Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3220–3230
Marseille, 20-25 June 2022

ArCovidVac: Analyzing Arabic Tweets About COVID-19 Vaccination

Hamdy Mubarak[1], Sabit Hassan[2], Shammur Absar Chowdhury[1] and Firoj Alam[1]
[1]Qatar Computing Research Institute, HBKU, Qatar
[2]University of Pittsburgh, USA
{hmubarak, shchowdhury, fialam}@hbku.edu.qa, sah259@pitt.edu

Abstract

The emergence of the COVID-19 pandemic and the *first global infodemic* have changed our lives in many different ways. We relied on social media to get the latest information about COVID-19 pandemic and at the same time to disseminate information. The content in social media consisted not only health related advise, plans, and informative news from policymakers, but also contains conspiracies and rumors. It became important to identify such information as soon as they are posted to make an actionable decision (e.g., debunking rumors, or taking certain measures for traveling). To address this challenge, we developed and publicly released the first largest manually annotated Arabic tweet dataset, *ArCovidVac*, for the COVID-19 vaccination campaign, covering many countries in the Arab region. The dataset is enriched with different layers of annotation, including, *(i)* Informativeness (more vs. less important tweets); *(ii)* fine-grained tweet content types (e.g., advice, rumors, restriction, authenticate news/information); and *(iii)* stance towards vaccination (pro-vaccination, neutral, anti-vaccination). Further, we performed in-depth analysis of the data, exploring the popularity of different vaccines, trending hashtags, topics and presence of offensiveness in the tweets. We studied the data for individual types of tweets and temporal changes in stance towards vaccine. We benchmarked the ArCovidVac dataset using transformer models for informativeness, content types, and stance detection.

Keywords: COVID-19, Vaccination, Stance Detection, Arabic Tweets, Tweet Classification

1. Introduction

Social media are integrated with our daily life. We share and access information through social media platforms making it the most prominent form of communication. Due to its reach to a larger and international population, many organizations and individuals use them to circulate their contents. Thus also making these platforms a constitutive part of online news distribution and consumption (Mitchell and Page, 2014).

In Figure 1(a) and 1(b) we demonstrate how online users share information (e.g., rumors, plan, travel restriction, personal experience). The post containing advice is important to reduce the spread of COVID-19 as, for example, vaccinated people can be also a carrier. Identifying such types of content from social media become important to the government, international and local organisation for understanding psychological and physical well being along with public reactions to every taken actions. Such an understanding can *(i)* aid decision making by governments; and *(ii)* prevent rumours and fake cures that can bring harm to the society. Research studies have been conducted using numbers of COVID-19 datasets collected from Twitter. The research focused on: unlabeled (Chen et al., 2020; Banda et al., 2021; Alqurashi et al., 2020; Haouari et al., 2021a), automatically labeled (Abdul-Mageed et al., 2020; Qazi et al., 2020b), labeled using distant supervision (Cinelli et al., 2020; Zhou et al., 2020), and small manually annotated (Song et al., 2020; Vidgen et al., 2020; Shahi and Nandini, 2020; Pulido et al., 2020) datasets.

Despite Arabic being one of the dominant languages on Twitter (Alshaabi et al., 2021), a very few research targeted toward aiding governments and international organisations in their decision making and understanding public perspective towards the vaccine, in the Arab region.

To aid such decision making process, in this study, we designed and publicly released the largest manually annotated COVID-19 tweets regarding its vaccine and vaccination campaigns in the Arab region. Our contributions can be summarized as follows:

- We developed a large manually annotated COVID-19 vaccine infodemic dataset, covering different countries in the Arab region, targeted to aid the policymakers and the society as a whole. To the best of our knowledge, this is the first dataset about COVID-19 vaccine in Arabic with diverse type of annotations.
- We annotated the tweets with fine-grained content types (10 classes) including: plan, request, advice, restrictions, rumors, authenticate news or information, personal experience among others.
- We also annotated tweets specifying their stance towards vaccine/vaccination process. We classify them into positive (pro-vaccination), negative (against vaccination) or as neutral stance.
- We analysed the tweets for different labels and explore what topics the content covers, top hashtags in each country, common sources that users post their tweets in different countries, etc. Moreover, we analysed the temporal changes in public stance on vaccination over time.
- We benchmarked the released dataset for sev-

(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.

eral tasks. The benchmark experiments includes *(i)* discriminating informative tweets from not-informative ones; *(ii)* fine-grained multi-class tweet type classification; and *(iii)* stance detection using transformer architectures.

- Finally, we made the dataset freely available for research purposes only.[1]

The rest of the paper is organized as follows. Section 2 provides a brief overview of previous work. Sections 3 and 4 provides a detail about data collection and annotation procedures. An in depth analysis of the annotated dataset is provided in Section 5. Section 6 presents the experiments and classification results. Section 7 provides an analysis on the classification results. Finally, Section 8 concludes the paper with possible directions for future work.

