ArCorona: Analyzing Arabic Tweets in the Early Days of Coronavirus (COVID-19) Pandemic

Hamdy Mubarak, Sabit Hassan
Qatar Computing Research Institute
HBKU, Qatar
{hmubarak, sahassan2}@hbku.edu.qa

Abstract

Over the past few months, there were huge numbers of circulating tweets and discussions about Coronavirus (COVID-19) in the Arab region. It is important for policy makers and many people to identify types of shared tweets to better understand public behavior, topics of interest, requests from governments, sources of tweets, etc. It is also crucial to prevent spreading of rumors and misinformation about the virus or bad cures. To this end, we present the largest manually annotated dataset of Arabic tweets related to COVID-19. We describe annotation guidelines, analyze our dataset and build effective machine learning and transformer based models for classification.

1 Introduction

As the Coronavirus (COVID-19) crippled lives across the world, people turned to social media to share their thoughts, news about vaccines or cures, personal stories, etc. With Twitter being one of the popular social media platforms in the Arab region, tweets became a major medium of discussion about COVID-19. These tweets can be indicators of psychological and physical well being, public reactions to specific actions taken by the government and also public expectation from governments. Therefore, identifying types of tweets and understanding their content can aid decision making by governments. It is also important for governments to identify and prevent rumours and bad cures since they can bring harm to society.

While there have been many recent works about tweets related to COVID-19, there are a very few targeted toward aiding governments in their decision making in the Arab region despite Arabic being one of the dominant languages on Twitter (Alshaabi et al., 2020). Some of the existing works use automatically collected datasets (Alqurashi et al., 2020). Manually labeled datasets are either small in size (few hundred tweets) (Alam et al., 2020) or target a different task such as sentiment analysis (Haouari et al., 2020). To fill this gap, we present and publicly share the largest (to our best knowledge) manually annotated dataset of Arabic tweets collected from early days of COVID-19, labeled for 13 classes. We present our data collection and annotation scheme followed by data analysis, identifying trends, topics and distribution across countries. Lastly, we employ machine learning and transformer models for classification.

2 Related Work

Much of recent works on COVID-19 rely on queries to Twitter or distant supervision. This allows a large number of tweets to be collected. Chen et al. (2020) collect 123M tweets by following certain queries and accounts on Twitter. GeoCoV19 (Qazi et al., 2020) is a large-scale dataset containing 524M tweets with their location information. Banda et al. (2020) collected 152M tweets at the time of their writing. Li et al. (2020) identifies situational information about COVID-19 and its propagation on Weibo. Other works include propagation of misinformation (Huang and Carley, 2020; Shahi et al., 2020), cultural, social and political impact of misinformation (Leng et al., 2020) and rumor amplification (Cinelli et al., 2020).

For Arabic, we see a similar trend where few datasets are manually labeled. Alqurashi et al. (2020) provide a large dataset of Arabic tweets containing keywords related to COVID-19. Similarly, ArCOV-19 (Haouari et al., 2020) is a dataset of 750K tweets obtained by querying Twitter. Alam et al. (2020) annotate a small number of English (currently 504) and Arabic tweets (currently 218) for (i) existence of claim and worthiness of fact-checking (ii) harmfulness to society, and (iii) relevance to governments or policy makers. Yang et al.

Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 1–6
April 19, 2021.

(2020) annotate 10K Arabic and English tweets for the task of fine-grained sentiment analysis.

Alsudias and Rayson (2020) collected 1M unique Arabic tweets related to COVID-19 between December 2019 and April 2020. They used K-means algorithm from Scikit-learn Python package to cluster tweets into 5 clusters, namely: statistics, prayers, disease locations, advising, and advertising. They also annotated random 2000 tweets for rumor detection based on the tweets posted by the Ministry of Health in Saudi Arabia.

3 Data Collection

We used twarc search API[1] to collect tweets having the Arabic word كورونا (Corona) in Feb and March 2020. We collected 30M tweets in total. The reason behind selecting this word is that it's widely used by normal people, news media[2] and official organizations [3] as opposed to 19-كوفيد (COVID-19) which is rarely used by normal people in different Arab countries based on our observations. We aimed to increase diversity of tweet sources. Our collection covers the period from Feb 21 until March 31 in which Coronavirus was reported for the first time in All Arab countries except United Arab Emirates (AE)[4] (Jan 29) and Egypt (EG) (Feb 14). The date of the first reported Coronavirus case in Lebanon (LB) was Feb 21, and in Iraq (IQ), Bahrain (BH), Oman (OM), and Kuwait (KW) was Feb 24, in Qatar was Feb 29, and in Saudi Arabia (SA) was March 2. All other Arab countries came later.

