ArCovidVac: Analyzing Arabic Tweets About COVID-19 Vaccination

Hamdy Mubarak, Sabit Hassan,
Shammur Chowdhury, Firoj Alam

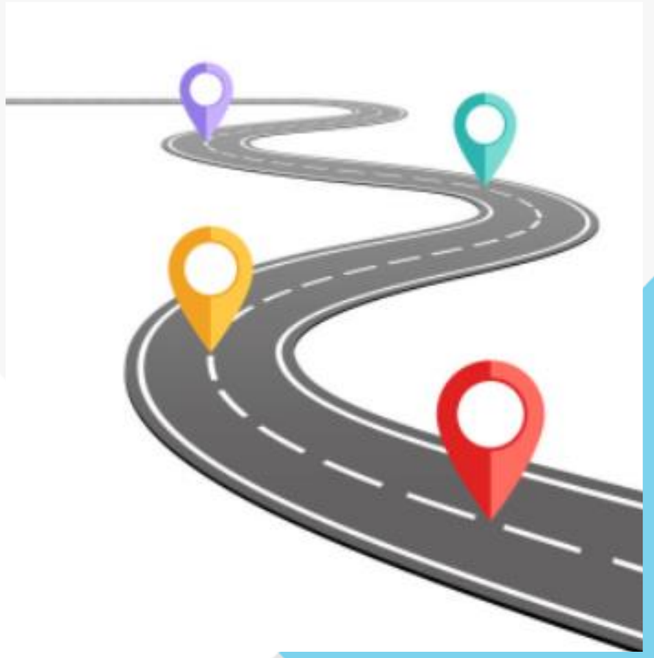
Arabic Language Technologies, QCRI

LREC 2022, June 20-25, Marseille, France

QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

Overview

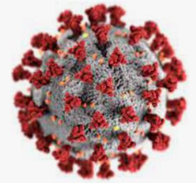
1. Introduction
2. Related Work
3. Data Collection
4. Data Annotation
5. Analysis
6. Experiments
7. Conclusion



01

Introduction

Introduction



- COVID-19 pandemic is the first global infodemic that changed our lives in many different ways

- Users share information on social media about COVID-19
  (e.g., rumors, vaccination plan, travel restrictions, personal experience, etc.)

- Identify content type is important to the government, international and local organizations

- Understanding users can:
  - aid decision making by governments, and
  - prevent rumors and fake cures that can bring harm to the society

02

Related Work

Related Work

- English Datasets
  - Many available and large datasets related to COVID-19 and its vaccine for: sentiment, misinformation, stances, factuality, hate speech, etc.

- Arabic Datasets

The manually labeled datasets are relatively few or small, ex:

- 2,000 tweets annotated for rumor detection based on posts by the MoH in SA (Al-sudias & Rayson, 2020)
- CheckThat! Lab for disinformation, factuality, check-worthiness and harmfulness of tweets (200 tweets) (Alam et al. 2021)
- 8,000 tweets collected from the early days of COVID-19 labeled for different types of content such as report, advice, seek action, etc. (Mubarak and Hassan, 2021)

✓ We present the first dataset about COVID-19 vaccine in Arabic with diverse type of annotations

03

Data Collection

Data Collection

- Collect tweets using Twitter API, keywords: لقاح، تطعيم، مطعوم (vaccine, vaccination)
- Collection timeline: Jan 5th and Feb 3rd 2021
- Many Arab countries already started COVID-19 vaccination campaigns
  Ex: Vaccine rollout in Saudi Arabia (SA)* started in mid Dec 2020

- Total: 550K unique tweets
- Consider “important” tweets: liked or retweeted at least 10 times => 14K tweets
- Tweets with large number of likes/retweets are the most important ones (highest attention from users)

- 10K tweets were randomly chosen for manual annotation

* ISO 3166-1 alpha-2 for country codes

04

Data Annotation

Data Annotation

- Appen crowdsourcing platform
- 3 annotations for each tweet
- QC: 150 test questions, 70% success threshold

- Types are identified based on engagement with MoH and policymakers

- Annotations:
  - Content type (10 categories)
  - Stance towards vaccine: +ve, -ve, neutral

- Cohen's kappa coefficient = 0.82 (high annotation quality)

Fine-grained Content Types:

