Social Network Analysis and Mining (2022) 12:71
https://doi.org/10.1007/s13278-022-00902-y

ORIGINAL ARTICLE

A large-scale analysis of COVID-19 tweets in the Arab region

Aya Mourad[1] · **Shady Elbassuoni**[1]

Received: 13 March 2022 / Revised: 6 June 2022 / Accepted: 8 June 2022 / Published online: 2 July 2022

Abstract
The COVID-19 virus has spread rapidly to the Arab World, affecting the public health and economy. As a result, people started communicating about the pandemic through social media such as Twitter. This paper utilizes text mining to extract useful insights into people's perceptions and reactions to the pandemic. First, we identified 11 general topics under which COVID-19 tweets emerging from the Arab region fall. Next, we generated training data consisting of English, multidialectal Arabic, and French tweets that were manually classified into one or more of the identified 11 topics via crowdsourcing. These training data were then used to train various deep learning models to automatically classify a tweet into one or more of the 11 topics. Our best performing models were then used to perform a large-scale analysis of COVID-19 tweets emerging from the Arab region and spanning a period of over one year. Our analysis indicates that the majority of the tweets analyzed emerged from Saudi Arabia, UAE, and Egypt and that the majority of the tweets were generated by males. We also observed a surge in tweeting about all the topics as the pandemic broke followed by a slow and steady decline over the following months. We finally performed sentiment analysis on the analyzed tweets, which indicated a strong negative sentiment until mid of September 2020, after which we observed a strong positive sentiment that coincided with the surge in tweeting about vaccines.

Keywords COVID-19 · Corona virus · Twitter · Analysis · Arab region

1 Introduction

In March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak as a pandemic. COVID-19 originated in China and has rapidly spread worldwide, affecting humans' daily routines. Its spread has impacted several sectors, mainly the global economy, public and private sectors, governments decisions, and people's mental health. The Arab region, which is home to a total population of around 436 million (WorldBank 2020), has been heavily hit by the pandemic. As the number of infections and deaths caused by COVID-19 intensified with no treatment or vaccine as of 2020, Arab governments implemented various measures to combat the pandemic, including enforcing curfews, closing public businesses, banning social gatherings, shutting down airports, and enforcing public health measures such as social distancing and masks. As a result, people started to communicate their thoughts, concerns, beliefs, and information related to COVID-19 through several social media platforms, including Twitter.

Since the beginning of the crisis, users have tweeted about COVID-19 symptoms, patients' stories, causes of infection, WHO announcements, COVID-19 transmission, among other topics. Some users tweeted about COVID-19 statistics, covering the daily cases and deaths, and public health measures such as wearing masks, gloves, and washing hands to spread awareness among citizens. Moreover, people used Twitter to disseminate governmental measures and political parties' actions related to the pandemic, such as hours of curfews, obligations to wear a mask, and social distancing. As these measures were implemented to prevent the spread of the virus, they resulted in new social norms. For instance, many people turned to working from home and in turn Twitter users started sharing their experiences on their new daily routines. On the other hand, several conspiracy theories and fake treatment news about COVID-19 have started and continue to spread on Twitter. Finally, as

Aya Mourad and Shady Elbassuoni have contributed equally to this work.

✉ Aya Mourad
aam63@mail.aub.edu

Shady Elbassuoni
se58@aub.edu.lb

[1] Computer Science Department, American University of Beirut, Bliss Street, Beirut, Lebanon

the COVID-19 vaccine production began, people started expressing their opinions about the vaccine. Some recommended receiving it, and others are still hesitant about it, leading to the rise of the anti-vaxxers community.

Our goal in this paper is to conduct a large-scale analysis of the COVID-19 discourse on Twitter, specifically taking place in the Arab region. To this end, we utilize a large dataset consisting of tweets related to COVID-19 that are geotagged and that span the period from February 1, 2020, until April 30, 2020 (Qazi et al. 2020). We used this initial dataset to identify the different topics under which the discourse surrounding COVID-19 in the Arab region falls. To identify the different topics, we relied on sampling tweets from the dataset followed by manual inspection of the sampled tweets and insight from the literature about COVID-19 discourse. Using such strategy, we were then able to identify 11 different topics, under which most tweets related to COVID-19 fall. These topics include *Economics, Stocking Up, Vaccine, COVID-19 Statistics, COVID-19 Information, Politics, Public Health Measures, Governmental Measures, Fake Treatment, and Conspiracy Theory*. The 11th topic pertains to tweets that are personal in nature, and does not fall under any of the previously mentioned topics, and we refer to it as the *Non-Informative* category.

Once these topics were identified, three labeled datasets were generated by sampling tweets from the dataset using the central limit theorem. The three datasets consisted of tweets generated by users in the Arab region in either one of the three most commonly used languages in the region, namely Arabic, English and French. To obtain labels for the tweets, we relied on crowdsourcing using the Labelbox platform[1] to associate each tweet in each dataset with one or more of the identified topics mentioned above. The final labeled datasets consisted of 5,600 tweets in English, 4,725 tweets in Arabic, and 5,496 tweets in French.

The three labeled datasets described above were then used to train multiple deep learning models, including CNN, BiLSTM, and BERT to automatically classify a tweet related to COVID-19 into one of the 11 topics we have identified. The best classifier BERT was then applied on a second geotagged dataset of tweets that also contains tweets related to COVID-19 spanning the period from February 1, 2020, to March 31, 2021 (Imran et al. 2021).

Finally, we performed a large-scale analysis of this automatically labeled dataset to understand what Twitter users in the Arab region tweet about when it comes to the COVID-19 pandemic. Our analysis indicates that the majority of the tweets analyzed emerged from Saudi Arabia, UAE, and Egypt, and that the majority of the tweets were generated by males. The analysis also shows that males mainly tweeted about conspiracy theory and governmental measures, whereas females mainly tweeted about politics, stocking up, and fake treatment. We also observed a surge in tweeting about all the topics as the pandemic broke followed by a slow and steady decline over the following months, except for a sudden surge in tweets about vaccines after October 2020, which continued to increase onwards. We also performed sentiment analysis on the analyzed tweets, which indicated a strong negative sentiment until mid of September 2020, after which we observed a strong positive sentiment that coincided with the surge in tweeting about vaccines. Overall, the analysis showed that positive sentiment increased over time with some countries having predominantly positive sentiment compared to negative or neutral ones such as Saudi Arabia, Kuwait, Bahrain and Jordan.

Our contributions in this paper can thus be summarized as follows:

1. we identified 11 general topics that cover the spectrum of tweets about COVID-19 generated in the Arab region through an intensive sampling of tweets from a large geotagged COVID-19 dataset spanning the period from February 1, 2020, to April 30, 2020,
2. we built three labeled datasets consisting of 5,600 English, 4,725 Arabic, and 5,496 French tweets that were manually classified into one or more of the identified 11 topics via crowdsourcing,
3. we used our three datasets to train various deep learning models, including CNN, BiLSTM, and BERT, to automatically classify a tweet into one or more of the 11 topics,
4. we used our best performing models to annotate a large dataset of tweets related to COVID-19 generated by users in the Arab region and spanning the period from February 1, 2020, to March 31, 2021, and
5. we performed a large-scale analysis on the automatically labeled tweets to understand the discourse surrounding COVID-19 in the Arab region, and how it changes over time, across countries and with respect to gender, and we also performed sentiment analysis to understand the overall sentiment of the Arab population towards the pandemic.

The paper is organized as follows. Section 2 gives an overview of annotated COVID-19 datasets and the different attempts to utilize machine learning for analyzing such datasets. Section 3 describes how we generated our annotated datasets and how we used them to train various deep learning models to automatically classify a tweet into one or more identified topics. In Sect. 4, we provide a large-scale analysis over a COVID-19 dataset that spans almost a year and that was automatically annotated by our deep learning models. Finally, we conclude and present future directions in Sect. 5.

[1] https://labelbox.com/.

2 Literature review

Since its outbreak in late 2020, various attempts have been carried out to track and analyze the discourse surrounding the Corona virus in social media such as Twitter. For example, Kumar et al. (2021) proposed classifying tweets into four classes, namely Irrelevant, Conspiracy, True Information, and False Information. To do so, they manually annotated an English dataset consisting of 1,970 tweets and used it to train various deep learning models. Similarly, Memon and Carley (2020) annotated and analyzed an English dataset consisting of a total of 4,573 tweets categorized into 17 different classes including True Treatment, True Prevention, Correction/Calling Out, Sarcasm/Satire, True Public Health Response Conspiracy, Fake Cure, Fake Treatment, False Fact or Prevention, and False Public Health Response.

Alam et al. (2020) also classified a total of 504 English and 218 Arabic tweets into various fine-grained classes. The classes cover whether a tweet contains facts, false information, is of interest to the public, is harmful to a social entity, needs verification by specialists, and whether it needs the government's consideration. Their annotated data were then used to train various classifiers to automatically classify a tweet into one of the aforementioned classes or categories. Additionally, Xue et al. (2020b) analyzed 4 million tweets related to the COVID-19 pandemic from March 1 to April 21 in 2020. They used Latent Dirichlet Allocation (LDA) to identify popular salient topics and themes in the dataset, which included public health measures, social stigma, coronavirus news cases and deaths, COVID-19 in the USA, and COVID-19 cases in the rest of the world. Xue et al. (2020a) also applied the LDA technique for topic modeling and identified ten themes using a dataset consisting of 1.9 million English tweets gathered from January to 7 March 2020. The ten themes include: updates about confirmed cases, COVID-19 related death, cases outside China (worldwide), COVID-19 outbreak in South Korea, early signs of the outbreak in New York, Diamond Princess cruise, economic impact, preventive measures, authorities, and supply chain. They also performed sentiment analysis, which indicated that fear of the unidentified nature of COVID-19 was dominant in all topics.

Hung et al. (2020) also applied machine learning techniques to analyze and study Twitter users' sentiments during the COVID-19 crisis. Their dataset consisted of English tweets geolocated in the USA from 20 March to 19 April 2020. They identified five prevalent themes of COVID-19 discussion via topic modeling using LDA with sentiments ranging from positive to negative. The themes included health care environment, emotional support, business economy, social change, and psychological stress. They performed sentiment analysis for the 902,138 tweets, where 48.2% were classified as positive, 31.1% as negative, and 20.7% as neutral. Imran et al. (2020) utilized different deep learning models to analyze tweets related to COVID-19 focused on six countries, including Pakistan, India, Norway, Sweden, USA, and Canada. They used two tweet datasets to detect sentiment polarity and emotion recognition, including the Sentiment140 and Emotional Tweets datasets. The latter were trained using BERT, GloVe, BiLSTM, and GRU models.

