

AMERICAN UNIVERSITY OF BEIRUT

A LARGE SCALE ANALYSIS OF COVID-19
TWEETS IN THE ARAB REGION

by

AYA AHMAD MOURAD

A thesis

submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Department of Computer Science
of Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
January 2022

AMERICAN UNIVERSITY OF BEIRUT

A LARGE SCALE ANALYSIS OF COVID-19
TWEETS IN THE ARAB REGION

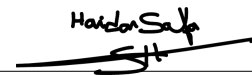
by
AYA AHMAD MOURAD

Approved by:



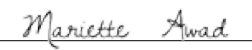
Dr. Shady Elbassuoni, Associate Professor
Computer Science

Advisor



Dr. Haidar Safa, Professor
Computer Science

Member of Committee



Dr. Mariette Awad, Associate Professor
Electrical and Computer Engineering

Member of Committee

Date of thesis defense: January 26, 2022

ACKNOWLEDGEMENTS

I would like first to thank my advisor, Dr. Shady Elbassuoni, for his constant support and motivation to present this research work. Without his guidance and mentorship, this thesis would not have been possible.

I would also like to thank my committee members, Dr. Haidar Safa and Dr. Mariette Awad, for their guidance and remarks.

I would also wish to express my gratitude to my family, especially my mom and dad, for their endless love and support throughout my graduate studies. A big thanks, to my sister Roya, for her help and for being by my side through this journey. To my friend Ghinwa, thank you for all the moments we spent together through this process.

ABSTRACT

OF THE THESIS OF

Aya Ahmad Mourad for Master of Science
Major: Computer Science

Title: A Large Scale Analysis of COVID-19 Tweets in the Arab Region

Since the first case was discovered in Wuhan, China, in December 2019, the coronavirus disease (COVID-19) has caused harm worldwide. It has spread rapidly to the Arab World, affecting public health, the economy, and mental health. To combat its spread, the Arab governments have announced many states of emergency and curfews. As a result, most people started communicating about the pandemic through social media platforms such as Twitter. This thesis proposes a suite of text mining tasks to extract useful insights into people's perceptions and reactions to the pandemic. We have identified 11 relevant topics based on an intensive sampling of randomly selected tweets from a large dataset consisting of 6,710,598 spanning from February 1, 2020, to April 30, 2020, combined with extensive literature review. The tweets in the dataset are geolocated multilingual

tweets emerging from the Arab region in English, Arabic, and French. Consequently, we defined an annotation schema to classify the tweets into misinformative and fine-grained informative tweets consisting of 10 different classes. The resulting labeled datasets composed of 5600 English, 4725 Arabic, and 5496 French tweets were then fed to different deep learning and transformer models, including CNN, BiLSTM, and Bert, to conduct multi-label classification. The models' performance evaluation shows that the BERT-based model outperformed deep learning models in classifying English, multi-dialect Arabic, and French tweets with an F1-Micro score of 0.84, 0.81, and 0.87, respectively. We also tested the BERT-based models and performed a large-scale analysis on an unlabeled dataset that spans from February 1, 2020, to March 31, 2021. The tweets distribution was the highest in Saudi Arabia (23%), UAE (20%), and Egypt (8%). The analysis by gender shows that Arab region males mainly discussed conspiracy theory and governmental measures topics, making up 68.5% of the total tweets. The topics debated showed a remarkably similar pattern of the rapid rise and slow decline across the region. A sudden surge in the vaccine topic was noticed after Oct 2020 and continues to increase afterward. The Arab region conversation reacts strongly negatively until mid of Sep 2020, where the positive sentiment starts dominating, coinciding with the vaccine topic's discussion period. Overall, the analysis shows that optimistic feelings increased over time. Surprisingly, Saudi Arabia (41.7%) and other countries, including Kuwait (36.5%), Bahrain (36.5%), and Jordan (35.6%), had higher positive sentiment than negative.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	1
ABSTRACT	2
ILLUSTRATIONS	6
TABLES	8
1 INTRODUCTION	9
2 LITERATURE REVIEW	13
3 Multi-Label Classification	17
3.1 Datasets	18
3.2 Classifiers	24
3.2.1 Convolutional Neural Network	27
3.2.2 Bidirectional Long Short-Term Memory	28
3.2.3 Bidirectional Encoder Representations	31
3.2.4 Experiments and Results	34
4 Analysis	39
4.1 Tweets Statistics and Distribution	40

4.2	Predicted Topics Using BERT-based Model Statistics	41
4.2.1	Topics Distribution	41
4.2.2	Topics Trend Analysis	42
4.3	Sentiment Analysis	44
5	Conclusion	51
A	Abbreviations	53
B	List of Tables	54
	Bibliography	58