1. **Info-news:** Information and news about vaccine and conditions of taking
2. **Celebrity:** Vaccination of celebrities such as politicians, artists, and public figures
3. **Plan:** Governments' vaccination plans, vaccination progress and reports
4. **Requests:** Requests from governments, e.g., speedup vaccination process
5. **Rumors:** Rumors and refute rumors
6. **Advice:** Advice or instructions related to the virus or its vaccination
7. **Restrictions:** Restrictions and issues that will be affected by taking vaccine, e.g., travel

8. **Personal:** Personal story or opinion about the vaccine, e.g., thank government
9. **Unrelated:** Unrelated to vaccination process. This includes also spam and ads
10. **Others:** Related to vaccine but not listed in the above classes

Informativeness

- More informative
- Less informative

Data Annotation

Stance: For identifying stance, we use the following labels:

- **Positive:** Support vaccination, encourage people to take vaccine, and remove their fears.
  Example: متحدث الصحة: المشككون في فعالية لقاح كورونا سوف يأتون لأخذ اللقاح
  Health spokesperson: Those who doubt the effectiveness of the Corona vaccine will come and get it
- **Negative:** Oppose vaccination and fear people from vaccine.
  Example: قلق بالغ في النرويج بسبب وفاة ٢٣ شخصا بعد تلقيهم لقاح فايزر
  Extreme concern in Norway because 23 people have died after receiving the Pfizer vaccine
- **Neutral/Unclear:** Neither clearly support nor oppose vaccination.
  Example: توتر العلاقات بعد رفض بريطانيا تسليم فرنسا ١٥ مليون من لقاح كورونا
  Relations are strained after Britain refused to deliver 15 million doses of the Corona vaccine to France

(a) Celebrity, info-news, plan, requests and advice

(b) Rumors, restrictions, others, unrelated and personal

Figure 1: Examples of tweets reporting different fine-grained categories

Data Annotation

Class	Count	Class	Count
Fine-grained		**Informativeness**	
Info-news	5,225	More Informative	7,891
Celebrity	1,398	Less Informative	2,109
Plan	860	**Total**	10,000
Requests	172	**Stance**	
Rumors	118	Positive	7,968
Advice	94	Negative	636
Restrictions	24	Neutral/Unclear	1,396
Personal	1,430	**Total**	10,000
Unrelated	450		
Others	229		
Total	10,000		

Table 1: Distribution of the annotated class labels.

05

Analysis

Analysis

- Vaccine Popularity

Vaccine	Top Hashtags	#	CC
Pfizer	فايزر، لقاح_فايزر، بيونتيك، بايونتيك, Pfizer	184	US
AstraZeneca	استرازينيكا، اوكسفورد، استرازينكا	94	UK
Sputnik V	سبوتنيك، سبوتنيك5	65	RU
Moderna	موديرنا، مودرنا, Moderna	43	US
BBIBP-CorV	سينوفارم, Sinopharm	24	CN
CoronaVac (Sinovac)	سينوفاك، كورونافاك	10	CN
Johnson & Johnson	جونسون، جونسون_اند_جونسون	5	US
Novavax	نوفافاكس	2	US

Table 2: Vaccine hashtag frequencies (# represent the number of times they are found in the corpus). Arabic hashtags are mainly different transliterations of vaccine names. CC: Country Code of the manufacturing company.

- Trending Hashtags

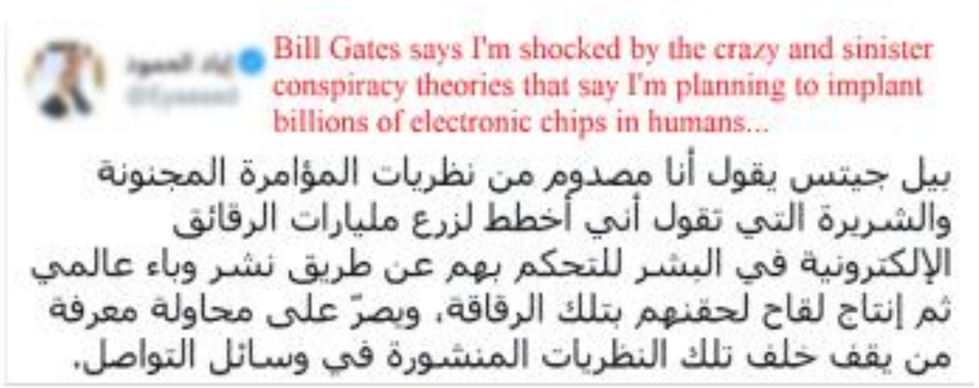
Country	Hashtags	Translation	#
IQ	نريد_لقاح_آمن	We want a safe vaccine	288
SA	الملك_يتلقي_لقاح_كورونا، نعود_بحذر	The king takes COVID vaccine, We return cautiously	174
LB	لقاح_آمن، خليك_بالبيت	Safe vaccine, Stay home	157
AE	يدا_بيد_نتعافى، اخترت_التطعيم	Hand in hand we recover, I chose vaccination	151
EG	معا_نطمئن	Together we can rest assured	7
MA	نبقاو_على_بال	We remain alert	7
OM	عمان_تواجه_كورونا، التحصين_وقاية	Oman fights Corona, Vaccination is protection	6
JO	المطعوم_وقاية، صحتك_بتهمنا	Vaccine is protection, Your health is important to us	5