Moreover, Chandrasekaran et al. (2020) compiled a COVID-19-related dataset of 13.9 million English tweets covering the period from January 1 to May 9, 2020, and studied variations of ten topics extracted using LDA and sentiment scores obtained using VADER ? Average negative sentiment was found in posts discussing the symptoms, spread and growth of cases, racism, source of the outbreak, and political impact of COVID-19. In contrast, a reversal of sentiments from negative to positive was shown in tweets debating prevention, impact on the economy and markets, government response, impact on the health care industry, and treatment and recovery.

Yin et al. (2020) collected 13 million tweets related to COVID-19 over two weeks to analyze the topics discussed and sentiment dynamics. The results showed that positive sentiment was higher than the negative sentiment. They also performed a topic-level sentiment analysis, where positive sentiment was shown in the 'stay safe home' topic and negative sentiment in the 'people death' topic. Abd-Alrazaq et al. (2020) identified 12 topics grouped into four themes: the origin of the virus; its sources; its impact on people, countries, and the economy; and ways of mitigating the risk of infection. They also performed sentiment analysis, where positive sentiment was shown in ten topics and negative in two topics: deaths caused by COVID-19 and increased racism.

Some studies like ours presented here in this paper focused on the Arab region. However, the majority of them conducted small-scale analysis over a short period of time and mostly focusing on tweets only in the Arabic language. For instance, Aljabri et al. (2021) performed sentiment analysis on Arabic tweets about distance learning emerging from Saudi Arabia. Similarly, Alqurashi et al. (2021) constructed an Arabic dataset with over 8,000 tweets related to COVID-19 and manually classified the tweets into two categories: misinformation or not. The dataset was then used to train various machine learning models. Unlike Alqurashi et al. that classified the tweets into only two classes, Ameur and Aliane (2021a) classified over 10,000 Arabic COVID-19 tweets into the following 10 classes: hate, cure, advice, morals raising, news or opinions, dialectal, blame and negative speech, factual, fact-checking worthy, and fake information. The annotated dataset was then used to train several

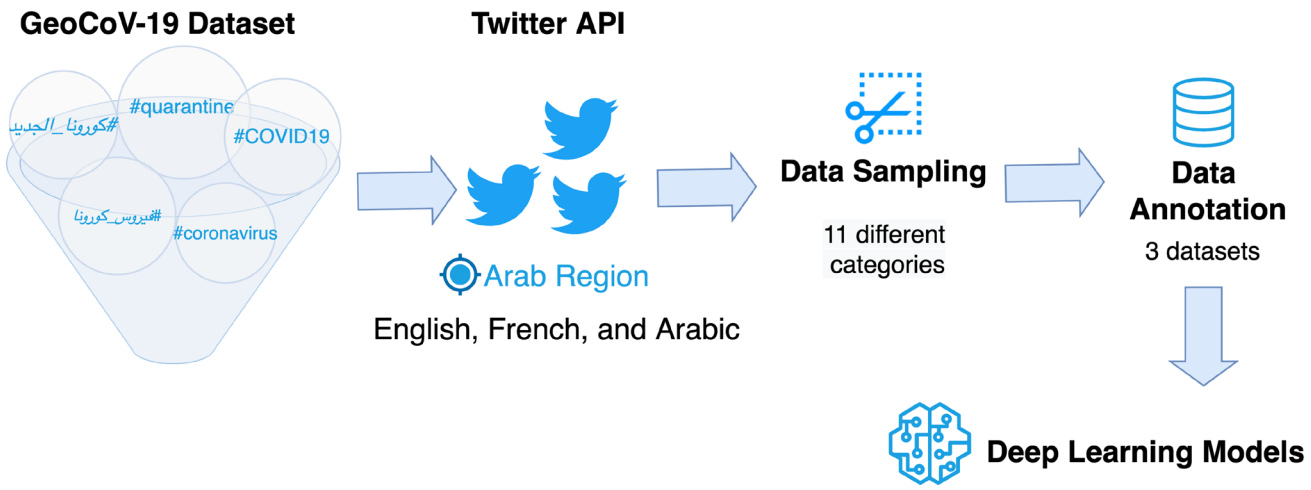

Fig. 1 Overview of approach

machine learning models. The same authors, Ameur and Aliane (2021b), also constructed an Arabic COVID-19 tweets dataset consisting of 5,000 tweets that were manually classified into sarcastic or not and used it to train various machine learning models and perform sentiment analysis.

Finally, Alsudias et al. (2020) identified topics discussed during the pandemic on Twitter using the K-means clustering, which included COVID-19 statistics, prayers, COVID-19 locations, advice, and prevention. They also performed rumors detection by manually sampling 2,000 tweets and classifying them as correct information, false information, or unrelated. Finally, they trained different machine learning models to automatically classify tweets into one or more of the previously mentioned topics.

Most of the above-surveyed studies and analyses were conducted on limited or small-scale data that either lack geolocation information, rely on very general topics for analysis, or both. Only a few conducted large-scale analyses; however, they focused their analysis on a single language. Our work differs from all surveyed work as follows. First, we used a multilingual geolocated dataset with millions of tweets spanning over a year since the pandemic outbreak. Second, we focused our analysis on the Arab region and identified 11 fine-grained topics via rigorous sampling and we generated multiple annotated datasets via crowdsourcing in the three most prominently used languages in the Arab region. Finally, we used our annotated datasets to train various deep learning models to automatically classify tweets into of the identified topics, and used the best performing models to perform a large-scale analysis over the whole dataset to study the discourse on Twitter surrounding COVID-19 in the Arab region.

3 Multilabel classification

In this section, we describe how we identified the various topics under which COVID-19 tweets emerging from the Arab region fall. We also describe how we generated three annotated datasets in English, French and Arabic via crowdsourcing and how we used these datasets to train various deep learning models to automatically classify a COVID-19 tweet into one or more of the identified classes. An overview of our approach to carry out these aforementioned tasks is depicted in Fig. 1. As can be seen in the figure:

1. we used the GeoCoV-19 dataset (Qazi et al. 2020) and Twitter API[2] to retrieve all the tweets related to COVID-19 from the period of February 1, 2020, to April 30, 2020,
2. we then retained only tweets written in English, French, or Arabic, the three most prominent languages used in the Arab region, and that were generated by users located in the Arab region (Fig. 2) using the geolocations of the tweets,
3. we then sampled the tweets extracted as described above and identified 11 categories under which most of these sampled tweets fall,
4. we then annotated three different sampled datasets of tweets in the three relevant languages using crowdsourcing, and
5. the annotated datasets were then used to train various deep learning models to automatically categorize a given COVID-19 tweet into one or more of the 11 identified categories.

[2] https://developer.twitter.com/en/docs/twitter-api.

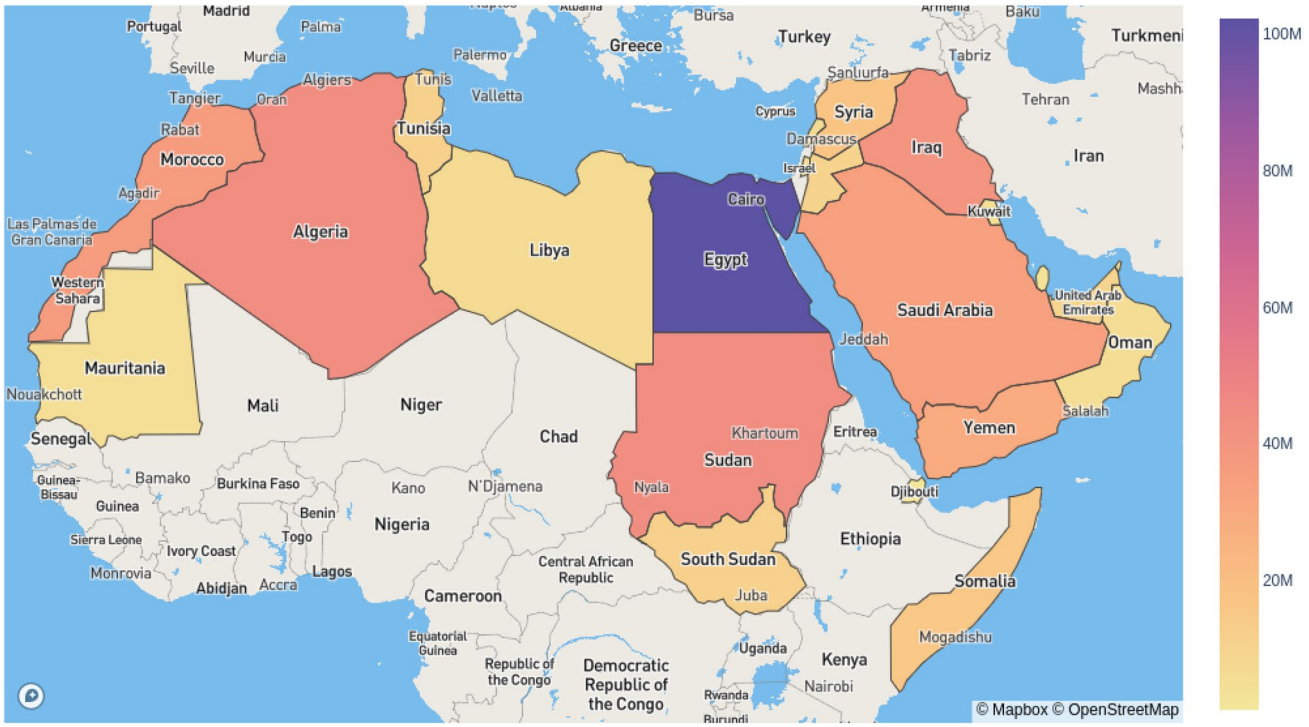

Fig. 2 The Arab region and its population distribution

3.1 Datasets

Our datasets are all retrieved from the GeoCoV-19 dataset (Qazi et al. 2020). It contains more than 524 million multilingual tweets about COVID-19 collected from February 1, 2020, until April 30, 2020, and with geolocations inferred either from the tweet location field, user location field provided in the user profile, or the tweet content. Since our goal is to perform a large-scale analysis of the COVID-19 tweets in the Arab region, we filtered out all the tweets in the GeoCOV-19 dataset whose inferred location is not one of the countries in the Arab region (Fig. 2).

Adhering to Twitter data redistribution policies, GeoCoV-19 does not contain the actual tweets. Instead, the dataset only contains tweet ids and user ids, along with geolocation information for each tweet. Therefore, we used the Twitter API to retrieve the actual tweets we kept (i.e., the ones originating from the Arab region), resulting in 6,710,598 tweets. We then dissected those retrieved tweets by language. We ended up with the distribution shown in Fig. 3 among the three prominently spoken languages in the region (i.e., Arabic, English, and French).