2. Related Work

Research studies on COVID-19 focused on sentiment analysis (Yang et al., 2020), propagation of misinformation (Huang and Carley, 2020; Shahi et al., 2020), credibility check (Cinelli et al., 2020; Pulido et al., 2020; Zhou et al., 2020), detecting racial prejudices and fear (Medford et al., 2020; Vidgen et al., 2020) along with situational information, e.g., caution and advice (Li et al., 2020). Moreover, studies also include detecting mentions and stance with respect to known misconceptions (Hossain et al., 2020); determining the stance of each document with respect to the claim then making a prediction about the factuality of the claim (Baly et al., 2018); development of Arabic corpus containing true vs. false claims and claim-evidence pairs (Khouja, 2020); automatic generation of fake stories from true stories and models for detecting such fake stories (Nagoudi et al., 2020); automatically annotated (Arabic/English) COVID-19 tweets collected from well-known sources (e.g., UNICEF, and UN) and pre-checked verified facts from different fact-checking websites (Elhadad et al., 2020); and manually annotated Arabic tweets related to COVID-19 consisting of 138 verified claims from popular fact-checking websites and identified 9.4K relevant tweets to those claims (Haouari et al., 2021b). For COVID-19 related disinformation, factuality, check-worthiness and harmfulness of tweets, notable recent efforts include CheckThat! Lab initiatives (Nakov et al., 2021a; Nakov et al., 2021b; Shaar et al., 2021; Nakov et al., 2022).

These studies rely mostly on the social media content – mainly Twitter using queries or some distant supervision. Most of the large-scale COVID-19 datasets are unlabeled tweet collection, including multi-lingual dataset of 123M tweets (Chen et al., 2020), 152M tweets (Banda et al., 2021), a billion multilingual tweets (Abdul-Mageed et al., 2020) and GeoCoV19 (Qazi et al., 2020a) containing 524M tweets with their location information. In addition to the unlabeled data, some datasets are created using distant supervision (Cinelli et al., 2020; Zhou et al., 2020) and some are manually annotated (Song et al., 2020; Vidgen et al., 2020; Shahi and Nandini, 2020; Pulido et al., 2020).

For Arabic, we see a similar trend in developing datasets. The Arabic dataset, studied in (Alqurashi et al., 2020) provide a large dataset of Arabic tweets containing keywords related to COVID-19. Similarly, ArCOV-19 proposed in (Haouari et al., 2021a), contains 750K tweets obtained by querying Twitter. The manually labeled datasets are relatively fewer and also diversity of annotated labels is little to none. Authors in (Alam et al., 2021b; Alam et al., 2021a) manually annotated tweets in multiple languages for fact-checking, harmfulness to society, and the relevance of the tweets to governments or policymakers. Another study in (Alsudias and Rayson, 2020) collected 1M unique Arabic

[1] https://alt.qcri.org/resources/ArCovidVac.zip

tweets related to COVID-19 in the early 2020, among which a random 2000 tweets were annotated for rumor detection based on the tweets posted by the Ministry of Health in Saudi Arabia. Authors in (Mubarak and Hassan, 2021a) annotated 8K tweets collected from the early days of COVID-19 (top 200 retweeted tweets in 40 continuous days) and labeled them for different types of content such as report, advice, seek action, rumor, etc. In (Yang et al., 2020), author also annotated 10K Arabic and English tweets for the task of fine-grained sentiment analysis.

Our Dataset: Prior studies are mainly focused on one or two aspects of actionable information (e.g., factuality, or rumor detection). In comparison, our work is focused on *Informativeness*, *fine-grained content types* and *stance towards vaccination* with manual annotation of $10K$ Arabic tweets. Such a diversity of labels enables the community to design and develop models in a multitask learning setup (i.e., fine-grained content types and stance in one model).

In addition, we specifically developed the dataset targeting the vaccination campaigns in the Arab region, covering many countries. We also analysed different annotated classes, explored topics and temporal changes in stance regarding vaccines, and studied classification errors for the stance and tweet type classification.

3. Data Collection

Fighting the pandemic as well as infodemic requires to identify and understand the content shared on social media either to reduce the spread of harmful content, health related disinformation or to make an actionable decision (e.g., attention worthy content for policymakers) (Alam et al., 2021b; Alam et al., 2021a). Such an understanding can help in identifying concerns and rumors about vaccination, public sentiment, requests from governments and health organizations, etc, while facilitating policymaking. This is a challenging given that manually annotated language specific (e.g., Arabic) datasets are scarce.

To address this challenge, we collected Arabic tweets and manually annotated them. To collect the tweets we used the following keywords: تطعيم ، لقاح، مطعوم (vaccine, vaccination) between *Jan 5th* and *Feb 3rd 2021*.[2] We used twarc search API[3] to collect these tweets specifying Arabic language. Our data collection timeline coincides with the phase where many Arab countries already started their COVID-19 vaccination campaigns.[4] For example, Saudi Arabic (SA)[5] started vaccine rollout in the middle of Dec 2020.

We collected 550K unique tweets in total. After considering only tweets that were liked or retweeted at least 10 times, we ended up with 14K tweets. We assume that tweets with large number of likes or retweets are the most important ones as they get highest attention from Twitter users. Out of them, 10K tweets were randomly chosen for manual annotation.