4 Data Annotation

During the period of our study (40 days), we extracted the top retweeted 200 tweets in each day (total of 8000). We assume that the top retweeted tweets are the most important ones which get highest attention from Twitter users. Annotation was done manually by a native speaker who is familiar with Arabic dialects according to class descriptions shown in Table 1. To measure quality, we annotated 200 random tweets by a second annotator. Inter-annotator agreement was 0.85 using Cohen's kappa coefficient which indicates high quality given that annotation is not trivial and some classes are close to each other. Examples of annotation classes are shown in Figures 1 and 2.

Note: If a tweet has multimedia (image or video) or an external link (URL or another tweet), the annotator was asked to open it and judge accordingly to consider the full context of the tweet. For example, if a tweet has a text about a prayer and the attached image is about number of new cases, this should be classified as REP not PRAYER.
Data can be downloaded from this link[5]: `https://alt.qcri.org/resources/ArCorona.tsv`.

4.1 Limitation

We found that ≈10% of the tweets can take more than one class, e.g. a tweet reports new cases and a medical advice. We plan to allow multiple labels in future. In the current version, such tweets take the label of the first "important" class. We consider the first 8 classes in Table 1 to be important and the last 5 classes (PRSNL, SUPPORT, PRAYER, UNIMP and NOT_ARB) to be less important. These classes will be merged into LessImportant class.

5 Analysis

Class timeline is shown in Figure 3. We can observe the following important notes:

- Large portion of tweets can be considered as LessImportant to many people (≈ 30%).
- Reports (REP) and actions taken by governments (ACT) are the most retweeted tweets.
- Information about the virus (INFO) get less attention with time and there is an increasing number of tweets about volunteering (VOLUNT).
- There are continuous requests for governments to take actions (SEEK_ACT) – especially in the beginning (≈ 15%), and few tweets are about rumors (≈ 5%) and cures (≈ 2%).

We took a random sample of 1000 tweets and annotated them for their topics. Figure 4 shows that, in addition to health, the virus affected many aspects of people's lives such as politics, economy, education, etc. We found also that 7% of tweets have hate speech, e.g. attacking China and Iran for spreading the virus as shown in Figure 5.

[1] https://github.com/DocNow/twarc
[2] https://www.aljazeera.com/topics/events/coronavirus-outbreak.html
[3] https://www.who.int/ar/emergencies/diseases/novel-coronavirus-2019
[4] We use ISO 3166-1 alpha-2 for country codes
[5] We share tweet id, date and class.

Class	Description	Count
1. REP	Reports and announcements such as number of infections, recovery cases and deaths.	1664
2. ACT	Measures or actions taken by governments such as curfew, closing of country borders, shops and worship places. This includes discussions and consequences of these measures.	1383
3. INFO	Information about the virus, symptoms, incubation period, how it spreads, mask types, etc.	300
4. RUMOR	Rumor or refute rumor. A rumor is a circulating story or report of uncertain or doubtful truth.	421
5. ADVICE	Advice such as washing hands, staying at home, wearing masks and avoiding travel.	1047
6. SEEK_ACT	Seek actions from governments such as closing airports, and controlling prices of goods.	587
7. CURE	News about good and bad cure, diagnosis, ventilators, supportive medical equipment, etc.	116
8. VOLUNT	Volunteering efforts or donation of money, goods or services.	133
9. PRSNL	Personal story or opinion.	453
10. SUPPORT	Support or praise governments, medical staff, celebrities, etc.	386
11. PRAYER	Prayer.	563
12. UNIMP	Unrelated or unimportant such as spams or advertisements.	786
13. NOT_ARB	Not Arabic, e.g. Persian.	161
Total		**8000**

Table 1: Annotation classes and distribution: Important classes (top) and LessImportant classes (bottom)