To determine the different categories under which the tweets in our dataset fall, we sampled tweets from each dataset and identified 11 relevant categories using a careful inspection of the sampled tweets as well as thorough literature review (Kumar et al. 2021; Memon and Carley 2020; Alam et al. 2020; Xue et al. 2020b). The identified categories were *Economics, Stocking Up, Vaccine/Cure, COVID-19 Statistics, COVID-19 Information, Politics, Public Health Measures, Governmental Measures, Fake Treatment, Conspiracy Theory, and Non-Informative*. Table 1 shows an example tweet that falls under each one of our 11 identified classes.

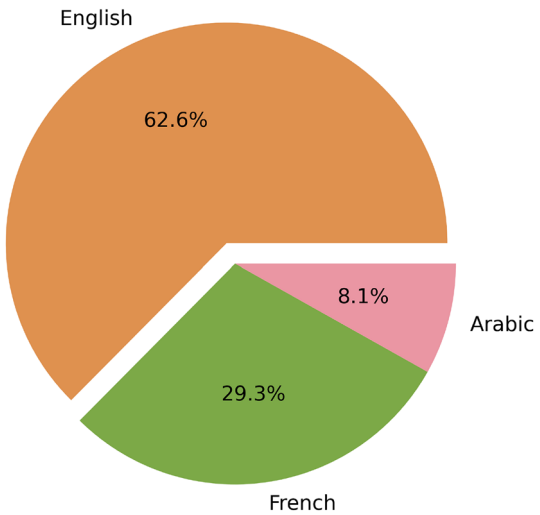

Fig. 3 Language distribution in the relevant retrieved tweets from the GeoCov-19 dataset

Table 1 Identified topics and example tweets

Topic	Tweet
Non-informative	Lol people out here worrying about the coronavirus and all I'm worried about is what pair of gloves will go well with my new knife https://t.co/CY2pR5613Y
Economics	Stocks continue nosedive Dow plunges another 1,193 points amid coronavirus fears https://t.co/IFCsclRLHD https://t.co/v8woQa1M4U
Stocking Up	RT @Royal_Creme: Scenes from my local supermarket in Basiglio, south of Milan. Panic stocking up on food because of the #coronavirus
Vaccine/Cure	Dr. Bob Sears, an anti-vax doc, announced on Facebook that his business has been slow due to #coronavirus urged patients to come in for routine annual physicals. This is despite the fact that most Californians are under a stay-at-home order. https://t.co/lDtXwbMFod #vaccines
COVID-19 statistics	In addition to Canada, France reported 6 cases coming from Egypt, and Taiwan reported 1 infected individual who visited both Egypt & Dubai (UAE). the Egyptian government still remains silent.. Could Egypt be the new Iran? https://t.co/zWtA8NcYJZ
COVID-19 information	Looks like its mutated already! #CoronaVirus is now #coronavrius!
Politics	#Iran s security official accused the #US of withholding information about an Iranian missile attack on a US base in #Iraq. The claim follows @SecPompeo's accusation that Iran is withholding information on the spread of #coronavirus https://t.co/3EKy3XoRYz
Public health measures	Any mask is better... than no mask at all... Everyone should have at least one mask and make sure you stay away from people who are coffee maker and people are washing and hand sanitizing their hands and if their set call the authorities CDC https://t.co/n1CMj0ZrM1
Governmental Measures	RIYADH: Saudi Arabia has placed a temporary ban on Umrah pilgrims in an attempt to ensure public safety by preventing the spread of the coronavirus. https://t.co/n35BdM5JJW @NAHCONCEO @HouseNGR @NGRPresident @MFA_Nigeria
Fake Treatment	RT @momblogger: Coconut oil eyed as possible treatment for #coronavirus infection
Conspiracy Theory	Federal law enforcement document reveals white supremacists discussed using coronavirus as a bioweapon https://t.co/X4T1GjqWCg

Next, we extracted a random sample from each of the three datasets and annotated them using the crowdsourcing platform Labelbox. The sample size was determined using the Central Limit Theorem (Monkey 1999) as follows:

$$Sample\ Size = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + (\frac{z^2 \times p(1-p)}{e^2 N})}$$

where:

- N is the total population size (i.e., the total number of tweets to be sampled from) with $N = 4,214,256$ for English, $N = 1,952,784$ for French, and $N = 543,558$ for Arabic,
- $e = 1.32\%$ is the margin of error, which is a percentage that describes how much the behavior of the sample is likely to deviate from the total population,
- $z = 1.96\%$ corresponding to a confidence level of 95%, with the latter being a measure of how reliable the behavior of the sample is, and
- $p = 50\%$ is the percentage value, which is the expected distribution of the tweets into the various categories, and it is advised to put it at 50% when such expected value is not known.

Our sample size rounded up to: 5, 600 tweets in English, 4, 725 tweets in Arabic, and 5, 496 tweets in French.

Each tweet in the samples above was then annotated using three different people through the Labelbox platform. More specifically, each tweet was displayed to the three annotators independently along with the 11 identified classes. The annotator was asked to assign one or more classes to the tweet based on its content. To ensure high-quality annotations, gold-standard tweets that we annotated were interjected among the tweets annotated on Labelbox without their labels. We then measured the agreement of the annotators with the ground-truth labels annotated by us on the gold-standard tweets. The annotators achieved an average of 97%, 95%, 94% accuracy with respect to the ground-truth labels (Table 2).

The distribution of the annotated tweets across the 11 topics is shown in Fig. 4. As shown in the figure, the Arabic dataset only contains 10 out of the 11 categories as none of the Arabic tweets sampled were about *Stocking Up*. The most prominent topic tweeted for the English dataset was *COVID-19 Statistics*, and the least one was *Conspiracy Theory*. On the other hand, *Governmental Measures* was the most tweeted topic for both French and Arabic datasets. In contrast, the least discussed topic was

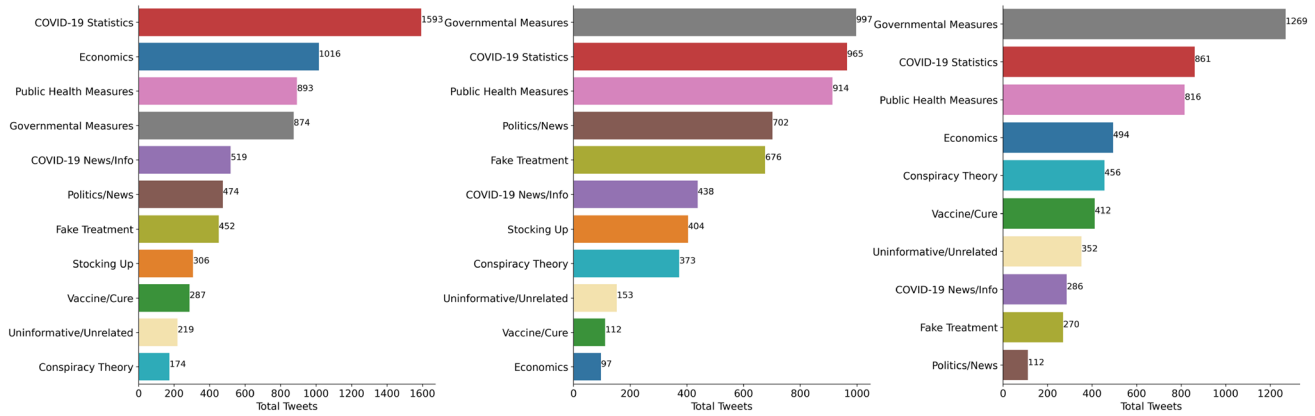

Fig. 4 Distribution of the annotated tweets across topics (English First, then French, and then Arabic)

Table 2 Annotators' agreement with ground truth

	English	Arabic	French	Average
Annotator 1	96%	97%	98%	97%
Annotator 2	93%	95%	97%	95%
Annotator 3	94%	92%	97%	94.33%

Economics in the French dataset and *Politics* in the Arabic dataset.

3.2 Models

We used our three annotated datasets described above to train various deep learning models to automatically classify a COVID-19 tweet into one or more of our 11 identified classes. Before training the classifiers, the tweets in each dataset were preprocessed using the following techniques:

1. removing non-ASCII characters for English tweets,
2. normalizing Arabic letters by deleting repeated characters and removing diacritics,
3. eliminating the HTML tags and URL links,
4. removing numbers, emojis, punctuations, and extra white spaces,
5. removing special characters (@, $, *, #),
6. normalizing English and French letters into lower case, and
7. eliminating English, Arabic, and French stopwords.

The final cleaned text for each language was then fed into three deep learning models, including a Convolutional Neural Network (CNN), a Bidirectional Long Short-Term Memory Network (BiLSTM), and a Bidirectional Encoder Representations from Transformers (BERT). For the CNN and BiLSTM models, we used the Keras library (using TensorFlow backend) on an Anaconda environment, and we used PyTorch for the BERT models. All models were trained through Python 3 on an Intel(R) Core (TM) i7 6th generation processor with 16GB RAM and Nvidia Tesla K80 GPU. We provide a detailed description for each model next.

3.2.1 CNN model

The architecture of our first model, the CNN model, is depicted in Fig. 5. CNNs are well-known for their excellent performance in various Natural Language Processing (NLP) tasks (Lai et al. 2015). As shown in Fig. 5, the CNN model consists of a word embedding layer as the first layer. We used pre-trained FastText word embeddings (Bojanowski et al. 2017) to represent a tweet as it can seamlessly handle out-of-vocabulary words. The embedding layer is followed by a 1D convolution layer, whose input dimension is $n \times 300$, where n is the number of words in the input tweet and 300 is the dimension of a word embedding vector. The convolution layer is followed by a 1D global maximum pooling layer, which is used to down-sample the features of the convolution layer. The pooling layer is followed by a number of fully connected layers and dropout layers to regularize the model and avoid overfitting. Finally, the output layer is a fully connected layer consisting of 11 Sigmoid units for English and French and 10 units for Arabic. Since each tweet can belong to one or more of our relevant classes, we use independent Sigmoid units rather than a single Softmax unit and binary cross-entropy error, the custom in multilabel classification (Glorot et al. 2011), as a loss function. For all hidden layers, we use Relu as an activation function. Note that we only use 10 units for Arabic tweets as our Arabic dataset consists only of 10 classes as explained earlier.