[2]Words used in different countries in the Arab World.

[3]https://github.com/DocNow/twarc

[4]https://tinyurl.com/mtm4wtrh

[5]ISO 3166-1 alpha-2 for country codes: https://tinyurl.com/mubpbjx6

4. Data Annotation

4.1. Annotation Task and Labels

We manually analyzed random samples of selected tweets to understand the data at hand, and to design and define the annotation task and class labels. We identified different types of class labels based on our engagement with the ministry of public health and some policymakers. We manually annotated two categories: *fine-grained content types* and *stance* with respect to vaccine, while the *informativeness* type labels are inferred from fine-grained content types. Note that identifying informativeness type in the first place is helpful to reduce information overload for the decision-makers because social media content is higher in volume and it is important to filter/remove irrelevant or less important content. Below, we define the class labels with examples, which are given to the annotators as instructions.

Fine-grained Content Types (Class labels):

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: For informativeness, the former seven class labels from the fine-grained content types are considered as *more informative* and the later three class labels are considered as *less informative*.

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 effec-

tiveness 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

4.2. Manual Annotation

For the manual annotation, we opted to use Appen crowdsourcing platform[6]. One of the challenges with crowdsourced annotation is to find a large number of qualified workers while filtering out low-quality workers or spammers (Chowdhury et al., 2015; Chowdhury et al., 2014). To deal with this problem and to ensure the quality of the annotation we followed standard evaluation (Chowdhury et al., 2020b), i.e., we used 150 gold standard test tweets. Based on these gold standard test tweets, each annotator needed to pass at least 70% of the tweets to participate in the annotation task. Given that the content of the tweet is in Arabic, therefore, we only allowed Arabic speaking participants from all Arab countries. While annotators needed to pass such criteria in order to annotate each tweet, we also designed the annotation task to label each tweet by three annotators so that final label can be selected based on the majority agreement. In total, 245 annotators participated in the annotation task from different Arab countries.[7]

We selected the final label for each tweet based on the label agreement score[8] greater than or equal to 70%. In Table 1, we report the distribution of the dataset. As mentioned earlier, the class labels for *Informativeness* are inferred from fine-grained labels.

Annotation agreement: We computed the annotation agreement using Cohen's kappa coefficient, and found an agreement score of 0.82, which indicates high annotation quality.

5. Analysis

We conducted an in-depth analysis of the *ArCovidVac* dataset to understand (i) the vaccine popularity; (ii) the most common trending hashtags in the dataset for different countries; (iii) the common rumors and requests from governments; (iv) topic country distribution of tweets and the most common sources in each country; (v) the distribution of stance; and (vi) the popularity of mobile applications used to fight the spread of the virus. We believe this kind of analysis gives a broad understanding of how people are reacting to the vaccine campaigns in different Arab countries. In addition, it can also help in understanding the reasons behind the vaccine hesitancy. For the analysis of different aspects, we computed the frequency of their appearance in the dataset.

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.

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.

Vaccine Popularity: Table 2 shows the list of vaccine hashtags mentioned in the dataset. This can give a rough estimate about vaccine popularity in the Arab countries during the period of our study. More details about these vaccines can be found at: https://en.wikipedia.org/wiki/COVID-19_vaccine

Trending Hashtags: The most frequent hashtags in different countries are listed in Table 3. The main messages in these hashtags show worry from vaccination, advice to take precautionary measures, and reassure people that vaccine is safe.

Rumors: Identifying rumors are very important and requires more attention from governments and policymakers. False claims about vaccines can negatively affect public trust in vaccination campaigns. This may

[6]www.appen.com

[7]We paid more than $15 per hour of work to conform to the minimum wage rate in US.

[8]https://success.appen.com/hc/en-us/articles/360038386492-How-to-Calculate-Overall-Unit-Agreement

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 _يتهمنا	Vaccine is protection, Your health is important to us	5

Table 3: Most frequent hashtags in some Arab countries.

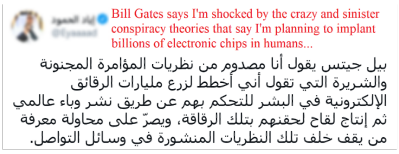
cause a threat to global public health. We analyzed all rumors in our dataset and classified them into the following main topics:

- **Vaccine is unsafe and ineffective:** (i) causes death and has side effects especially on elderly; (ii) manipulates genes; and (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; and (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.

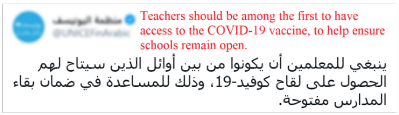
Requests from Governments: We analyzed all requests from governments and classified their main topics into the following classes:

- **Safe vaccine:** (i) wait until studies and other countries prove vaccine effectiveness and safety; (ii) prefer US vaccines over their Chinese counterparts; and (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; and (vi) Israel and Palestinians.
- **Vaccination process:** (i) speedup; (ii) transparency in plans and contract details; (iii) finding alternative companies and cheaper vaccines; and (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.