Table 3: Most frequent hashtags in some Arab countries.

Analysis

- Rumors

- **Vaccine is unsafe and ineffective:** (i) causes death and has side effects especially on elderly; (ii) manipulates genes; (iii) causes infertility in women.
- **Conspiracy theory:** (i) big countries or companies created the virus and its vaccine for commercial purposes; (ii) vaccine has chips to monitor and control people; (iii) vaccine is a biological weapon; (iv) question about finding vaccines within a year. Figure 2(a) shows the most retweeted and targeted tweet in this category.
- **Doubts** about government statistics, plans, and vaccination process.

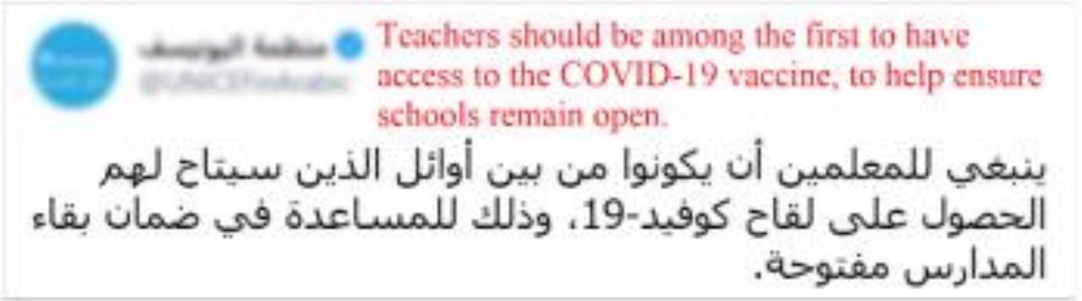


(a) Conspiracy theory

Analysis

• Requests from Governments

- **Safe vaccine:** (i) wait until studies and other countries prove vaccine effectiveness and safety; (ii) prefer US vaccines over their Chinese counterparts; (iii) refuse vaccine from the US (especially in Iraq).
- **Fair access to vaccine:** (i) rich and poor countries and people; (ii) males and females; (iii) citizens, expats and refugees; (iv) cities and regions in the same country; (v) politicians and common people; (vi) Israel and Palestinians.
- **Vaccination process:** (i) speedup; (ii) transparency in plans and contract details; (iii) finding alternative companies and cheaper vaccines; (iv) allow private sector to sell vaccines.
- **Give priority:** to some professionals such as doctors, teachers, players, and natives. Figure 2(b) shows one of the most common tweets that asks to give priority to the teaching professionals.



(b) A request to give priority to the teaching professional.

Analysis

- Vaccine Announcements

- Countries: TR, SA, EG and IR
- None of those vaccines was used in any Arab countries until the date of our study

Analysis

- Top sources of news

(mainly news sources and agencies)

- Mobile Applications

(show health status, report violations of precautionary measures, book medical services, track medicines, facilitate travel/visa process)

CC	%	Top Accounts
SA	25	sabqorg, Akhbaar24, KSA24, ajlnews
AE	14	cnnarabic, AlArabiya_Brk, skynewsarabia, AlHadath
LB	11	AlMayadeenNews, ALJADEEDNEWS, JamalCheaib
EG	8	youm7, AlMasryAlYoum, RassdNewsN, Extranewstv
GB	5	aawsat_News, AlarabyTV, IndyArabia, Mhd_AlObaidi
KW	5	liferdefempire, WhistleBlowerQ8, gucciya234, TfTeeeSH
JO	4	AlMamlakaTV, alrai, khaberni, RoyaTV
TR	4	TRTArabi, aa_arabic, TurkPressMedia, YeniSafakArabic
DZ	3	ennaharonline, El_Bilade, radioalgerie_ar, elkhabarlive
RU	3	RTarabic, RTarabic_Bn

Table 4: Distribution of top accounts across different countries. CC: Country Code.