ILLUSTRATIONS

3.1	The methodology followed in this study	18
3.2	Sampled dataset languages distribution	19
3.3	The study area of the work: arab region countries per population	20
3.4	Annotated tweets topic distribution: English, French, and Arabic	24
3.5	The architecture of the proposed CNN model for multi-label classification	28
3.6	LSTM cell structure	30
3.7	The architecture of the proposed BiLSTM model for multi-label classification	31
3.8	The Architecture of BERT Model	32
3.9	The Architecture of Transformers	33
3.10	The architecture of the proposed BERT model for multi-label classification	35
4.1	Tweets Distribution Per Language	41
4.2	Tweets Distribution Per Country	41
4.3	Geotagged tweets in the Arab Region normalized by country's population (per 100,000 persons)	42
4.4	Distribution of tweets' sentiment per gender	43

4.5	Gender Distribution	43
4.6	Monthly trends of the identified topics	43
4.7	Monthly trends of the sentiments	44
4.8	Monthly trend of sentiments per language	45
4.9	Distribution of tweets' sentiment per gender	46
4.10	Arab Region sentiment based on normalized scores of the labeled sentiment in each country	47
4.11	Topic Sentiments Distribution	48
4.12	Sentiments distribution per country	48
4.13	Topic Distribution	48
4.14	Geotagged Topics in the Arab Region normalized by country's pop- ulation (per 100,000 persons)	49
4.15	Topics sentiment in the Arab Region based on normalized scores of the labeled sentiment in each country	50

TABLES

3.1	Annotators benchmark score	20
3.2	Tweets topics examples	25
3.3	Comparison between BERT transformers	34
3.4	Deep learning models best parameters experimental settings	36
3.5	Performance of the proposed models	38
B.1	List of geotagged weighted average score sentiments in the Arab Region	54
B.2	List of geotagged tweets in the Arab Region normalized by coun- try’s population (per 100,000 persons)	55
B.3	List of normalized tweets per topic by country’s population (per 100,000 persons)	56
B.4	List of geotagged weighted average score sentiments in the Arab Region per topic	57

CHAPTER 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 [1], have been hit by the pandemic in escalating numbers. 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, implementing 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 information. 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 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 thesis 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 [2]. 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 [3] to associate each tweet in each dataset with one or more of the identified topics mentioned above. The final resulting datasets are 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, to automatically classify a tweet related to COVID-19 into one of the 11 topics we have identified. The best classifier 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 [4]. Next, 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 contributions in this thesis can thus be summarized as follows:

- We identify a set of topics that cover the spectrum of tweets about COVID-19 generated in the Arab region and build three labeled datasets in three languages that include tweets that fall under one or more of the identified topics
- Train various deep learning models to automatically label COVID-19 tweets into one or more of the identified relevant topics

- Use our deep learning model to annotate a large dataset of tweets related to COVID-19 generated by users in the Arab region and perform a large scale analysis of such annotated dataset

This thesis is organized as follows. Chapter 2 gives an overview of related work of the annotated COVID-19 datasets and the different machine and deep learning techniques employed for topic classification. Chapter 3 describes the datasets we utilized with a fine-grained definition of the annotations used. We also define the various classifiers CNN, BiLSTM, and BERT models to conduct multi-label classification. In the same chapter, we provide the results of the experiments of the parameter tuning of deep learning models and the results of the models' performances. Chapter 4 provides a large-scale analysis aggregated per time, topic, sentiments, and gender. Finally, we conclude and present future directions in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

The objective of this literature survey is to give a general overview about some important concepts related to the thesis work.

Various research has been conducted throughout the literature to track Coronavirus topics and discussions through Twitter and develop multiple deep learning models. Some have focused on specific topics. For example, Kumar et al. (2021) [5] proposed classifying tweets by manually annotating an English dataset consisting of 1970 tweets categorized into four classes as follows Irrelevant, Conspiracy, True Information, and False Information. A comparative analysis of various language models was conducted using: Convolution Neural Networks (TextCNN), Recurrent neural networks (Bi-LSTM, LSTM), three variants of BERT, three variants of RoBERTa, and two variants of ALBERT. They also proposed two ensemble deep learning models, a CNN-RNN model by stacking CNN layer over RNN layer (Bi-LSTM) and an RNN-CNN model where a single Bi-LSTM layer is employed over the top of the 1D-CNN layer. The results show that RoBERTa-large gives the best F1-score 76% among the different variants of transformer language models since it is trained on a bigger corpus than other models. As