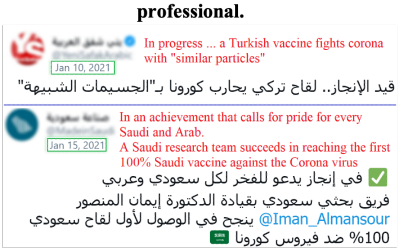
Vaccine Announcements: We spotted many news, posted in Jan 2020, about successful vaccines coming from research labs in different countries in the MENA region (e.g., TR, SA, EG, and IR), but in reality none of those vaccines was used in any Arab countries until the date of our study. Examples of such announcements are shown in Figure 2(c). We suspect these news were posted for political or social purposes.



(a) Conspiracy theory



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



(c) Vaccine announcements in Turkey and Saudi Arabia

Figure 2: Examples of tweets reporting conspiracy theory, request and vaccine announcements.

Topic and Country Distribution: We took a random sample of 1,000 tweets and manually categorized them for their main topics, such as health, politics, society and economy. Additionally for all tweets, we used ASAD[9] (Hassan et al., 2021), which achieves 88.1% F1 score on the UL2C dataset (Mubarak and Hassan, 2021b) for country prediction of original tweet authors based on their user locations in their profiles. Figure 3 shows that in addition to the health topics in most of the tweets, one third of tweets talk about the vaccine from different aspects (e.g., attacking politicians or countries). We also found that 7% of tweets have hate speech or offensive language.

Country distribution of tweets and top accounts that users share their posts the most in each country are shown in Table 4. Analysis of such accounts shows that people retweet posts mainly from online news agencies and newspapers in their countries, and less from some journalists or activists. Most of those accounts are verified. Surprisingly, accounts of ministries of health were not among the top four sources in the listed countries. We anticipate one reason for that might be due to the

[9]https://asad.qcri.org/

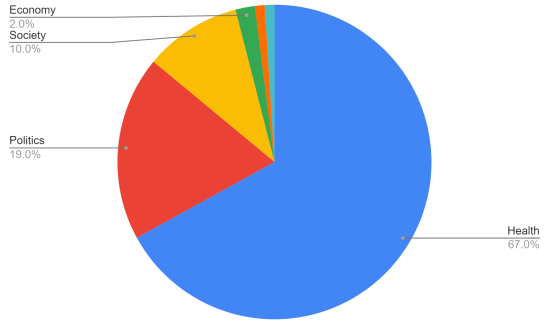


Figure 3: Distribution of *topics* labeled on a random sample of 1,000 tweets.

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.

less amount of posts from ministries of health compared to the large volumes of tweets that come from news agencies and newspapers.

Distribution of Stance: Figure 4 shows timeline of stance towards vaccine during the period of our study. We observe a big increase of positive stance (pro vaccine) in Jan 8th when media announced that the king of Saudi Arabia took the vaccine. This can show the effect of sharing news about celebrity vaccination on public opinion. On the opposite side, we found an increase of negative stance (anti vaccine) in Jan 12th than other days due to wide adoption of a hashatg against US vaccines among activists especially in Iraq.

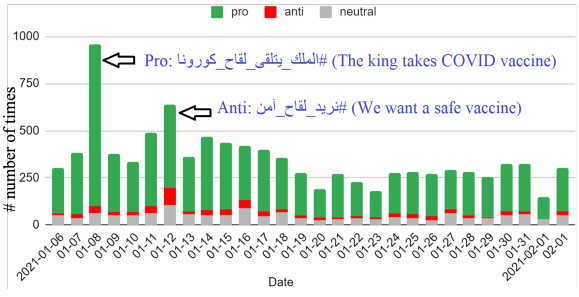


Figure 4: Distribution of *stance* towards vaccine over time. pro: positive stance, anti: negative stance.

Uses of Mobile Applications: We spotted also discussions about the mobile applications listed in Table 5 that help fighting the spread of COVID-19 virus. The purposes of these applications vary from showing the health status of application users, reporting violations of precautionary measures, booking and following-up medical services, tracking medicines, and facilitating travel/visa process. It is worth to mention that there are other applications used in different Arab countries but did not appear in our dataset, e.g., احتراز Ehterhaz "Precaution" (Qatar, released in Apr'20), أمان Aman "Safety" (Jordan, Aug'20), etc.

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: Number of downloads at Google Store in May'20.

6. Experiments

For the experiments, we randomly split the data into train, dev and test sets with 7,000, 1,000 and 2,000 tweets, respectively. Table 6 shows the distribution of categories across the three label sets, defined as three tasks, which include (i) *Task 1:* distinguish important tweets from less important ones, (ii) *Task 2:* fine-grained classification of important tweets, and (iii) *Task 3* stance of the tweets.

We trained several models, SVM with different features combinations, and different transformers models as discussed below.

To measure the performance of the models we report macro-averaged Precision (P), Recall (R) and F1 score

Class	Train	Dev	Test
Informativeness			
More informative	5,482	819	1,590
Less informative	1,518	181	410
Fine-grained categorization			
Info-news	3,623	545	1,057
Celebrity	977	145	276
Plan	606	82	172
Requests	112	20	40
Rumors	79	15	24
Advice	67	10	17
Restrictions	18	2	4
Personal	1,027	128	275
Unrelated	324	36	90
Others	167	17	45
Stance			
Negative	439	70	127
Neutral	1,017	126	253
Positive	5,544	804	1,620

Table 6: Distribution of labels for different tasks.

along with Accuracy (Acc) on the test set. We used F1 score for comparison.