Application (and meaning)	Arabic Name	CC	Date	#	DL
Tawakkalna (We Trust in God)	توكلنا	SA	May'20	35	10M
Sehhaty (My Health)	صحتي	SA	Dec'20	32	5M
Kuwait Mosafer (Kuwait Traveller)	كويت مسافر	KW	Feb'21	3	5K
DHA (Dubai Health Authority)	صحة دبي	AE	Dec'20	2	500K
Al Hosn UAE (The Fort)	الحصن	AE	Apr'20	1	1M

Table 5: Applications used to fight COVID-19 in some Arab countries. DL: Downloads at Google Store in May'20.

Analysis

- Stance timeline

Figure 4: Distribution of *stance* towards vaccine over time. pro: positive stance, anti: negative stance.

06

Experiments

Experiments

- BERT models (AraBERT and QARiB) outperform SVMs significantly
- Most errors stem from Plan class misclassified as Info-news

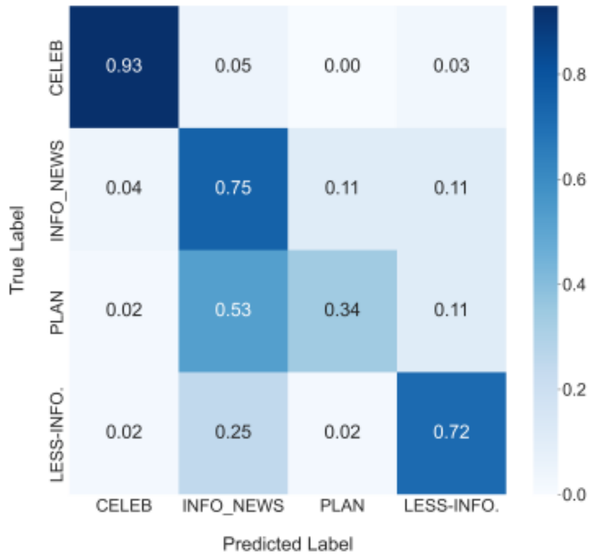


Figure 5: Confusion matrix of fine-grained classification normalized over true labels.

Model	Features	Acc.	P	R	F1
Informativeness (binary)					
Majority		79.5	39.8	50.0	44.3
SVM	W[1-3]	84.0	75.7	73.1	74.3
SVM	C[2-7]	84.9	77.6	72.9	74.8
SVM	C[2-7] + W[1-3]	84.6	76.8	73.0	74.6
QARiB		86.0	78.4	80	79.1
AraBERT		**86.4**	**78.9**	**81.3**	**80.0**
Fine-grained categorization (multiclass)					
Majority		54.4	13.6	25.0	17.6
SVM	W[1-3]	70.2	66.4	57.9	59.0
SVM	C[2-7]	71.6	66.7	58.0	58.8
SVM	C[2-7] + W[1-3]	72.0	68.7	59.3	60.5
QARiB		72.1	66.2	**68.2**	**67.1**
AraBERT		**75.4**	**69.2**	65.1	64.3
Stance Detection (multiclass)					
Majority		81.0	27.0	33.3	29.8
SVM	W[1-3]	81.6	60.8	48.6	52.1
SVM	C[2-7]	82.5	65.8	47.9	52.3
SVM	C[2-7] + W[1-3]	**82.5**	62.6	47.7	51.4
QARiB		81.6	**64.3**	62.7	**63.1**
AraBERT		82.2	61.0	**65.1**	62.5

Table 7: Results for different classification tasks.

07

Conclusion

Conclusion

- We present the first large manually annotated Arabic tweet dataset for COVID-19 vaccines
- 10k tweets covering many Arab countries
- Annotations:
  - informativeness of the tweets,
  - fine-grained tweet content types with 10 classes, and
  - stance towards vaccine

- In-depth analysis of the dataset to consider different aspects

- Download link: https://alt.qcri.org/resources/ArCovidVac.zip