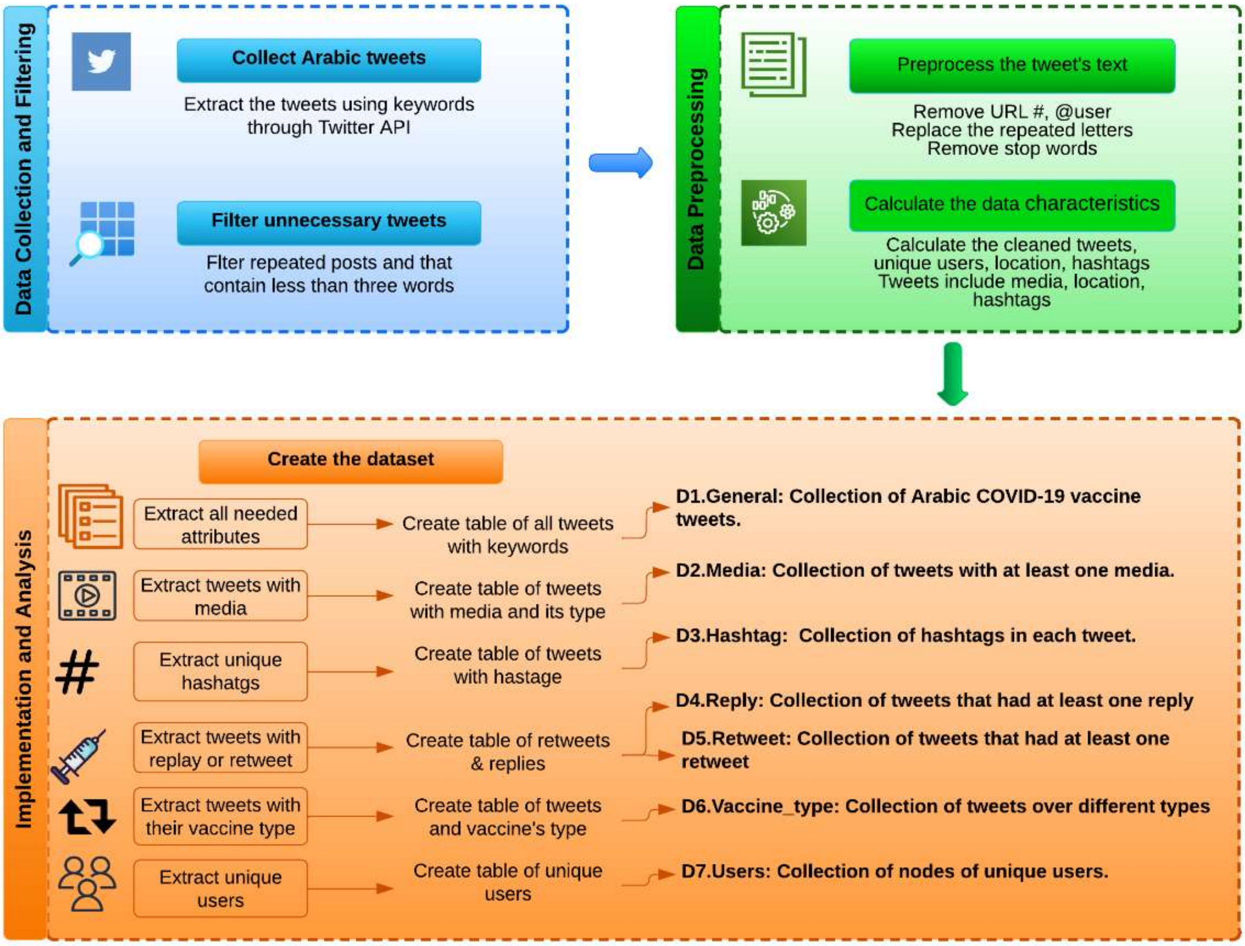
Data Collection and Filtering
Collect Arabic tweets
Extract the tweets using keywords through Twitter API
Filter unnecessary tweets
Flter repeated posts and that contain less than three words
Data Preprocessing
Preprocess the tweet's text
Remove URL #, @user
Replace the repeated letters
Remove stop words
Calculate the data characteristics
Calculate the cleaned tweets, unique users, location, hashtags
Tweets include media, location, hashtags
Implementation and Analysis
Create the dataset
Extract all needed attributes
Create table of all tweets with keywords
D1.General: Collection of Arabic COVID-19 vaccine tweets.
Extract tweets with media
Create table of tweets with media and its type
D2.Media: Collection of tweets with at least one media.
Extract unique hashatgs
Create table of tweets with hastage
D3.Hashtag: Collection of hashtags in each tweet.
Extract tweets with replay or retweet
Create table of retweets & replies
D4.Reply: Collection of tweets that had at least one reply
D5.Retweet: Collection of tweets that had at least one retweet
Extract tweets with their vaccine type
Create table of tweets and vaccine's type
D6.Vaccine_type: Collection of tweets over different types
Extract unique users
Create table of unique users
D7.Users: Collection of nodes of unique users.