Figure 1: Examples for REP, ACT, INFO, RUMOR, ADVICE and SEEK_ACT classes

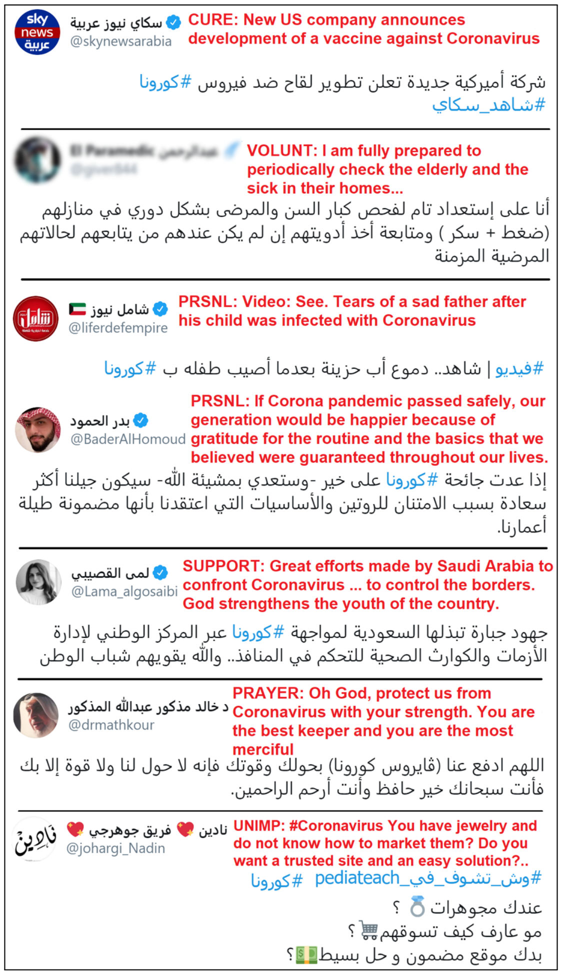

Figure 2: Examples for CURE, VOLUNT, PRSNL, SUPPORT, PRAYER and UNIMP classes

Table 2 shows country distribution and top accounts for the original authors of tweets.

Typically, people retweet tweets from ministry of health in their countries in addition to famous news agencies and celebrities. Most of these accounts are verified.