6.1. Classification Models

Support Vector Machines (SVMs) SVMs are known to perform decently for Arabic text classification tasks, with imbalanced class distribution, in tasks such as offensiveness detection (Hassan et al., 2020; Chowdhury et al., 2020b), text categorisation (Chowdhury et al., 2020a) or dialect identification (Abdelali et al., 2020). Due to its popularity and efficacy among machine learning algorithms, this is one of the algorithm we explored in this study for the aforementioned classification tasks.

Using SVM, we experimented with character and word n-gram features weighted by term frequency-inverse term document frequency (tf-idf). We report results for only the most significant ranges, namely, word [1-3] and character [2-7]. As for the classifier training, we used LinearSVC implementation by scikit-learn[10] with its default parameters.

Deep Contextualized Transformer Models (BERT) Transformer-based pre-trained contextual embeddings, such as BERT (Devlin et al., 2019), have outperformed other classifiers in many NLP tasks. We used AraBERT (Antoun et al., 2020), a BERT-based model trained on Arabic news and QARiB (Abdelali et al., 2021), another BERT-model trained on Arabic Wikipedia and Twitter data. We used ktrain library (Maiya, 2020) that utilizes Huggingface[11] implementation to fine-tune AraBERT and QARiB. We used learning rate of 8e-5, truncating length of 54 and fine-tuned for 3 epochs.

6.2. Results

Task 1: Informativeness For discriminating between more *vs.* less informative tweets, we designed binary classifiers and report the results in Table 7. For baseline, we used majority approach where we assign the label of most frequent class. We observed the fine-tuned BERT models, AraBERT and QARiB outperform the SVMs significantly. We noticed AraBERT achieves the highest macro F1 score of 80%.

Task 2: Fine-grained Content Types Classification We trained the models with fine-grained labels using the multiclass classification setting. Due to skewed class distribution, we merged scarce classes (see Table 1) and use the hierarchical representation for further classification. For this, we merged *Restrict* and *Request* classes with *Plan*, *Advice* with *Info-news*, due to their similarity in nature. We exclude *Rumor* class since detecting rumors is difficult without any fact-checked information or other contextual features.

The merging process results in four classes: (i) Info-news, (ii) Celebrity, (iii) Plan, and (iv) Less Informative. From the results reported in Table 7, we noticed all classifiers outperform the majority baseline. Moreover, we noticed that once again, the fine-tuned BERT models, AraBERT and QARiB outperform the simple SVMs. With an F1 score of 67.1, QARiB outperforms AraBERT (F1 score of 64.3) by 2.8%. From the confusion matrix (see Figure 5), we observed a confusion for the class *Plan* with *Info-News*. Such a confusion is indeed expected due to the similarity in nature of the tweets. For example, plans introduced by the government are very much similar to the tweets that are discussing the vaccine news or condition to take it.

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.

Task 3: Stance Detection For predicting the stance of the user (tweet), we designed a multiclass classifier using the aforementioned algorithms (see Section 6.1). To identify the stance of the tweets, we designed the classifier using 3 classes: *positive*, *negative* and *neural*. From our results, in Table 7, we observed a similar pattern to Task 2, where transformers outperform SVMs by a significant margin of about 10%. QARiB achieves the best results with an F1 score of 63.1%. A relatively low performance suggests that stance detection is a difficult task for classifiers. From the per class performance (see Figure 6), we noticed that both neutral and negative stances are confused with the positive ones (the major class).

7. Error Analysis and Findings

7.1. Error Analysis

To understand the designed model behaviour, we analyzed the errors and confusion made by the best classifier, fine-tuned QARiB for Task 2 (fine-grained content type classification) and Task 3 (stance detection).

[10] https://scikit-learn.org/

[11] https://huggingface.co/

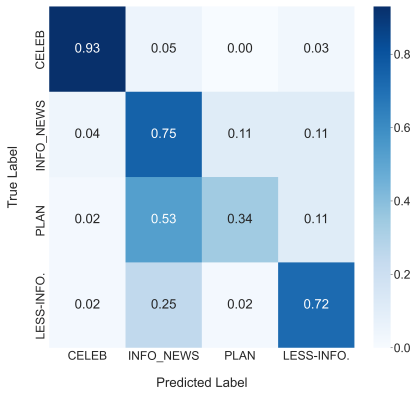


Figure 5: Confusion matrix of fine-grained classification normalized over true labels.

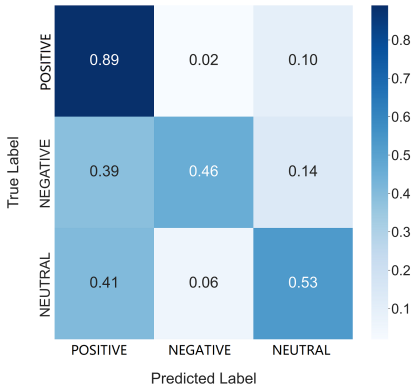


Figure 6: Confusion matrix of stance detection normalized over true labels.