3.2.2 BiLSTM model

The architecture of our second model, the BiLSTM model, is depicted in Fig. 6. A BiLSTM network is a type of

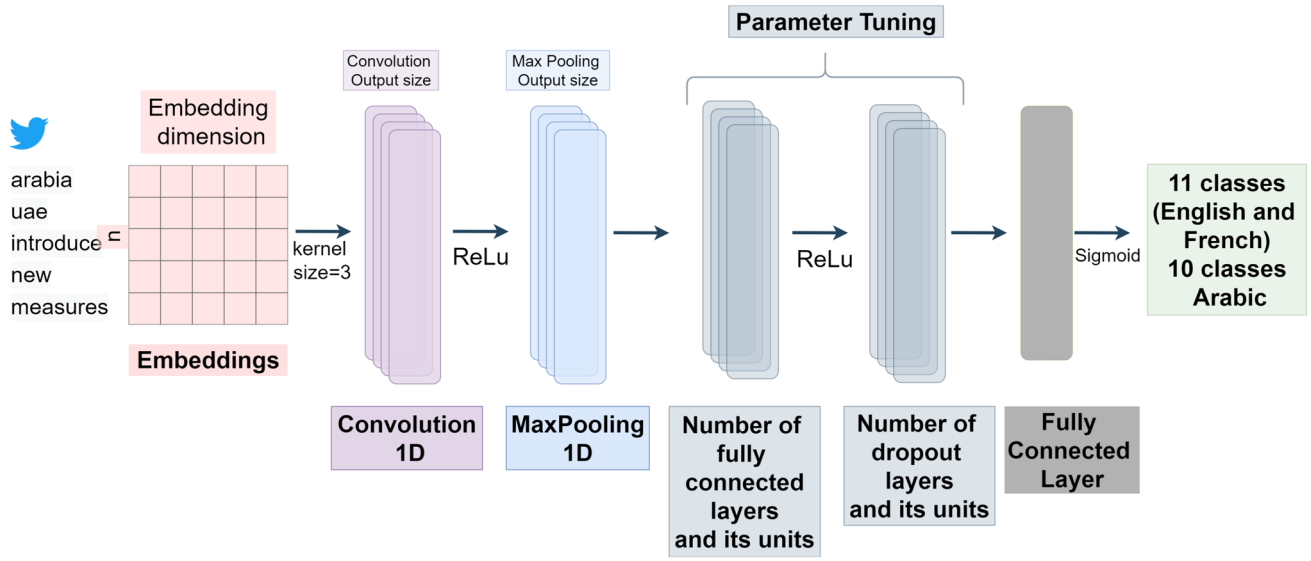

Fig. 5 CNN Model architecture

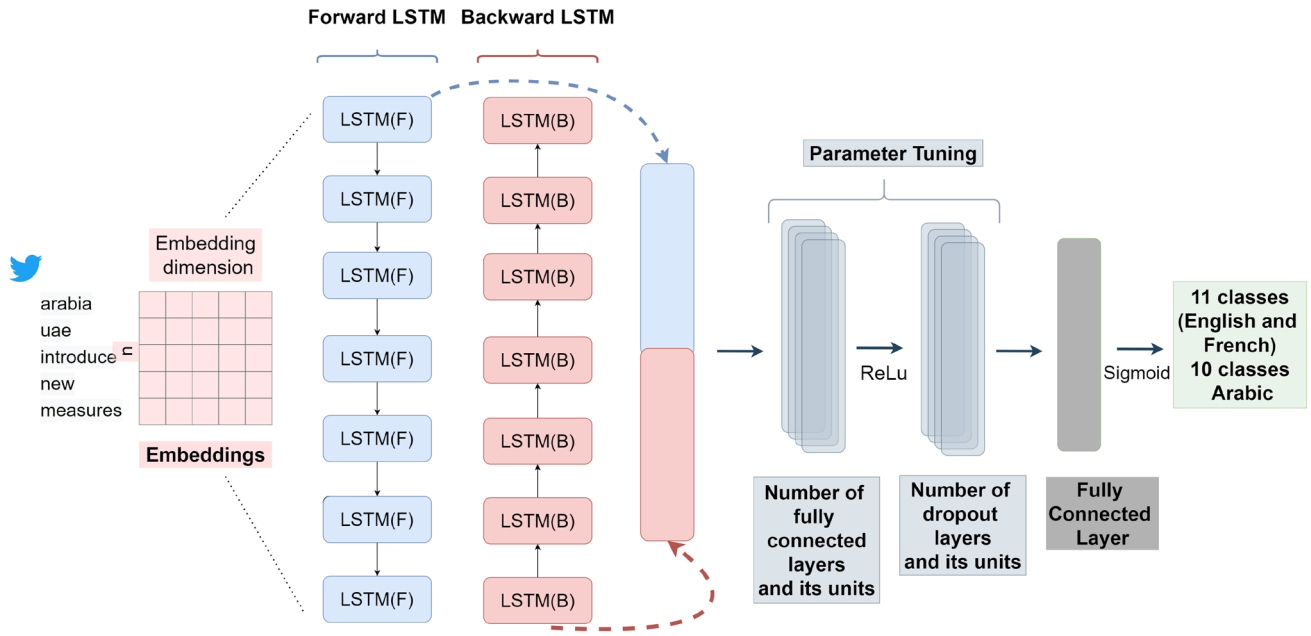

Fig. 6 BiLSTM model architecture

recurrent neural networks (RNNs) proposed by Hochreiter and Schmidhuber (1997) as a solution to the vanishing gradient problem. The model consists of a word embedding layer, and similar to the case of the CNN model, we again used pre-trained FastText word embeddings to represent a tweet. The embedding layer is then followed by a bidirectional LSTM layer, a number of fully connected layers and a number of dropout layers to regularize the model. Finally, the output layer is again a fully connected layer consisting of 11 Sigmoid units for English and French and 10 units for Arabic. Similar to the CNN model, we use Relu as an activation function for all hidden layers, and we use binary cross-entropy error as a loss function.

3.2.3 BERT model

The architecture of our third and final model, the BERT model, is shown in Fig. 7. BERT is a pretrained

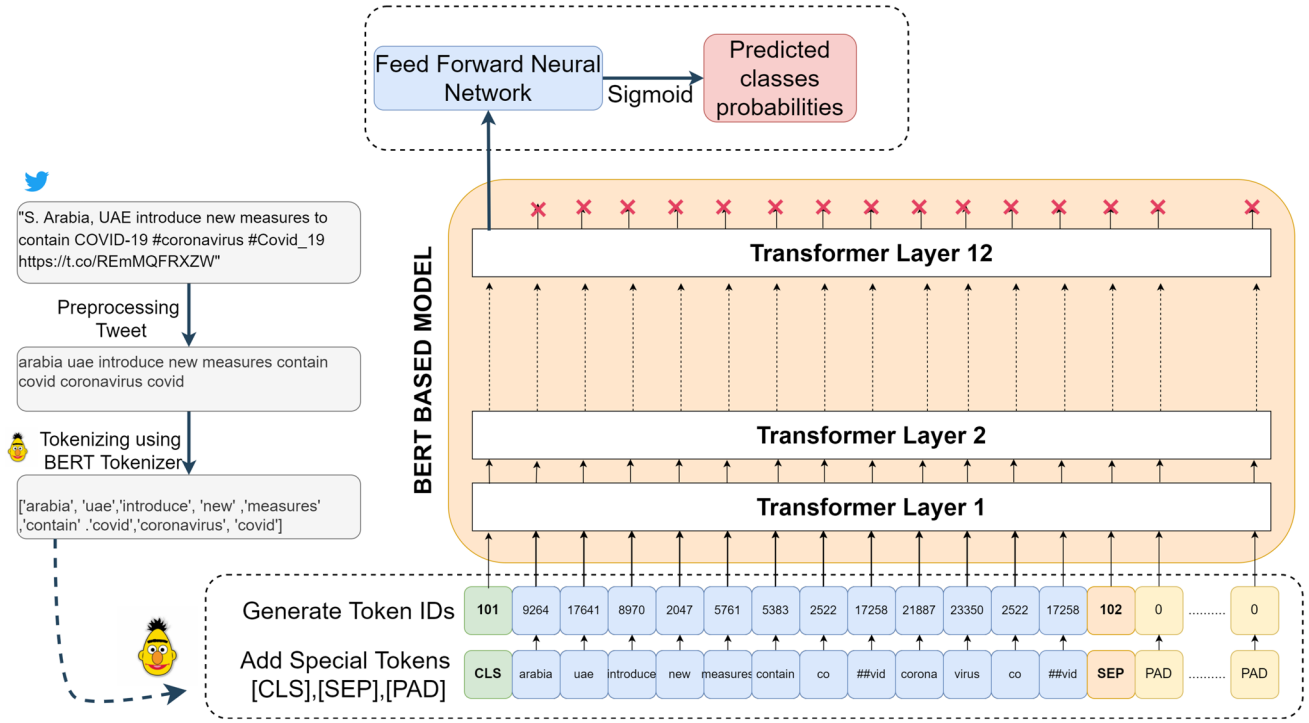

Fig. 7 The BERT model architecture

transformer model (Devlin et al. 2018b). Recently, transformers have demonstrated state-of-the-art results for various NLP tasks (Wolf et al. 2020). Unlike other deep learning models like CNNs and RNNs, transformers such as BERT use contextualized word embeddings. For the English dataset, our BERT model is based on the BERT_{base} uncased model (Devlin et al. 2018a), which is a pretrained transformer model composed of 12 attention layers, 12 attention heads, and with a hidden layer dimension of 758 and a maximum sequence length of 100. For the Arabic and French datasets, we employed Arabic-BERT (Safaya et al. 2020) and CamemBERT (Martin et al. 2020) pretrained transformer models, respectively. The two models have the same architecture as the BERT_{base} model used for the English dataset.

We fine-tuned our three BERT models using our three labeled datasets. To tokenize the input tweets, we used BERT Tokenizer for the BERT_{base} uncased model and we used WordPiece tokenizer (Wu et al. 2016) both the Arabic-BERT and the CamemBERT models. We then added special tokens typically used for fine-tuning BERT models as follows:

- The [CLS] token at the beginning of each tweet to mark the start of the input sequence
- The [SEP] token at the end of each tweet to mark the end of the input sequence
- The [PAD] token to maintain a uniform input sequence length across the entire training data

To be able to use the BERT models to classify a tweet into one of our identified classes, we add a fully connected layer consisting of a number of Sigmoid units equal to the number of classes and we use binary cross-entropy error as a loss function.

3.3 Training, validation and testing

For each of the above three described models, we train three different versions, one for each language (i.e., English, French and Arabic). Each one of our three datasets corresponding to the three languages was divided into 80% training, 10% validation, and 10% test sets. All the models were trained using their corresponding training data with TensorFlow (Abadi et al. 2015) and their hyperparameters tuned using Keras Tuner (O'Malley et al. 2019). Table 3 shows the best hyperparameters for the CNN and the BiLSTM models. For the BERT models, we used adaptive moment estimation (AdamW) (Loshchilov and Hutter 2017) as a learning algorithm with a learning rate of 2e-5, batch size equals to 32 and a max sequence length of 100. The rest of the hyperparameters of the BERT models were set to their default values.

