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
Overview

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

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: المسكون في مجالة لقاح كورونا سوف يأتي. أحد 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 23 شخصًا 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 15 مليون لقاح كورونا
Relations are strained after Britain refused to deliver 15 million doses of the Corona vaccine to France

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

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

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

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

Table 1: Distribution of the annotated class labels.
05

Analysis
Analysis

- Vaccine Popularity

<table>
<thead>
<tr>
<th>Vaccine</th>
<th>Top Hashtags</th>
<th>#</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfizer</td>
<td>Pfizer, فايزر, بيونتيك, بایوتنیک</td>
<td>184</td>
<td>US</td>
</tr>
<tr>
<td>AstraZeneca</td>
<td>استرازینجا, اوکسفورد, استارزنیکا</td>
<td>94</td>
<td>UK</td>
</tr>
<tr>
<td>Sputnik V</td>
<td>سبوتنيک, سبوتنيک5</td>
<td>65</td>
<td>RU</td>
</tr>
<tr>
<td>Moderna</td>
<td>مودرنا, مودرنایا</td>
<td>43</td>
<td>US</td>
</tr>
<tr>
<td>BBIBP-CorV</td>
<td>Sinopharm, سینوفارم</td>
<td>24</td>
<td>CN</td>
</tr>
<tr>
<td>CoronaVac (SinoVac)</td>
<td>سینوفارم، کورونافاکس</td>
<td>10</td>
<td>CN</td>
</tr>
<tr>
<td>Johnson &amp; Johnson</td>
<td>جونسون، جونسون، جونسون</td>
<td>5</td>
<td>US</td>
</tr>
<tr>
<td>Novavax</td>
<td>نوفاکوکس</td>
<td>2</td>
<td>US</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Country</th>
<th>Hashtags</th>
<th>Translation</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>IQ</td>
<td>نريد _نفاح _آمن</td>
<td>We want a safe vaccine</td>
<td>288</td>
</tr>
<tr>
<td>SA</td>
<td>الله _لنا يفني _نفاح _كرونا، تمود _يضر</td>
<td>The king takes COVID vaccine, We return cautiously</td>
<td>174</td>
</tr>
<tr>
<td>LB</td>
<td>نفاح _آمن، خليك _بالبيت</td>
<td>Safe vaccine, Stay home</td>
<td>157</td>
</tr>
<tr>
<td>AE</td>
<td>يدا _في _اليد _نتجز، يمتحن _الإمام</td>
<td>Hand in hand we recover, I choose vaccination</td>
<td>151</td>
</tr>
<tr>
<td>EG</td>
<td>معا _الطم</td>
<td>Together we can rest assured</td>
<td>7</td>
</tr>
<tr>
<td>MA</td>
<td>نفقوأ _على _بال</td>
<td>We remain alert</td>
<td>7</td>
</tr>
<tr>
<td>OM</td>
<td>عمان _نواجه _كرونا، التحصين _نوفا</td>
<td>Oman fights Corona, Vaccination is protection</td>
<td>6</td>
</tr>
<tr>
<td>JO</td>
<td>اللعوم _نوفا، صحتك _يهمنا</td>
<td>Vaccine is protection, Your health is important to us</td>
<td>5</td>
</tr>
</tbody>
</table>

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.
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)

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

<table>
<thead>
<tr>
<th>Application (and meaning)</th>
<th>Arabic Name</th>
<th>CC</th>
<th>Date</th>
<th>#</th>
<th>DL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tawakkalna (We Trust in God)</td>
<td>توكا</td>
<td>SA</td>
<td>May’20</td>
<td>35</td>
<td>10M</td>
</tr>
<tr>
<td>Selhaty (My Health)</td>
<td>صحتي</td>
<td>SA</td>
<td>Dec’20</td>
<td>32</td>
<td>5M</td>
</tr>
<tr>
<td>Kuwait Mosafir (Kuwait Traveller)</td>
<td>كويت مسافر</td>
<td>KW</td>
<td>Feb’21</td>
<td>3</td>
<td>5K</td>
</tr>
<tr>
<td>DHA (Dubai Health Authority)</td>
<td>صحة دبي</td>
<td>AE</td>
<td>Dec’20</td>
<td>2</td>
<td>500K</td>
</tr>
<tr>
<td>Al Hosa UAE (The Fort)</td>
<td>الخصن</td>
<td>AE</td>
<td>Apr’20</td>
<td>1</td>
<td>1M</td>
</tr>
</tbody>
</table>

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.
Experiments
Experiments

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

![Confusion matrix](image)

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

**Table 7:** Results for different classification tasks.
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