Figure 9. The research framework.

5.1. Data Collection

The data regarding the COVID-19 vaccine was collected in the Arabic language over a year. The tweets were streamed for about twelve months from January 2021 to December 2021 using a list of Arabic crawling keywords. To select this list of keywords, we started with the keyword (لقاح) as the seed of our search. Then, we collected tweets containing this word. Next, from these collected tweets, we identified the co-occurrence words with this seed. Finally, these words were reviewed by the author and compared with the vaccines that have been used in Arabic-speaking countries, resulting with these keywords (كورونافاك, سينوفارم, سبوتنيك في, موديرنا, أسترازينكا, أكسفورد, فايزر, لقاح). The translations of these keywords are (Vaccine, Pfizer, Oxford, Astrazeneca, Sputink v, Moderna, Sinopharm, Coronavac). We collected the tweets using the "snscrape" Python package [36] that interacts with Twitter API. We searched for a set of Arabic keywords that collected a total of 1,125,446 tweets. We further filtered the tweets, resulting in 1,101,349. All the collected data is in the Arabic language since we restricted the search with the lang option "ar" to return only Arabic tweets. Figure 10 shows the distribution of the collected tweets associated with the keywords.

Table 5 presents the summary of the dataset. It shows that the total number of collected tweets is about 1.13 M posts and the volume of the data after filtering is about 1.10 M posted by 322,328 unique users. The table indicates that 26.22% of the tweets include hashtags, and 20.04% include media, with 3.35% videos and 16.94% photos. We found that 62.78% of the metadata have value in the location attribute; however, we noticed that some values in this field were non-standard locations.

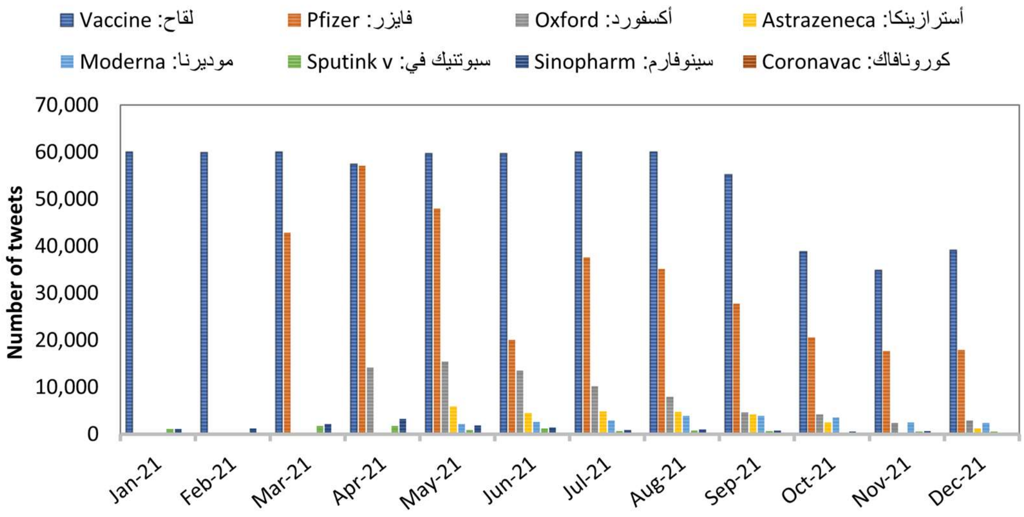

Figure 10. The distribution of the tweets by keyword.