Table 3 Best hyperparameters for the CNN and the BiLSTM models

Dataset	Models	# Hidden Layers	#HL units	# Dropout Layers	Dropout rate	# LSTM units	Learning rate
English	CNN	1	64	2	0.34	None	0.01
	BiLSTM	None	None	3	0.4	256	4.5e−05
French	CNN	1	64	2	0.3	None	0.001
	BiLSTM	None	None	1	0.77	64	0.0005
Arabic	CNN	1	32	2	0.41	None	0.01
	BiLSTM	1	16	2	0.72	64	0.004

Table 4 Test performance of the different models

Dataset	Model	Accuracy	J-Accuracy	F1-Score	LRAP	Hamming Loss
English	CNN	0.77	0.68	0.76	0.89	0.05
	BiLSTM	0.79	0.72	0.78	0.90	0.04
	$BERT_{base}$	**0.80**	**0.81**	**0.84**	**0.92**	**0.04**
French	CNN	0.81	0.78	0.84	0.93	0.03
	BiLSTM	0.78	0.75	0.80	0.90	0.04
	CamemBERT	**0.85**	**0.85**	**0.87**	**0.93**	**0.03**
Arabic	CNN	0.75	0.67	0.75	0.87	0.05
	BiLSTM	0.70	0.64	0.70	0.83	0.06
	Arabic-BERT	**0.78**	**0.79**	**0.81**	**0.89**	**0.04**

To choose our best performing model for each dataset, we use the test data and we report the performance of each model using various multilabel classification metrics such as accuracy, Jaccard accuracy, Micro-averaged F1-score, label ranking average precision score (LARP), and Hamming loss (Yang et al. 2020). As shown in Table 4, the transformer-based models, $BERT_{base}$, CamemBERT and Arabic-BERT outperformed the CNN models and the BiLSTM models for all three datasets in terms of all metrics. We thus use the BERT based models to conduct a large-scale analysis of COVID-19 tweets in the Arab region, which we describe in the next section.

4 Large-scale analysis

To study the discourse around COVID-19 in the Arab region, we use the TBCOV dataset (Imran et al. 2021), an extension of the GeoCoV-19 dataset we sampled and annotated to train our models. TBCOV consists of two billion multilingual and geolocated tweets about COVID-19 spanning from February 1, 2020, until March 31, 2021. Each tweet in the TBCOV dataset is associated with a sentiment label (positive, negative or neutral), a sentiment score, and a confidence level of that score, which were all obtained using the XLM-T model (Barbieri et al. 2021), a transformer-based model. In addition, some of the tweets are associated with an author's gender (Female or Male). Therefore, only tweets having a gender label were included in the analysis. Finally, since the focus of our study is on the Arab region, we extracted the tweets emerging from the Arab region using their geolocations. We ended up with 10, 635, 996 tweets in English, Arabic, and French.

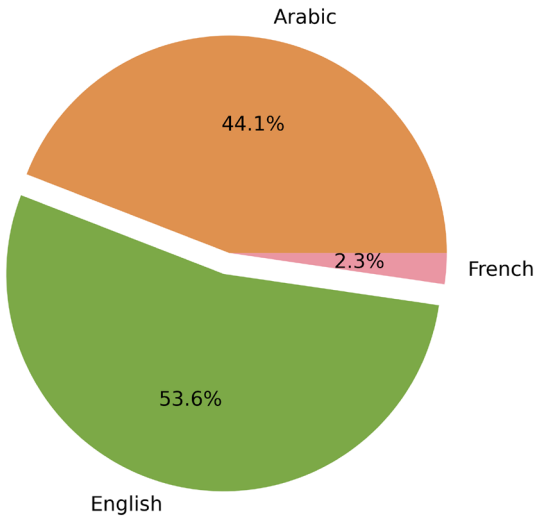

Fig. 8 Distribution of analyzed tweets per language

Figure 8 shows the distribution of the analyzed tweets per language. As can be seen from the figure, the majority of the tweets in the TBCOV dataset coming from the Arab region were in English with a total of around 5.7

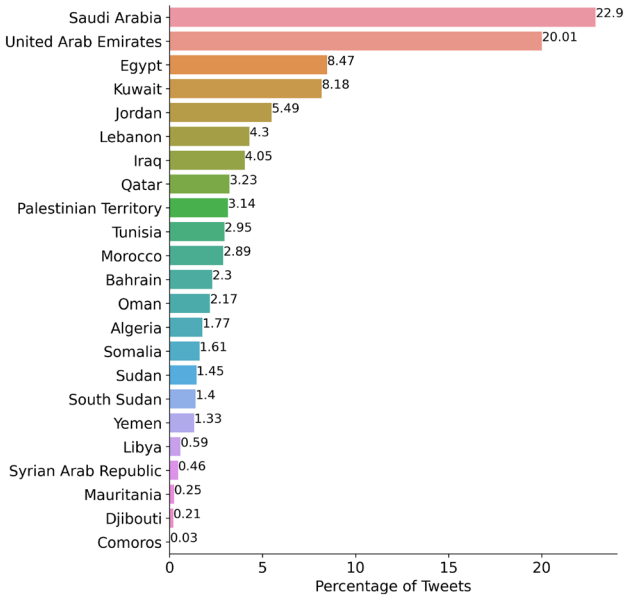

Fig. 9 Distribution of analyzed tweets per country

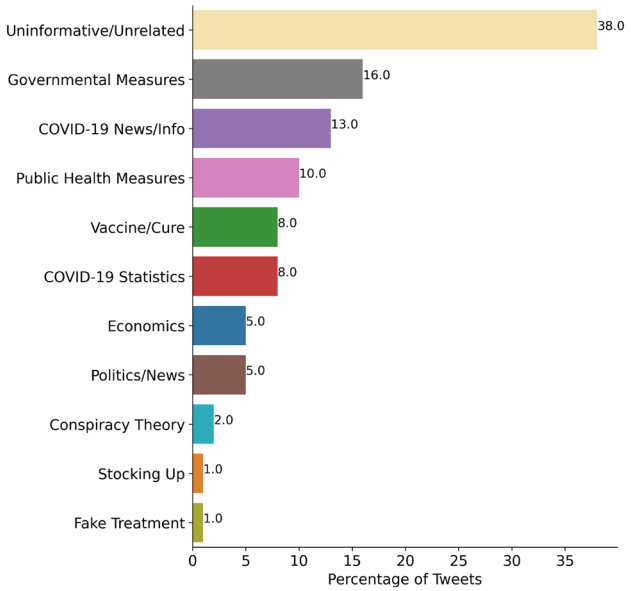

Fig. 11 Distribution of analyzed tweets per topic

Fig. 10 Distribution of analyzed tweets per country normalized by the country's population

million tweets, followed by Arabic, with a total of about 4.7 million tweets, whereas French tweets only constitute a small chunk of all the tweets, with around 250,000 tweets in total. This coincides with the fact that the most prominently used languages in the Arab region are English and Arabic, and French being only frequently used in small part of the region, namely Tunisia, Algeria and Morocco, which are all former French colonies. Figure 9 shows the distribution of the analyzed tweets per country and as can be seen from the figure, the largest amount of the tweets took place in Saudi Arabia (2,435,206), then the United Arab Emirates (2,128,522), followed by Egypt(900,814), then Kuwait (869,571), then Jordan(584,391) and Lebanon (457,466). Since the population size varies differently among the different countries in the Arab region, with Egypt being the country with the most number of people, we also display the distribution of the analyzed tweets normalized by each countries population in Fig. 10. The figure shows the number of tweets per 100,000 people in each Arab country, which is computed as follows: $(\#\ tweets\ per\ country / country's\ population) \times 100,000$. As can be seen from the figure, the United Arab Emirates had around 21k tweets per 100,000 people, followed by Kuwait, Bahrain, Qatar, Saudi Arabia, Palestine, and Lebanon, which had 20k, 14k, 12k, 7k, 7k, 6.7k tweet per 100,000 people, respectively.

We next applied our best performing models, the BERT models described in the previous section, over all the analyzed tweets to classify each tweet into one or more of our 11 identified topics. The distribution of the analyzed tweets per topic is shown in Fig. 11. As can be seen from the figure, the largest portion of the analyzed tweets (32%) were classified as *Non-Informative* by our BERT models, and the topics with the least amount of tweets were *Conspiracy Theory*, *Stocking Up* and *Fake Treatment* with about 1–2% of the total number of analyzed tweets. The high percentage of non-informative tweets indicates that people in the Arab region were mostly tweeting about personal topics related to the pandemic. Nonetheless, the analyzed tweets still contain a good representation of other topics, such as *Governmental Measures*, *Public Health Measures*, *COVID-19 News/Information*, *Vaccines/Cure*, *Economics*, and *Politics*, with a percentage ranging between 5–16% of the overall number of analyzed tweets.

Figure 12 shows the tweets distribution per country for each topic separately, normalized by each country's population. The countries contributing the most in tweeting about all the 11 topics were the United Arab Emirates, Qatar,