Figure 5 shows the confusion matrix for the fine-grained classification by QARiB. From the confusion matrix, we can see that most errors stem from *Plan* class misclassified as *Info-news*. Figure 6 shows the confusion matrix for stance detection. The confusion matrix shows that most errors stem from Anti-vaccine and Neutral being tagged as Pro-vaccine due to high-class imbalance.

7.2. Classification Errors: Task 2

We picked 200 errors from our best classifier and analyzed them manually. We can summarize most important cases in the following categories:

- **Confusion between classes:** *Info-news* (vaccine) and *Plan* (vaccination process) in the reference or system prediction.
- **Annotation errors:** In some cases personal opinions about the vaccine are labeled as informative.
- **Multilabel:** Some tweets can have more than one class label. For example, announcement from government about the vaccine followed by details about vaccination plan. We plan to annotate multiple labels in the future. We found this case in 10% of the errors.
- **Contextual information:** Need to consider associated multimedia posted in tweet to get more accurate prediction. For example, a question about the vaccine and the answer is in an associated video.

7.3. Stance Errors: Task 3

Similarly, for a randomly selected 200 errors in stance prediction, we found the following issues:

- **Full context:** Need to understand full context including questions and associated multimedia. This includes also considering sarcasm and negation. For example, Is vaccine unsafe? Answer: No.
- **Annotation errors:** Labelling a question about taking the vaccine or not as positive stance.
- **Ambiguous content:** Errors are due to spam, unrelated or unclear content.
- **Mixed/Targeted stance:** For example, refusing vaccines from a certain country but want a safer vaccine.

7.4. Key Observations

In this study, we demonstrated the popularity of different vaccines, the common hashtags, e.g., 'safe vaccine', present in the data, indicating the main concern of the public towards the vaccine. We also observed different types of rumors spreading the doubts on the safety of vaccination, conspiracy theory and doubts in government assessments and plans. Meanwhile, we also noticed informative tweets confirming vaccine safety, promising fair access and priority and importance of front-liners vaccination. We observed the topics covered in the tweets are mainly health, politics and society centred. From stance timeline, we observed the reliability on the vaccine (pro-stance) increase when leaders/influencers (e.g., kings) takes the vaccine, to set examples.

As for the classification performance, for all the three tasks we noticed transformer models outperforms SVMs and present a high performance classifier even with imbalanced class levels. Such a performance indicates the efficacy of this data to aid automation of such process.

8. Conclusion

We presented and publicly released the first large manually annotated Arabic tweet dataset, *ArCovidVac*, for the COVID-19 vaccination campaign. The dataset consists of 10k tweets, covering many countries in Arab region, is enriched with different types of annotation, including, (i) informativeness of the tweets; (ii) fine-grained tweet content types with 10 classes; and (iii) stance towards vaccination identifying tweets with pro-vaccination (positive), neutral, anti-vaccination content (negative). We performed an in-depth analysis of the dataset considering diverse aspects and presented classification results, which can be used as a benchmark in future studies and aid policymakers in decision making process. In the future, we plan to study (i) the dynamics and changes in types/topics of the content and stance towards vaccination in long run, (ii) class imbalance issue, (iii) training the models by dividing the data in chronological order, and (iv) multilabel annotation.

References

Abdelali, A., Mubarak, H., Samih, Y., Hassan, S., and Darwish, K. (2020). Arabic dialect identification in the wild. arXiv, abs/2005.06557.

Abdelali, A., Hassan, S., Mubarak, H., Darwish, K., and Samih, Y. (2021). Pre-training bert on arabic tweets: Practical considerations. arXiv: abs/2102.10684.

Abdul-Mageed, M., Elmadany, A., Pabbi, D., Verma, K., and Lin, R. (2020). Mega-COV: A billion-scale dataset of 65 languages for COVID-19. arXiv:2005.06012.

Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., and Nakov, P. (2021a). Fighting the COVID-19 infodemic in social media: A holistic perspective and a call to arms. In Proceedings of the International AAAI Conference on Web and Social Media, ICWSM '21, pages 913–922, Online. AAAI.

Alam, F., Shaar, S., Dalvi, F., Sajjad, H., Nikolov, A., Mubarak, H., Da San Martino, G., Abdelali, A., Durrani, N., Darwish, K., Al-Homaid, A., Zaghouani, W., Caselli, T., Danoe, G., Stolk, F., Bruntink, B., and Nakov, P. (2021b). Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In Findings of the Association for Computational Linguistics, EMNLP (Findings) '21, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alqurashi, S., Alhindi, A., and Alanazi, E. (2020). Large arabic twitter dataset on COVID-19. arXiv preprint arXiv:2004.04315.

Alshaabi, T., Dewhurst, D. R., Minot, J. R., Arnold, M. V., Adams, J. L., Danforth, C. M., and Dodds, P. S. (2021). The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. EPJ data science, 10(1):15.