Table 5. Summary of the dataset.

Tweets	Counts	Percentage
Collected tweets	1,125,446	Collected tweets
Filtered tweets	1,101,349	
Tweets with location	691,461	62.78%
Include hashtag	288,803	26.22%
Include media	220,797	20.045%
Include videos	36,863	3.347%
Include photos	186,571	16.94%
Unique hashtags	94,407	
Unique locations	58,785	
Unique users	344,328	

5.2. Data Preprocessing

Preprocessing the content of the tweets is an essential step in content and textual analysis. We conduct two phases of preprocessing, including the exclusion of irrelevant tweets and cleaning the text of the tweets. We excluded repeated posts and tweets that contain less the three words, due to insufficient content. Then, we carried out tweet text preprocessing. We first extracted the tweets' text for each month, then we performed cleaning steps on the tweets' text that are important for eliminating noisy, incomplete, and uninformative data. Therefore, we applied the following steps using the NLTK Python library [37]:

Remove the URL from the text
Remove the mentions (@user)
Remove the hashtags
Replace the repeated letters with one letter.
Remove stop words in the Arabic language such as pronouns, articles, prepositions, etc.
Remove punctuation such as commas, brackets, and full stops.
Replace emojis with special tokens.

5.3. Implementation

To develop the dataset, we conducted a series of computations and statistical analyses using Python language version 3.8. We started by partitioning the remainder of the collected data after the preprocessing by date. Then, we performed different types of analysis on the data. Firstly, we computed some characteristics of the data as shown in Table 3. Secondly, we carried out a content analysis, including trending hashtags, sharing media, the time series of retweets and replies, and the most active users. In the following, we will describe the implementation and the analysis in detail.

A hashtag is a phrase starting with a hash symbol (#); using hashtags on Twitter allows users to engage in conversations about specific topics or trends. Hence, we created a hashtag database called "D3.Hashtag" as described in Table 2. In this dataset, 26.22% of the tweets (about 288,803) include one or more hashtags. However, we filtered the unique hashtags, and 94,407 were used more than one time. To extract the most frequent hashtags, we counted the occurrence of each unique hashtag in the whole corpus, then we selected the top 50 hashtags.

Twitter allows users to embed media such as photos and videos in their posts, features that can increase the chance of the post being retweeted [38]. To identify the tweets that embed photos or videos, we created a database called "D2.Media", as shown in Table 2. We extracted the tweets that include media such as images or videos, then we filtered 223,434 tweets that contain at least one media. Next, we grouped each type of media in one cluster to find which is the most shared. To inspect the time series of sharing media, we counted their numbers each month.

The discussion on Twitter around different topics is enriched by a lot of interactions through retweets and replies to tweets. To create the retweet dataset, we extracted the tweets that had been retweeted, and then we counted the number of retweets for each tweet to construct the database called "D5.Retweet", as shown in Table 2. The same process was applied to create the database "D4.Reply". Both databases contain the date and time that the tweet was either retweeted or replied to.

The dataset includes 344,328 unique users who posted 1.10 M tweets. To investigate the more active users in terms of the number of tweets, we first extracted the unique users and created the database "D7.Users", as shown in Table 2. Then, we counted their posts through the entire period of our dataset. Subsequently, the 20 most active users were extracted and their countries located.

To look at the conversation taking place on Twitter regarding vaccination, we explored the content of the tweets by identifying the most frequent words. We excluded the set of the keywords that were used for extracting the corpus since it was expected to be highly frequent. To achieve this, we first counted the words in each month, then we extracted the words that occur more than 5000 times. Next, the top 20 terms were filtered for each month. Further, we selected the 10 most frequent terms over the whole period.