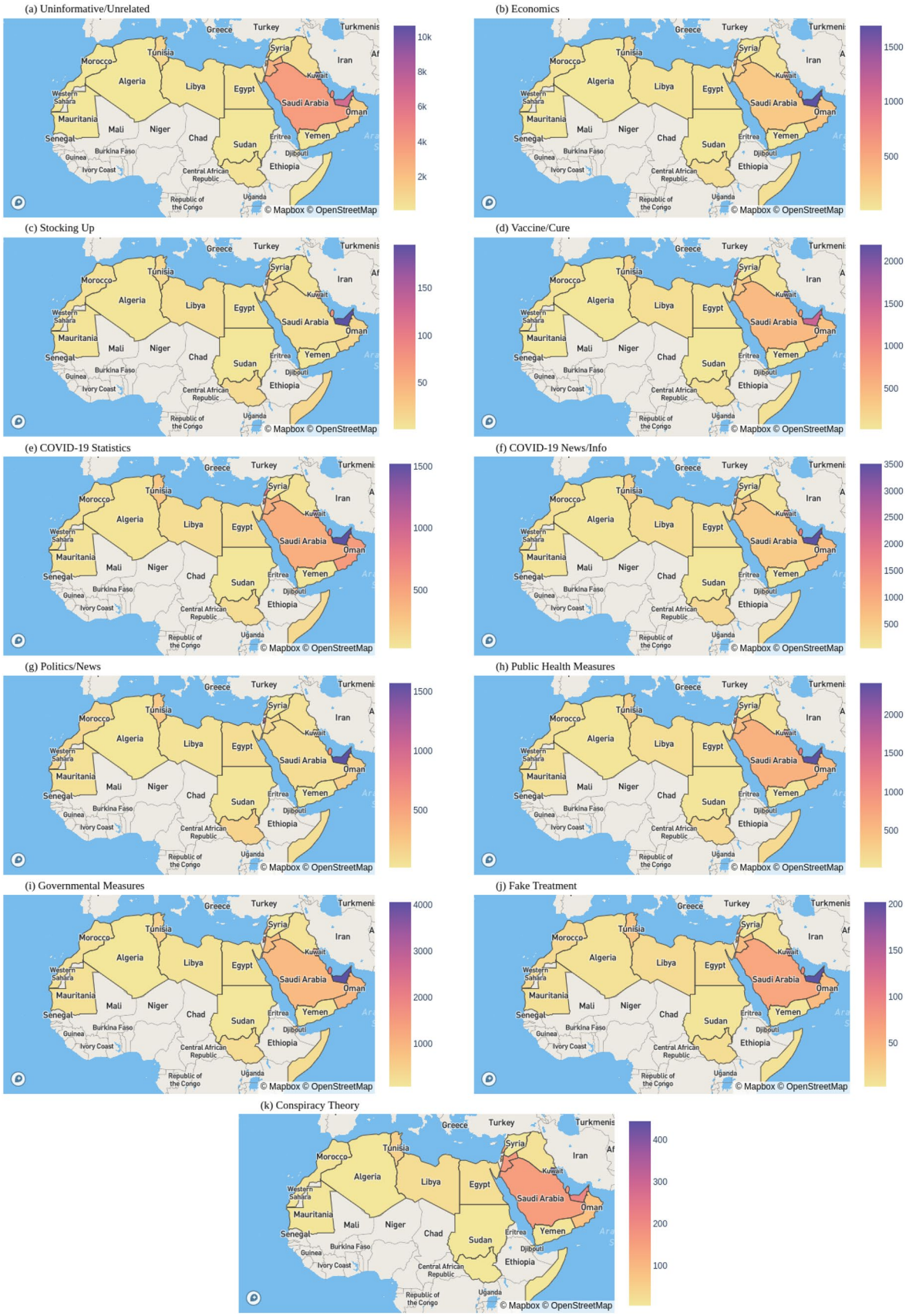

Fig. 12 Distribution of the analyzed tweets per country for each topic normalized by the country's population

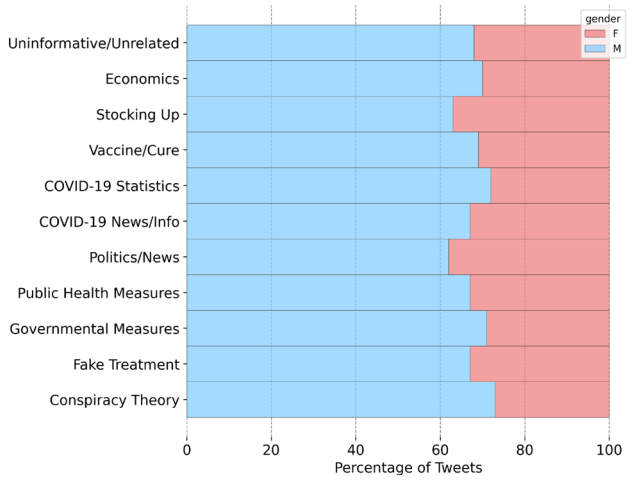

Fig. 13 Distribution of analyzed tweets per gender

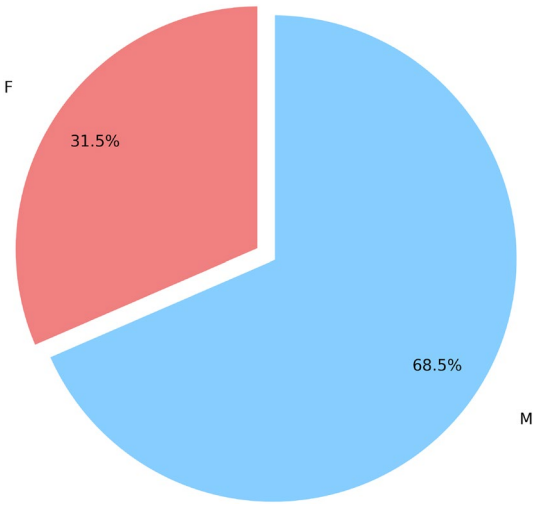

Fig. 14 Gender distribution

Kuwait, Palestine, Lebanon, Oman, Saudi Arabia, and Jordan, in that order.

We further analyzed the topic distribution per gender as can be seen in Fig. 13. The males, which represents an overall higher percentage in the dataset (68.5%) as illustrated in Fig. 14, were more actively tweeting about all the topics compared to females. The topics the males tweeted about the most compared to females were *Conspiracy Theory* (i.e., 73% of the conspiracy theory tweets were by males), followed by *Governmental Measures*, *COVID-19 Statistics*, and *Economics* with percentages ranging between 70–72%. In contrast, the females were more active in tweeting about *Politics*, *Stocking Up*, and *Fake Treatment*, with percentages ranging between 33–38% compared to males.

Figure 16 depicts the monthly trends of the topics. As can be seen from the figure, tweeting about all topics continued to increase during the first few months of the pandemic, starting from Feb 2020 and peaking in April 2020. We then observe a persistent decline in tweeting starting from April 2020 until Oct 2020, where the tweeting about all the topics begin to rise again for all the topics except for *Stocking Up*. Finally, a sudden increase in tweeting about the *Vaccine* topic can be noticed after Oct 2020, while tweeting about the other topics continued to plateau. This coincides with the development of the pandemic, where news about vaccines started to break in late 2020, and where panic buying and governmental measures started to decline in most countries.

We also performed sentiment analysis to understand the overall sentiment of the Arab population towards the pandemic. Recall that each tweet in the TBCOV is associated with a sentiment label, which we used for our sentiment analysis. Figure 17 shows the total number of positive, negative and neutral tweets per month. As expected, the negative sentiment dominates until September 2020. A significant rise of the negative sentiment is noticeable at the beginning of March, peaking in April and then declining during the later months. The negative sentiment peaks again twice after April 2020, but never reaching as high as the peak in April. On the other hand, the positive sentiment remained lower than the negative and neutral ones until September 2020. After that, the positive sentiment dominates, coinciding with news about potential vaccines for COVID-19. Finally, the neutral sentiment stays lower than the negative one and higher than the positive one until September 2020, and then, it starts to be dominated by the other two sentiments.

We also study the distribution of the different sentiments per gender. Overall, the largest percentage of the analyzed tweets were associated with a negative sentiment (37.8%), compared to 32.7% neutral and 29.5% positive. Breaking this down by gender, we observe the same trend, where tweets by both males and females were more associated with negative sentiment than neutral or positive ones as can be seen in Fig. 18.

Figure 19 shows the total number of positive, negative and neutral tweets per month for each of our three relevant languages (English on top, followed by Arabic and then French). Interestingly, the Arabic tweets show a dominance of the positive sentiment throughout the 14 months except for June 2020 and a few weeks in the beginning. On the other hand, the negative sentiment dominates the other two sentiments for English and French tweets with peaks in April and May 2020, respectively.

Breaking down the sentiment by topic, we observe an overwhelming negative sentiment for *Politics* (71.65%), *Stocking Up* (54.9%), *Government Measures* (52.3%), and *COVID-19 News/Info* (50.52%), with other topics being dominated by the neutral and positive sentiments as can be seen in Fig. 20. Breaking down the sentiment analysis by country (Fig. 21) shows an overwhelming negative