Alsudias, L. and Rayson, P. (2020). 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. Association for Computational Linguistics.

Antoun, W., Baly, F., and Hajj, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pages 9–15, Marseille, France. European Language Resource Association.

Baly, R., Mohtarami, M., Glass, J., Màrquez, L., Moschitti, A., and Nakov, P. (2018). Integrating stance detection and fact checking in a unified corpus. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 21–27, New Orleans, Louisiana. Association for Computational Linguistics.

Banda, J. M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E., and Chowell, G. (2021). A large-scale COVID-19 twitter chatter dataset for open scientific research—an international collaboration. Epidemiologia, 2(3):315–324.

Chen, E., Lerman, K., and Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. JMIR Public Health Surveill, 6(2):e19273.

Chowdhury, S. A., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., and Klasinas, I. (2014). Cross-language transfer of semantic annotation via targeted crowdsourcing. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore. ISCA.

Chowdhury, S. A., Calvo, M., Ghosh, A., Stepanov, E. A., Bayer, A. O., Riccardi, G., García, F., and Sanchis, E. (2015). Selection and aggregation techniques for crowdsourced semantic annotation task. In Proceedings of the 16th Annual Conference of the International Speech Communication Association, Dresden, Germany. ISCA.

Chowdhury, S. A., Abdelali, A., Darwish, K., Soon-Gyo, J., Salminen, J., and Jansen, B. J. (2020a). Improving arabic text categorization using transformer training diversification. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 226–236, Online. Association for Computational Linguistics.

Chowdhury, S. A., Mubarak, H., Abdelali, A., Jung, S.-g., Jansen, B. J., and Salminen, J. (2020b). A multi-platform Arabic news comment dataset for offensive language detection. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 6203–6212, Marseille, France. European Language Resources Association.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. (2020). The COVID-19 social media infodemic. Scientific reports, 10(1):1–10.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Elhadad, M. K., Li, K. F., and Gebali, F. (2020). Covid-19-fakes: a twitter (arabic/english) dataset for detecting misleading information on covid-19. In International Conference on Intelligent

Networking and Collaborative Systems, pages 256–268. Springer.

Haouari, F., Hasanain, M., Suwaileh, R., and Elsayed, T. (2021a). ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 82–91, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Haouari, F., Hasanain, M., Suwaileh, R., and Elsayed, T. (2021b). ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 72–81, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hassan, S., Samih, Y., Mubarak, H., and Abdelali, A. (2020). ALT at SemEval-2020 task 12: Arabic and English offensive language identification in social media. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, pages 1891–1897, Barcelona (online). International Committee for Computational Linguistics.

Hassan, S., Mubarak, H., Abdelali, A., and Darwish, K. (2021). ASAD: Arabic social media analytics and unDerstanding. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 113–118, Online. Association for Computational Linguistics.

Hossain, T., Logan IV, R. L., Ugarte, A., Matsubara, Y., Young, S., and Singh, S. (2020). COVIDLies: Detecting COVID-19 misinformation on social media. In Proceedings of the 1st Workshop on NLP for COVID-19, Online. Association for Computational Linguistics.

Huang, B. and Carley, K. M. (2020). Disinformation and misinformation on twitter during the novel coronavirus outbreak. arXiv preprint arXiv:2006.04278.

Khouja, J. (2020). Stance prediction and claim verification: An Arabic perspective. In Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), pages 8–17, Online. Association for Computational Linguistics.

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T., Duan, W., Tsoi, K. K., and Wang, F. (2020). Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo. IEEE Transactions on Computational Social Systems, 7(2):556–562.

Maiya, A. S. (2020). ktrain: A low-code library for augmented machine learning. arXiv preprint arXiv:2004.10703.

Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., and Lehmann, C. U. (2020). An "infodemic": Leveraging high-volume twitter data to understand public sentiment for the COVID-19 outbreak. medRxiv 2020.04.03.20052936.

Mitchell, A. and Page, D. (2014). State of the news media 2014: Overview. Pew Research Center.

Mubarak, H. and Darwish, K. (2016). Demographic surveys of arab annotators on crowdflower. In Proceedings of ACM WebSci16 Workshop "Weaving Relations of Trust in Crowd Work: Transparency and Reputation across Platforms, Hannover, Germany.

Mubarak, H. and Hassan, S. (2021a). ArCorona: Analyzing Arabic tweets in the early days of coronavirus (COVID-19) pandemic. In Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 1–6, online. Association for Computational Linguistics.

Mubarak, H. and Hassan, S. (2021b). UL2C: Mapping user locations to countries on Arabic Twitter. In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pages 145–153, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Nagoudi, E. M. B., Elmadany, A., Abdul-Mageed, M., and Alhindi, T. (2020). Machine generation and detection of Arabic manipulated and fake news. In Proceedings of the Fifth Arabic Natural Language Processing Workshop, pages 69–84, Barcelona, Spain (Online). Association for Computational Linguistics.

Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeno, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N., et al. (2021a). The clef-2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. In Proceedings of the European Conference on Information Retrieval, ECIR '21, pages 639–649, Online. Springer.