6. Potential Research Applications

Following the announcement of the distribution of the COVID-19 vaccine, people turned to social media to express their opinion, look for information, and report personal incidents. Therefore, this dataset offers a view of the dynamic of the vaccine conversation on Twitter. The dataset of 1.1 M Arabic posts is expected to help research in different fields such as data mining, natural language processing, social analysis, and healthcare. We expected that the data may be useful for several potential applications including but not limited to, social analytics, misinformation detection, and crisis management, as presented below:

Social analytics: social investigation analysis includes sentiment analysis, developing topic modeling, stance detection, monitoring retweet patterns, detecting sarcasm, hate speech, and many other applications.

Misinformation detection: with the fast development of the vaccine, many rumors are spread and have people's attention across the globe, which causes the spread of false information about vaccination. The dataset can support research on rumor detection, claims credibility, and community detection in retweets networks to identify fake news and vaccine stances.

Crisis Management: since the outbreak of the COVID-19 pandemic, health organizations around the world need to analyze information on the pandemic and gather information related to the disease and vaccination for analysis and thus obtain important insight. We believe our dataset can be useful in different tasks such as trend or event detection, content summarization, and information propagation. Insights and information detected in this situation can help in managing some potential crisis situation.

7. Conclusions

In this paper, we presented an Arabic Twitter dataset about the COVID-19 vaccine that offers the first look at the dynamics of vaccine conversation on Twitter. We performed a preliminary analysis that characterized the data in several ways, including trending hashtags, prominent terms, and time series of embedded media. Our results provide implications and insights into the dynamic of vaccine conversation on Twitter. First, the results show that hashtags are associated with a positive or negative stance on the vaccine, suggesting that we can understand people's opinions regarding vaccination distribution. Second, the timeline of retweeting and replying indicates how people are communicating about vaccination. Third, we can discover from the most dominant terms that the conversation continues to grow as the debate continues about vaccination. The dataset can be a source of data that facilitates research in different areas, including health informatics, topic modeling, natural language processing, social network analysis, information retrieval, and social science. In the future, we intend to update the data with newly collected tweets and to perform more in-depth analysis. We are currently working on investigating the content of the conversation to detect the side effects of different types of vaccines reported by people. We also plan to explore the relationship between the Twitter conversation and its effects on public health. Finally, we plan to model vaccination community detection.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are publicly available on Mendeley Data at: https://data.mendeley.com/datasets/zmwfnsms9n (accessed on 31 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sorensen, L. User managed trust in social networking—Comparing Facebook, MySpace and Linkedin. In Proceedings of the 2009 1st International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronics Systems Technology, Aalborg, Denmark, 17–20 May 2009; pp. 427–431. [CrossRef]
2. Kavada, A. Social Media as Conversation: A Manifesto. *Soc. Media Soc.* **2015**, *1*, 2056305115580793. [CrossRef]
3. Aslam, S. Twitter by the Numbers: Stats, Demographics & Fun Facts. Available online: https://www.omnicoreagency.com/twitter-statistics/ (accessed on 14 October 2022).
4. Aldekhyyel, R.N.; Binkheder, S.; Aldekhyyel, S.N.; Alhumaid, N.; Hassounah, M.; AlMogbel, A.; Jamal, A.A. The Saudi Ministries Twitter communication strategies during the COVID-19 pandemic: A qualitative content analysis study. *Public Health Pract.* **2022**, *3*, 100257. [CrossRef]
5. Michael, H. The use of Twitter by state leaders and its impact on the public during the COVID-19 pandemic. *Heliyon* **2020**, *6*, e05540. [CrossRef]
6. Roy, M.; Moreau, N.; Rousseau, C.; Mercier, A.; Wilson, A.; Atlani-Duault, L. Ebola and Localized Blame on Social Media: Analysis of Twitter and Facebook Conversations During the 2014–2015 Ebola Epidemic. *Cult. Med. Psychiatry* **2020**, *44*, 56–79. [CrossRef]
7. Kagashe, I.; Yan, Z.; Suheryani, I. Enhancing Seasonal Influenza Surveillance: Topic Analysis of Widely Used Medicinal Drugs Using Twitter Data. *J. Med. Internet Res.* **2017**, *19*, e315. [CrossRef] [PubMed]
8. Muñoz-Sastre, D.; Rodrigo-Martín, L.; Rodrigo-Martín, I. The Role of Twitter in the WHO's Fight against the Infodemic. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11990. [CrossRef] [PubMed]
9. Abbas, A.; Eliyana, A.; Ekowati, D.; Saud, M.; Raza, A.; Wardani, R. Data set on coping strategies in the digital age: The role of psychological well-being and social capital among university students in Java Timor, Surabaya, Indonesia. *Data Brief* **2020**, *30*, 105583. [CrossRef]