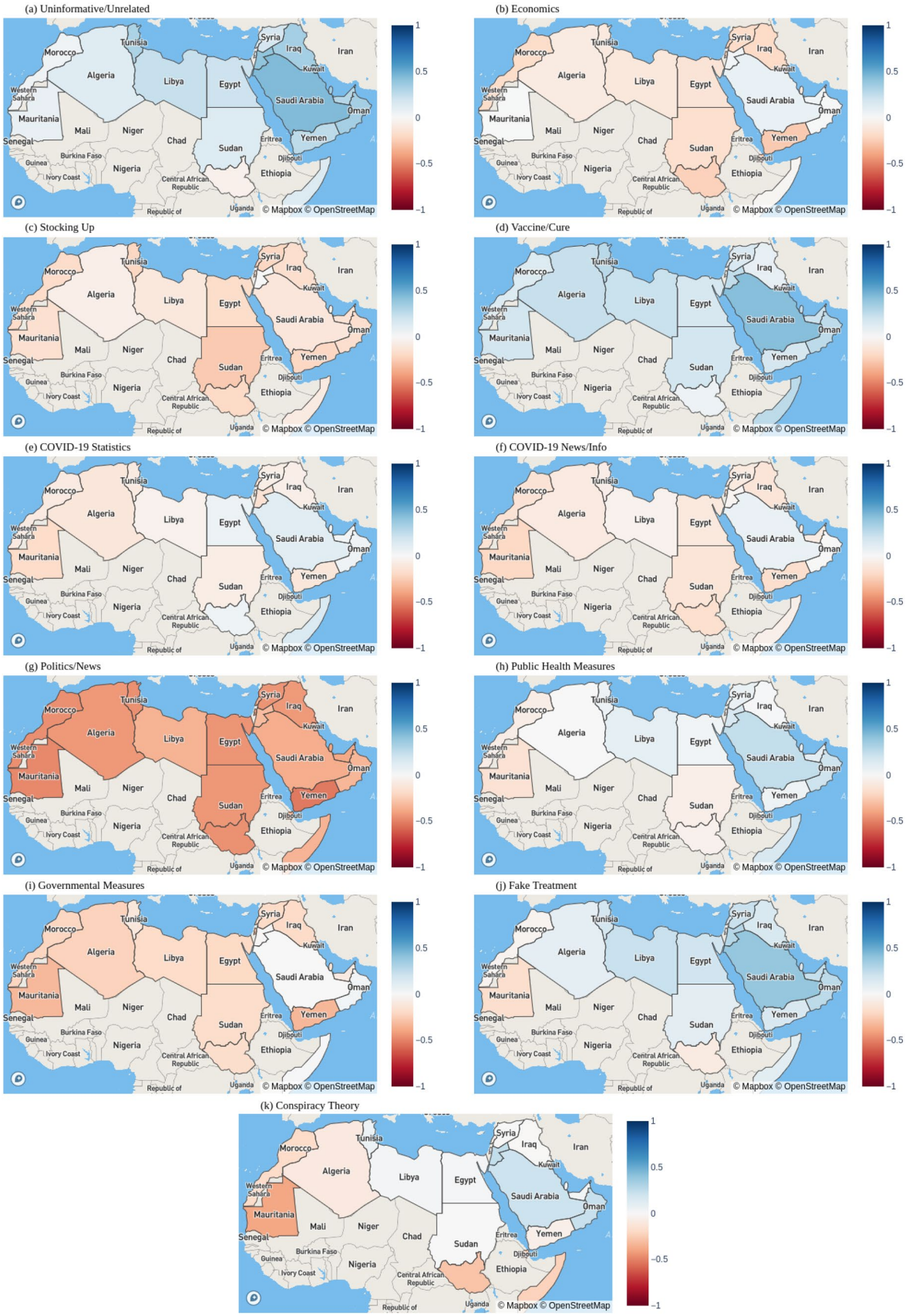

Fig. 15 Distribution of normalized sentiment scores per country for each topic

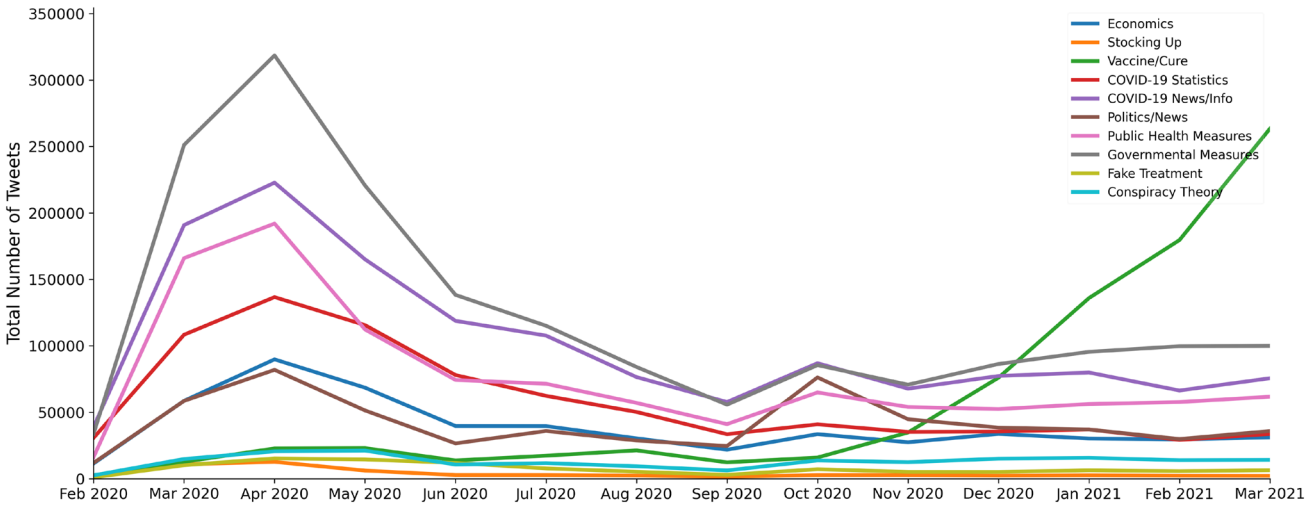

Fig. 16 Monthly trends of the topics

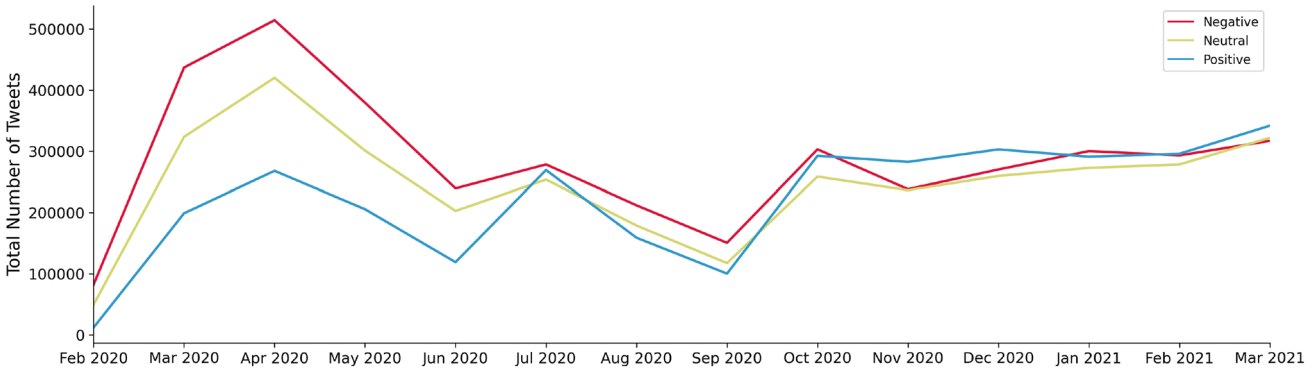

Fig. 17 Monthly trends of sentiment

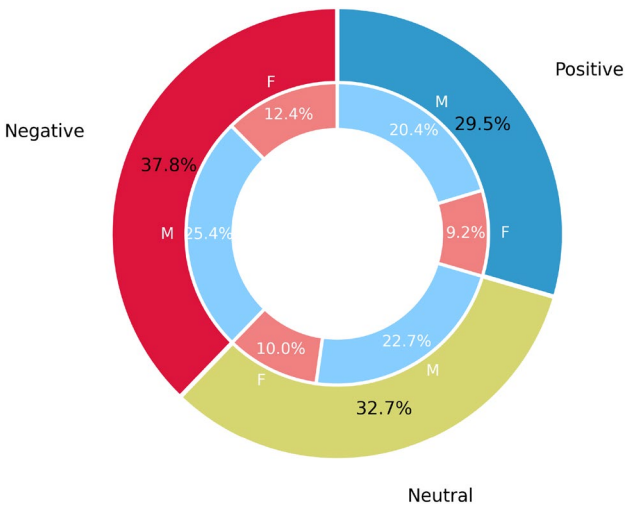

Fig. 18 Distribution of sentiment per gender

sentiment for tweets across the majority of the Arab countries. Surprisingly, Saudi Arabia, Kuwait, Bahrain and Jordan had a higher positive sentiment than negative (41.7%, 36.5%, 36.5% and 35.6%, respectively). The rest of the Arab countries, including Mauritania, Morocco, South Sudan and Palestine, show moderate to strong negative sentiment.

We also compute a sentiment score normalized by the total number of tweets disseminated from each country as follows:

$$S_c = \frac{\sum_{t_i^c\{pos,neut\}} \phi_i^c - \sum_{t_i^c\{neg\}} \phi_i^c}{N_c}$$

where S_c is the normalized sentiment score of country c, t_i^c is the sentiment score of tweet i form country c, ϕ_i^c is the confidence level for t_i^c, and N_c the total number of tweets per country c. The normalized sentiment scores for all the Arab