Nakov, P., Giovanni, D. S. M., Elsayed, T., Barrón-Cedeño, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Mansour, W., Hamdan, B., Ali, Z. S., Babulkov, N., Nikolov, A., Shahi, G. K., Struß, J. M., Mandl, T., Kutlu, M., and Kartal, Y. S. (2021b). Overview of the CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. LNCS (12880). Springer.

Nakov, P., Barrón-Cedeño, A., Da San Martino, G., Alam, F., Struß, J. M., Mandl, T., Míguez, R., Caselli, T., Kutlu, M., Zaghouani, W., Li, C., Shaar, S., Shahi, G. K., Mubarak, H., Nikolov, A., Babulkov, N., Kartal, Y. S., and Beltrán, J. (2022). The CLEF-2022 CheckThat! Lab on fighting the covid-19 infodemic and fake news detection. In Matthias Hagen, et al., editors, Advances in Information Retrieval, pages 416–428, Cham. Springer International Publishing.

Pulido, C. M., Villarejo-Carballido, B., Redondo-Sama, G., and Gómez, A. (2020). Covid-19 infodemic: More retweets for science-based information on coronavirus than for false information.

International sociology, 35(4):377–392.

Qazi, U., Imran, M., and Ofli, F. (2020a). Geocov19. SIGSPATIAL Special, 12(1):6–15.

Qazi, U., Imran, M., and Ofli, F. (2020b). GeoCoV19: A dataset of hundreds of millions of multilingual COVID-19 tweets with location information. SIGSPATIAL Special, 12(1):6–15.

Shaar, S., Alam, F., Da San Martino, G., Nikolov, A., Zaghouani, W., Nakov, P., and Feldman, A. (2021). Findings of the NLP4IF-2021 shared tasks on fighting the COVID-19 infodemic and censorship detection. In Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, pages 82–92, Online. Association for Computational Linguistics.

Shahi, G. K. and Nandini, D. (2020). FakeCovid – a multilingual cross-domain fact check news dataset for covid-19. In Workshop Proceedings of the 14th International AAAI Conference on Web and Social Media, Online. AAAI.

Shahi, G. K., Dirkson, A., and Majchrzak, T. A. (2020). An exploratory study of covid-19 misinformation on twitter. arXiv preprint arXiv:2005.05710.

Song, X., Petrak, J., Jiang, Y., Singh, I., Maynard, D., and Bontcheva, K. (2020). Classification aware neural topic model and its application on a new COVID-19 disinformation corpus. arXiv:2006.03354.

Vidgen, B., Hale, S., Guest, E., Margetts, H., Broniatowski, D., Waseem, Z., Botelho, A., Hall, M., and Tromble, R. (2020). Detecting East Asian prejudice on social media. In Proceedings of the Fourth Workshop on Online Abuse and Harms, pages 162–172, Online. Association for Computational Linguistics.

Yang, Q., Alamro, H., Albaradei, S., Salhi, A., Lv, X., Ma, C., Alshehri, M., Jaber, I., Tifratene, F., Wang, W., et al. (2020). SenWave: Monitoring the global sentiments under the covid-19 pandemic. arXiv preprint arXiv:2006.10842.

Zhou, X., Mulay, A., Ferrara, E., and Zafarani, R. (2020). ReCOVery: A multimodal repository for COVID-19 news credibility research. In Mathieu d'Aquin, et al., editors, Proceedings of the 29th ACM International Conference on Information and Knowledge Management, pages 3205–3212, Online. ACM.

Appendix

Ethics and Broader Impact

Dataset Collection

We collected the dataset using the Twitter API[12] with keywords that only use terms related to *COVID-19 vaccine*, without other biases. We followed the terms of use outlined by Twitter.[13] Specifically, we only downloaded public tweets, and we only distribute tweet text and label information. We release the dataset by maintaining Twitter data redistribution policy.

Biases

We note that some of the annotations are subjective. Thus, it is inevitable that there would be biases in our dataset. Yet, we have very clear instructions, which should reduce biases. We anticipate annotation errors are also to due to the fact that tweets come from different Arab countries (ex: Gulf region) and some may contain heavy dialects or need cultural background to be fully understood and annotated correctly. As noted in (Mubarak and Darwish, 2016), almost one third of Arab annotators on Appen (previously CrowdFlower) are from Egypt, the most populous Arab country, and around 80% of them are males. These factors can be other sources of unintended biases in the annotation process.

Misuse Potential

Most datasets compiled from social media present some risk of misuse. We, therefore, ask researchers to be aware that our dataset can be maliciously used to unfairly moderate text (e.g., a tweet) that may not be malicious based on biases that may or may not be related to demographics and other information within the text. Intervention with human moderation would be required in order to ensure this does not occur.

Intended Use

Our dataset can enable automatic systems for analysis of social media content, which could be of interest to practitioners, social media platforms, and policymakers. Such systems can be used to alleviate the burden for social media moderators, but human supervision would be required for more intricate cases and in order to ensure that the system does not cause harm.

[12] http://developer.twitter.com/en/docs

[13] http://developer.twitter.com/en/developer-terms/agreement-and-policy