10. Polack, F.P.; Thomas, S.J.; Kitchin, N.; Absalon, J.; Gurtman, A.; Lockhart, S.; Perez, J.L.; Pérez Marc, G.; Moreira, E.D.; Zerbini, C.; et al. Safety and Efficacy of the BNT162b2 mRNA COVID-19 Vaccine. *New Engl. J. Med.* **2020**, *383*, 2603–2615. [CrossRef]
11. Covid-19: Pfizer/BioNTech Vaccine Judged Safe for Use in UK. Available online: https://www.bbc.com/news/health-55145696. (accessed on 9 September 2022).
12. Kim, J.H.; Marks, F.; Clemens, J.D. Looking beyond COVID-19 vaccine phase 3 trials. *Nat. Med.* **2021**, *27*, 205–211. [CrossRef]
13. Tekumalla, R.; Banda, J.M. A Large-Scale Twitter Dataset for Drug Safety Applications Mined from Publicly Existing Resources. *arXiv* **2020**, arXiv:2003.13900v1. [CrossRef]
14. Stemmer, M.; Parmet, Y.; Ravid, G. What Are IBD Patients Talking about on Twitter? In *ICT for Health, Accessibility and Wellbeing*; Springer International Publishing: Cham, Switzerland, 2021; pp. 206–220.
15. Saniei, R.; Rodríguez Doncel, V. PHDD: Corpus of Physical Health Data Disclosure on Twitter during COVID-19 Pandemic. *SN Comput. Sci.* **2022**, *3*, 212. [CrossRef] [PubMed]
16. Singh, L.; Bansal, S.; Bode, L.; Budak, C.; Chi, G.; Kawintiranon, K.; Wang, Y. A first look at COVID-19 information and misinformation sharing on Twitter. *arXiv* **2020**, arXiv:2003.13907v1.
17. Chen, E.; Lerman, K.; Ferrara, E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill.* **2020**, *6*, e19273. [CrossRef] [PubMed]
18. Aguilar-Gallegos, N.; Romero-García, L.E.; Martínez-González, E.G.; García-Sánchez, E.I.; Aguilar-Ávila, J. Dataset on dynamics of Coronavirus on Twitter. *Data Brief* **2020**, *30*, 105684. [CrossRef]
19. Haouari, F.; Hasanain, M.; Suwaileh, R.; Elsayed, T. ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks. *arXiv* **2021**, arXiv:2004.05861v4 2021. [CrossRef]
20. Elhadad, M.K.; Li, K.F.; Gebali, F. COVID-19-FAKES: A Twitter (Arabic/English) dataset for detecting misleading information on COVID-19. In *Advances in Intelligent Networking and Collaborative Systems*; Barolli, L., Li, K.F., Miwa, H., Eds.; Springer: Cham, Switzerland, 2021; Volume 1263, pp. 256–268. [CrossRef]
21. Hayawi, K.; Shahriar, S.; Serhani, M.A.; Taleb, I.; Mathew, S.S. ANTi-Vax: A novel Twitter dataset for COVID-19 vaccine misinformation detection. *Public Health* **2022**, *203*, 23–30. [CrossRef] [PubMed]
22. Memon, S.A.; Carley, K.M. Characterizing COVID-19 misinformation communities using a novel twitter dataset. *arXiv* **2020**, arXiv:2008.00791. [CrossRef]
23. Mubarak, H.; Hassan, S.; Chowdhury, S.; Alam, F. ArCovidVac: Analyzing Arabic Tweets about COVID-19 Vaccination. *arXiv* **2022**, arXiv:2201.06496. [CrossRef]
24. Hu, T.; Wang, S.; Luo, W.; Zhang, M.; Huang, X.; Yan, Y.; Liu, R.; Ly, K.; Kacker, V.; She, B.; et al. Revealing public opinion towards COVID-19 vaccines with Twitter Data in the United States: A spatiotemporal perspective. *J. Med. Internet Res.* **2021**, *23*, e30854. [CrossRef]