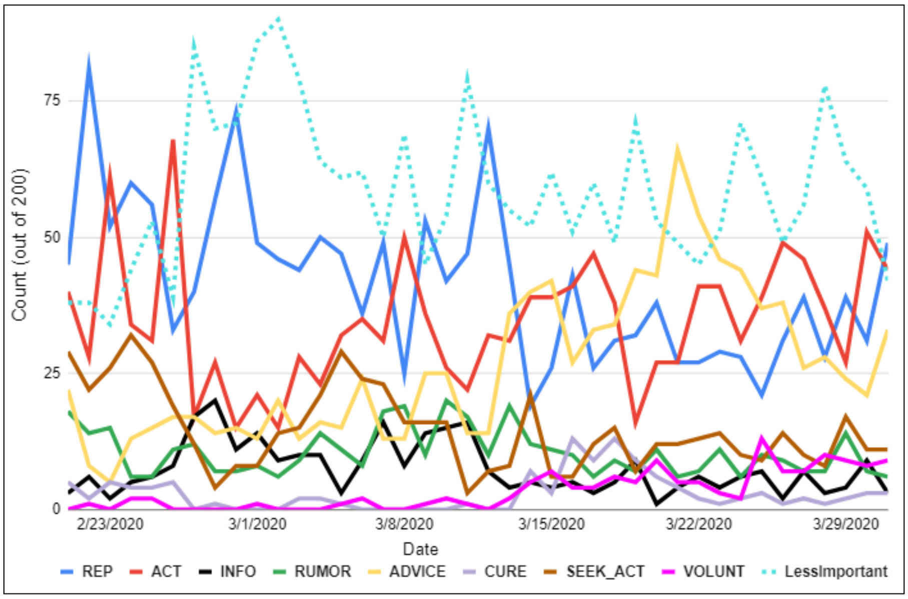

Figure 3: Timeline: Number of class tweets each day

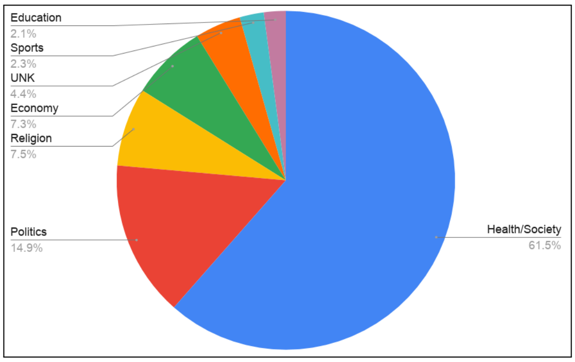

Figure 4: Tweet topics

Figure 5: Hate speech example

Country	%	Top Accounts
SA	59	SaudiNews50, SaudiMOH
OTH	13	MohamadAhwaze, amjadt25
OM	7	OmanVSCovid19, OmaniMOH
KW	7	Almajlliss, KUWAIT_MOH
QA	4	amansouraja, MOPHQatar
AE	4	AlHadath, AlArabiya_Brk
EG	3	RassdNewsN, mohpegypt

Table 2: Country distribution and top accounts

6 Experiments and Evaluation

We randomly split the data into sets of 6000, 1000 and 1000 tweets for train, dev and test sets respectively. We report macro-averaged Precision (P), Recall (R) and F1 score along with Accuracy (Acc) on test set [6]. We use F1 score as primary metric for comparison.

6.1 Features

N-gram features 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 [1-2] and character [2-5].

Mazajak Embeddings Mazajak embeddings are word-level skip-gram embeddings trained on 250M Arabic tweets, yielding 300-dimensional vectors (Abu Farha and Magdy, 2019).

6.2 Classification Models

Support Vector Machines (SVMs) SVMs have been shown to perform decently for Arabic text classification tasks such as spam detection (Mubarak et al., 2020), adult content detection (Mubarak et al., 2021; Hassan et al., 2021), offensiveness detection (Hassan et al., 2020b,a) or dialect identification (Abdelali et al., 2020; Bouamor et al., 2019). We experimented with i) word n-gram, ii) character n-gram and iii) Mazajak Embeddings. We used LinearSVC implementation by scikit-learn [7].

[6] Differences between dev and test sets are $\pm 2 - 3\%$ (F1).

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

Model	Features	Acc.	P	R	F1
Majority Class	-	72.5	36.3	50.0	42.0
SVM	W[1-2]	84.4	82.5	76.4	78.6
SVM	C[2-5]	**85.4**	**84.3**	77.4	**79.8**
SVM	Mazajak	83.9	80.5	77.7	78.9
AraBERT		83.9	80.0	**79.2**	79.6

Table 3: Binary classification results

Model	Features	Acc.	P	R	F1
Majority Class	-	22.7	1.7	7.7	2.8
SVM	W[1-2]	**62.8**	**64.3**	53.5	56.3
SVM	C[2-5]	59.0	**64.3**	49.4	51.8
SVM	Mazajak	60.0	55.1	51.5	52.4
AraBERT		62.7	61.6	**59.8**	**60.5**

Table 4: Fine-grained classification results

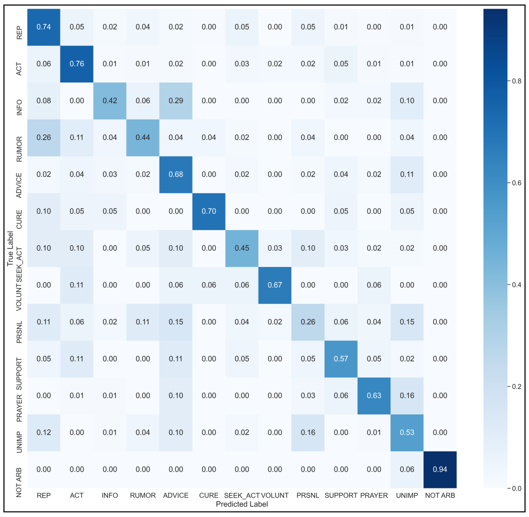

Figure 6: Confusion matrix normalized over true labels

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. We used ktrain library (Maiya, 2020) that utilizes Huggingface[8] implementation to fine-tune AraBERT. We used learning rate of 8e^{-5}, truncating length of 50 and fine-tuned for 5 epochs.

6.3 Binary Classification

First, we experiment to distinguish LessImportant tweets from others (see Section 4). From Table 3, we can see that SVMs with character [2-5]-gram outperformed others with F1 score of **79.8**, closely followed by AraBERT with **79.6** F1.