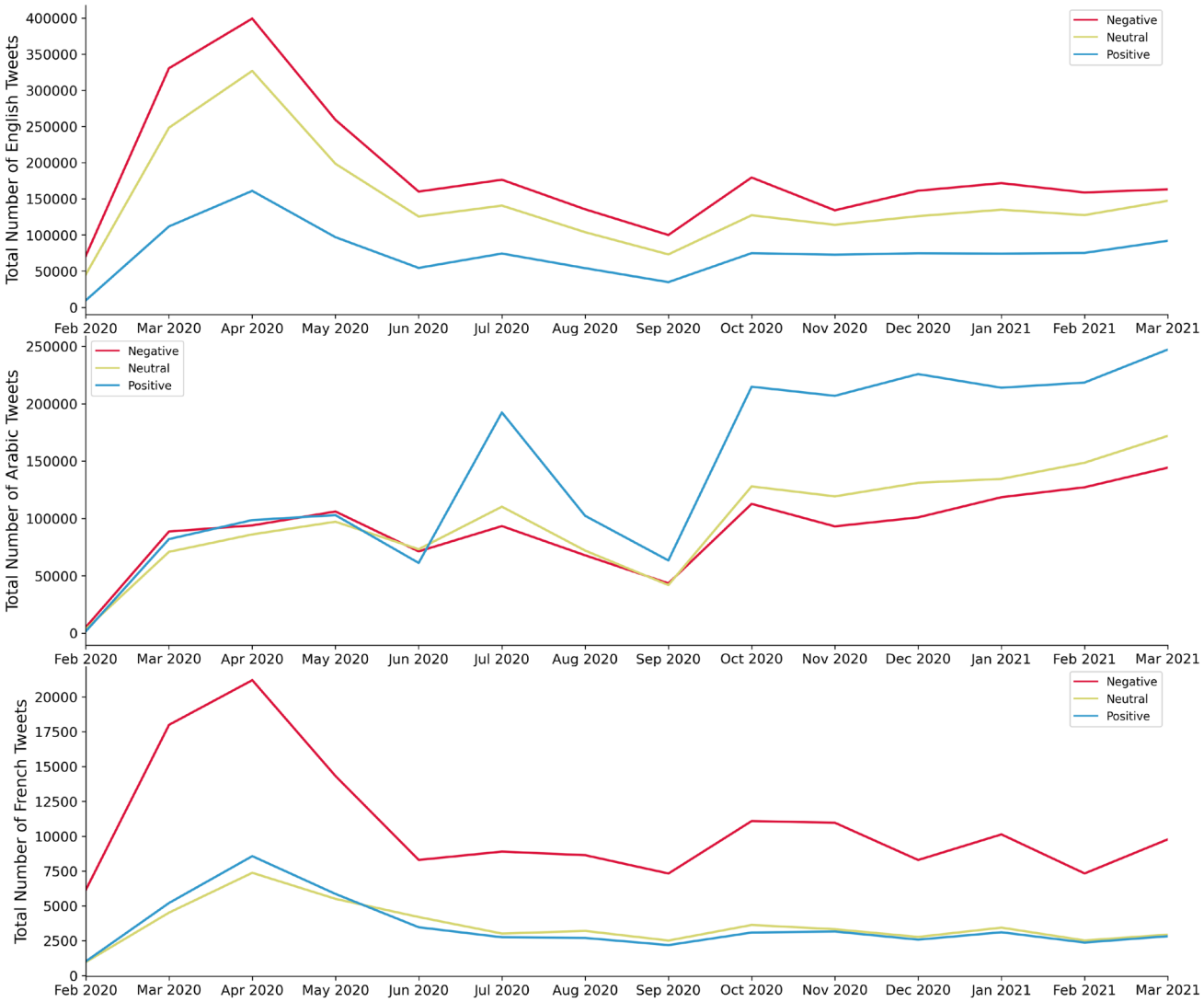

Fig. 19 Monthly trends of sentiment per language

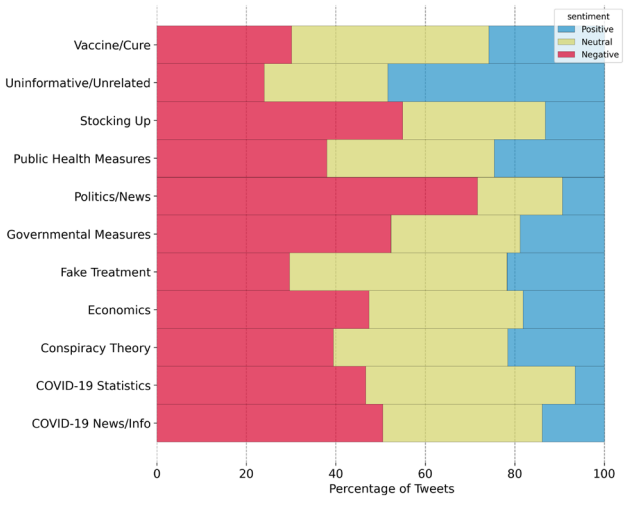

Fig. 20 Distribution of sentiment per topic

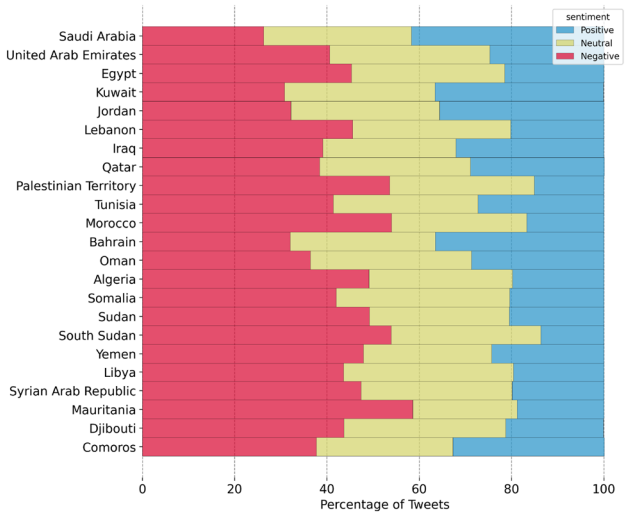

Fig. 21 Distribution of sentiment per country

Fig. 22 Normalized sentiment scores per country

countries are shown Fig. 22. As can be seen from the figure, the normalized sentiment scores of Saudi Arabia, Kuwait, Jordan, and Bahrain show a significant positive sentiment with normalized scores of 0.31, 0.24, 0.22, 0.22, respectively. This confirms the results shown in Fig. 21.

Finally, we also show the normalized sentiment score distribution per country for each of the 11 topics separately in Fig. 15. Overall, topics having *neutral, or slightly to highly negative sentiments* across all countries include *Economics*, *Stocking Up*, *Politics*, and *Governmental Measures* where the highest overall negative sentiment is observed in the political discourse. Topics showing *slightly positive, neutral to slightly negative sentiment* across all countries, except for Mauritania, South Sudan, and Somalia, are *Public Health Measures*, *Fake Treatment*, *COVID-19 Statistics*, *COVID-19 News*, and *Conspiracy Theory*. On the other hand, the *Vaccine* topic's sentiment varied between *highly positive and neutral*. For example, Saudi Arabia shows a high positive sentiment for the vaccine topic, whereas Palestine, South Sudan, and Iraq seem to discuss the topic more neutrally.

5 Conclusion

In this paper, we presented a large-scale analysis of COVID-19 tweets in the Arab region. First, we identified 11 general topics under which COVID-19 tweets emerging from the Arab region fall, through an intensive sampling of tweets from the GeoCoV-19 dataset, which consists of over 6 million geolocated COVID-19 tweets spanning the period from February 1, 2020, to April 30, 2020, in English, Arabic, and French. Next, we generated training data consisting of 5,600 English, 4,725 Arabic, and 5,496 French tweets that were manually classified into one or more of the identified 11 topics via crowdsourcing. These training data were then used to train various deep learning models, including CNN, BiLSTM, and BERT, to automatically classify a tweet into one or more of the 11 topics. Our best performing model BERT outperformed all other models for all languages in terms of various multilabel classification metrics.

We then used our BERT models to perform a large-scale analysis on tweets emerging from the Arab region and spanning the period from February 1, 2020, to March 31, 2021, which were obtained from the TBCOV dataset. Our analysis indicates that the majority of the tweets analyzed emerged from Saudi Arabia, UAE, and Egypt, and that the majority of the tweets were generated by males. The analysis also shows that males mainly tweeted about conspiracy theory and governmental measures, whereas females mainly tweeted about politics, stocking up, and fake treatment. We also observed a surge in tweeting about all the topics as the pandemic broke followed by a slow and steady decline over the following months, except for a sudden surge in tweets about vaccines after October 2020, which continued to increase onwards. We finally performed sentiment analysis on the analyzed tweets, which indicated a strong negative sentiment until mid of September 2020, after which we observed a strong positive sentiment that coincided with the surge in tweeting about vaccines. Overall, the analysis showed that positive sentiment increased over time with some countries having predominantly positive sentiment compared to negative or neutral ones such as Saudi Arabia, Kuwait, Bahrain and Jordan.

In future work, we aim to conduct more similar studies on later periods (i.e., after March 2021) and to compare our insights from the analysis of the Arab region discourse around COVID-19 on Twitter to other regions in the world. We also plan to involve public health practitioners and social scientists to better understand our insights and explore ways of utilizing them for the public good.

Acknowledgements This work is supported by the American University of Beirut Research Board (URB).

Author Contributions All authors contributed equally to this work.

Funding This work was funded by the American University of Beirut Research Board (URB).

Data Availability No data with regard to the direct content of tweets will be published or forwarded in line with the Developer Agreement and Policy of the Twitter API. All other data used in this work are publicly available, and their sources are referenced in the manuscript.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this manuscript.

Ethics approval This work did not include any experiments involving humans or animals.

Code availability Code is available upon request.

References

Abadi M, Agarwal A, Barham P, et al (2015) TensorFlow: large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, software available from tensorflow.org

Abd-Alrazaq A, Alhuwail D, Househ M (2020) Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. J Med Internet Res 22((4)):e19,016

Alam F, Dalvi F, Shaar S, et al (2020) Fighting the COVID-19 infodemic in social media: a holistic perspective and a call to arms. arXiv preprint arXiv:2007.07996

Aljabri M, Chrouf SM, Alzahrani NA et al (2021) Sentiment analysis of arabic tweets regarding distance learning in Saudi Arabia during the COVID-19 pandemic. Sensors 21(16):5431

Alqurashi S, Hamoui B, Alashaikh A, et al (2021) Eating garlic prevents COVID-19 infection: detecting misinformation on the arabic content of twitter. arXiv preprint arXiv:2101.05626

Alsudias L, 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, Seattle, WA, USA, 9–10 July 2020

Ameur MSH, Aliane H (2021a) Aracovid19-mfh: Arabic COVID-19 multi-label fake news and hate speech detection dataset. arXiv preprint arXiv:2105.03143

Ameur MSH, Aliane H (2021b) AraCovid19-ssd: Arabic COVID-19 sentiment and sarcasm detection dataset. arXiv preprint arXiv:2110.01948

Barbieri F, Anke LE, Camacho-Collados J (2021) Xlm-t: A a multilingual language model toolkit for twitter. arXiv preprint arXiv:2104.12250

Bojanowski P, Grave E, Joulin A et al (2017) Enriching word vectors with subword information. Transa Assoc Comput Linguist 5:135–146

Chandrasekaran R, Mehta V, Valkunde T et al (2020) Topics trends and sentiments of tweets about the COVID-19 pandemic: temporal infoveillance study. J Med Int Res 22(10):624

Devlin J, Chang M, Lee K, et al (2018a) BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805. arXiv:1810.04805

Devlin J, Chang MW, Lee K, et al (2018b) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805

Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics, JMLR Workshop and Conference Proceedings, pp 315–323.

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780

Hung M, Lauren E, Hon ES et al (2020) Social network analysis of COVID-19 sentiments: application of artificial intelligence. J Med Internet Res 22(8):e22,590

Imran AS, Daudpota SM, Kastrati Z, et al (2020) Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. Ieee Access 8:181,074–181,090

Imran M, Qazi U, Ofli F (2021) Tbcov: Two billion multilingual COVID-19 tweets with sentiment, entity, geo, and gender labels. arXiv preprint arXiv:2110.03664

Kumar S, Pranesh RR, Carley KM (2021) A fine-grained analysis of misinformation in COVID-19 tweets

Lai S, Xu L, Liu K, et al (2015) Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI conference on artificial intelligence

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101

Martin L, Muller B, Suárez PJO, et al (2020) Camembert: a tasty french language model. In: Proceedings of the 58th annual meeting of the association for computational linguistics

Memon SA, Carley KM (2020) Characterizing COVID-19 misinformation communities using a novel twitter dataset. arXiv preprint arXiv:2008.00791

Monkey S (1999) Calculating the number of respondents. https://help.surveymonkey.com/articles/en_US/kb/How-many-respondents-do-I-need

O'Malley T, Bursztein E, Long J, et al (2019) Keras Tuner. https://github.com/keras-team/keras-tuner

Qazi U, Imran M, Ofli F (2020) Geocov19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information. ACM SIGSPATIAL Special 12(1):6–15. https://doi.org/10.1145/3404111.3404114

Safaya A, Abdullatif M, Yuret D (2020) KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media. In: Proceedings of the fourteenth workshop on semantic evaluation. International committee for computational linguistics, Barcelona (online), pp 2054–2059, https://www.aclweb.org/anthology/2020.semeval-1.271

Wolf T, Debut L, Sanh V, et al (2020) Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, pp 38–45

WorldBank (2020) Arab region total population. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=1A

Wu Y, Schuster M, Chen Z, et al (2016) Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144

Xue J, Chen J, Chen C et al (2020a) Public discourse and sentiment during the COVID 19 pandemic: using latent dirichlet allocation for topic modeling on twitter. PloS one 15(9):e0239,441

Xue J, Chen J, Hu R, et al (2020b) Twitter discussions and emotions about COVID 19 pandemic: a machine learning approach. arXiv preprint arXiv:2005.12830

Yang Q, Alamro H, Albaradei S, et al (2020) Senwave: monitoring the global sentiments under the Covid-19 pandemic. arXiv preprint arXiv:2006.10842

Yin H, Yang S, Li J (2020) Detecting topic and sentiment dynamics due to covid-19 pandemic using social media. In: international conference on advanced data mining and applications, Springer, pp 610–623