25. Malagoli, L.G.; Stancioli, J.; Ferreira, C.H.G.; Vasconcelos, M.; da Silva, A.P.C.; Almeida, J.M. A look into COVID-19 vaccination debate on Twitter. In Proceedings of the 13th ACM Web Science Conference 2021, New York, NY, USA, 21–25 June 2021; p. 225e33. [CrossRef]
26. Alshaabi, T.; Dewhurst, D.R.; Minot, J.R.; Arnold, M.V.; Adams, J.L.; Danforth, C.M.; Dodds, P.S. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. *arXiv* **2021**, arXiv:2003.03667. [CrossRef]
27. Abdul-Mageed, M.; Elmadany, A.; Nagoudi, E.M.B.; Pabbi, D.; Verma, K.; Lin, R. Mega-Cov: A billion-scale dataset of 100+ languages for covid-19. *arXiv* **2021**, arXiv:2005.06012. [CrossRef]
28. Haouari, F.; Hasanain, M.; Suwaileh, R.; Elsayed, T. ArCOV19-Rumors: Arabic COVID-19 Twitter Dataset for Misinformation Detection. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, Kyiv, Ukraine, 9 April 2021; pp. 72–81.
29. Alam, F.; Shaar, S.; Dalvi, F.; Sajjad, H.; Nikolov, A.; Mubarak, H.; Martino, G.D.S.; Abdelali, A.; Durrani, N.; Darwish, K.; et al. Fighting the COVID-19 Infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 611–649. [CrossRef]
30. Yang, Q.; Alamro, H.; Albaradei, S.; Salhi, A.; Lv, X.; Ma, C.; Zhang, X. Senwave: Monitoring the global sentiments under the covid-19 pandemic. *arXiv* **2020**, arXiv:2006.10842. [CrossRef]
31. Alsudias, L.; Rayson, P. COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? In Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Online, 9–10 July 2020.
32. Zhou, X.; Mulay, A.; Ferrara, E.; Zafarani, R. ReCOVery: A multimodal repository for COVID-19 news credibility research. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, New York, NY, USA, 19–23 October 2020. [CrossRef]
33. Murić, G.; Wu, Y.; Ferrara, E. COVID-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Dataset of Anti-vaccine Content, Vaccine Misinformation and Conspiracies. *JMIR Public Health Surveill.* **2021**, *7*, e30642. [CrossRef] [PubMed]
34. DeVerna, M.R.; Pierri, F.; Truong, B.T.; Bollenbacher, J.; Axelrod, D.; Loynes, N.; Bryden, J. CoVaxxy: A Collection of English-Language Twitter Posts About COVID-19 Vaccines. In Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 6–9 June 2021; Volume 15, pp. 992–999.
35. Twitter. Developer Agreement and Policy. Available online: https://developer.twitter.com/en/developer-terms/agreement-and-policy (accessed on 9 September 2022).

36. Pypi, Snscrape. Available online: https://pypi.org/project/snscrape/ (accessed on 9 September 2022).
37. NLTK. Available online: https://github.com/linuxscout/pyarabic (accessed on 14 October 2022).
38. Christodoulou, G.; Georgiou, C.; Pallis, G. The Role of Twitter in YouTube Videos Diffusion. In *Web Information Systems Engineering–WISE 2012. WISE 2012. Lecture Notes in Computer Science*; Wang, X.S., Cruz, I., Delis, A., Huang, G., Eds.; Springer: Heidelberg/Berlin, Germany, 2012; Volume 7651. [CrossRef]