6.4 Fine-grained Classification

Our next set of experiments were designed for fine-grained classification for 13 classes. With F1 score of **60.5**, AraBERT outperformed others (Table 4).

Error Analysis: AraBERT confusion matrix (Figure 6) shows that PRSNL, INFO and RUMOR are the hardest classes to identify and the most common error is misclassifying INFO as ADVICE. We hypothesize these errors can be reduced if larger data set is being annotated.

7 Conclusion and Future Work

We present the largest publicly available manually annotated dataset of Arabic tweets for 13 classes that includes the most retweeted tweets in the early days of COVID-19. Followed by data analysis, we present models that can reliably identify important tweets and can perform fine-grained classification. In the future, we plan to compare our data to data from later days of the pandemic.

[8]https://huggingface.co/

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2020. Arabic dialect identification in the wild. *ArXiv*, abs/2005.06557.

Ibrahim Abu Farha and Walid Magdy. 2019. Mazajak: An online Arabic sentiment analyser. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 192–198, Florence, Italy. Association for Computational Linguistics.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, and Preslav Nakov. 2020. Fighting the covid-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. *arXiv:2005.00033*.

Sarah Alqurashi, Ahmad Alhindi, and Eisa Alanazi. 2020. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315*.

Thayer Alshaabi, David R. Dewhurst, Joshua R. Minot, Michael V. Arnold, Jane L. Adams, Christopher M. Danforth, and Peter Sheridan Dodds. 2020. The growing amplification of social media: Measuring temporal and social contagion dynamics for over 150 languages on twitter for 2009–2020. Available online at http://arxiv.org/abs/2003.03667.

Lama Alsudias and Paul Rayson. 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.

Wissam Antoun, Fady Baly, and Hazem M. Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *ArXiv*, abs/2003.00104.

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

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Emily Chen, Kristina Lerman, and Emilio Ferrara. 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.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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. Association for Computational Linguistics.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. ArCOV-19: The first Arabic COVID-19 twitter dataset with propagation networks. *arXiv:2004.05861*.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. Asad: Arabic social media analytics and understanding. In *Proceedings of the Software Demonstrations of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Sabit Hassan, Younes Samih, Hamdy Mubarak, and Ahmed Abdelali. 2020a. 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.

Sabit Hassan, Younes Samih, Hamdy Mubarak, Ahmed Abdelali, Ammar Rashed, and Shammur Absar Chowdhury. 2020b. ALT submission for OSACT shared task on offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 61–65, Marseille, France. European Language Resource Association.

Binxuan Huang and Kathleen M Carley. 2020. Disinformation and misinformation on twitter during the novel coronavirus outbreak. *arXiv preprint arXiv:2006.04278*.

Yan Leng, Yujia Zhai, Shaojing Sun, Yifei Wu, Jordan Selzer, Sharon Strover, Julia Fensel, Alex Pentland, and Ying Ding. 2020. Analysis of misinformation during the covid-19 outbreak in china: cultural, social and political entanglements. *arXiv preprint arXiv:2005.10414*.

L. Li, Q. Zhang, X. Wang, J. Zhang, T. Wang, T. Gao, W. Duan, K. K. Tsoi, and F. Wang. 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.

Arun S Maiya. 2020. ktrain: A low-code library for augmented machine learning. *arXiv preprint arXiv:2004.10703*.

Hamdy Mubarak, Ahmed Abdelali, Sabit Hassan, and Kareem Darwish. 2020. Spam detection on arabic twitter. In *Social Informatics*, pages 237–251, Cham. Springer International Publishing.

Hamdy Mubarak, Sabit Hassan, and Ahmed Abdelali. 2021. Adult content detection on arabic twitter: Analysis and experiments. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*.

Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19. *SIGSPATIAL Special*, 12(1):6–15.

Gautam Kishore Shahi, Anne Dirkson, and Tim A Majchrzak. 2020. An exploratory study of covid-19 misinformation on twitter. *arXiv preprint arXiv:2005.05710*.

Qiang Yang, Hind Alamro, Somayah Albaradei, Adil Salhi, Xiaoting Lv, Changsheng Ma, Manal Alshehri, Inji Jaber, Faroug Tifratene, Wei Wang, et al. 2020. Senwave: Monitoring the global sentiments under the covid-19 pandemic. *arXiv preprint arXiv:2006.10842*.