for the deep learning models, CNN-RNN performed the best with an F1- score of 71%. On the other hand, Memon & Carley (2020) [6] only annotated and analyzed an English dataset of a total of 4573 tweets categorized into 17 different classes grouped into informative and misinformative. The classes included 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) [7] also defined an annotation schema and detailed annotation instructions to classify a total of 504 English and 218 Arabic tweets. The annotations are prepared with seven questions. The questions include the following inquiries: whether the tweet contains a fact claim, includes 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. The data was fed to three different classifiers SVM with word-based, TF-IDF, FastText, and BERT-based models. The best model for English was BERT, and for Arabic, FastText was better. Additionally, Xue et al. (2020) [8] analyzed 4 million Twitter messages related to the COVID-19 pandemic from March 1 to April 21 in 2020. They used a machine learning approach, Latent Dirichlet Allocation (LDA), to identify popular salient topics and themes. They identified 13 discussion topics and categorized them into five different themes, including public health measures, social stigma, coronavirus news cases and deaths, COVID-19 in the United States, and coronavirus cases in the rest of the world. The results show that the dominant sentiments for the spread of coronavirus are anticipation that measures can be taken, followed by a mixed feeling of trust, anger, and fear for different topics.

Here we present studies conducted across the Arab region. Many of them considered general categories for tweets topic classification over a short period of

analysis and, at a small scale assessing only the Arabic language. For clarification, Aljabri et al. (2021) [9] collected Arabic tweets about distance learning with Saudi Arabia as a study area. They focused on studying the sentiment analysis, whether positive or negative, by building different machine learning classifiers with various features extraction techniques. The best accuracy is achieved (0.899) by the Logistic regression classifier with unigram and TF-IDF as a feature extraction method. Furthermore, Alqurashi et al. (2021) [10] constructed a large Arabic dataset (8,786) related to COVID-19 and annotated the tweets into two categories: misinformation or not. The dataset was fed into eight different machine and deep learning models, with varying features, including word embeddings and word frequency. Experiments show that Extreme Gradient Boosting (XGBoost) presents the highest accuracy. Unlike Alqurashi et al. that classified the tweets only for two categories, Ameer & Aliane (2021) [11] annotated multi-label Arabic COVID-19 tweets (10,828) into ten different labels. The labels study whether the tweets contain hate, talk about a cure, give advice, raise morals, news or opinion, written in Dialect, Blame and negative speech, factual, worth fact-checking, and contain fake information. The annotated dataset is then used to train and evaluate several classification models using AraBERT and mBERT transformers. The same authors, Ameer & Aliane [12], manually annotated Arabic COVID-19 tweets of 5,162. The classes include whether the tweet is sarcastic (yes or no) and positive, neutral, and negative sentiments. The dataset was then fed into different models, including AraBERT, mBERT, and XLM-Roberta. As well, Alsudias and Rayson (2020) [13] identified topics discussed during the pandemic using the K-means algorithm with $k=5$, including COVID-19 statistics, prayers for God, COVID-19 locations, advice, and education for prevention and advertising. They also performed rumors detection by manually sampling 2000 tweets and labeling

them with 1, -1, and 0 to denote correct information, false information, and unrelated, respectively. Finally, they applied three different machine learning models Logistic regression, support vector model (SVM), and Naïve Bayes, with 84% as the highest accuracy achieved by LR.

The above-summarized work discusses various topic identification methods using machine and deep learning models. Some of the presented papers performed the classification task on limited data that is not geolocated, and others utilized general categories in the process. However, few have conducted a large-scale analysis, and most of them focused on a single language dataset. Our work fills these gaps found in the literature. We used a multilingual geolocated dataset with fine-grained classes, including 11 different topics mentioned in Chapter 1. To the best of our knowledge, there are no English, Arabic, and French COVID-19 multi-label datasets covering various topics as large and as rich as the one we are releasing in this thesis.

CHAPTER 3

MULTI-LABEL CLASSIFICATION

This chapter describes the proposed models for multi-label classification of COVID-19 tweets. First, as illustrated in Figure 3.1, we used the GeoCoV-19 dataset [2] and Twitter API [14] to retrieve all tweets related to COVID-19 from the period of February 1, 2020, to the period of April 30, 2020. We then kept 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 (Figure 3.3). Next, we sampled the tweets extracted as described above and identified 11 categories under which most of these sampled tweets fall. We then annotated three different sampled datasets of tweets in the three relevant languages using crowdsourcing. Finally, the annotated datasets were used to train various deep learning models to automatically categorize a given COVID-19 tweet into one or more of the 11 identified categories.

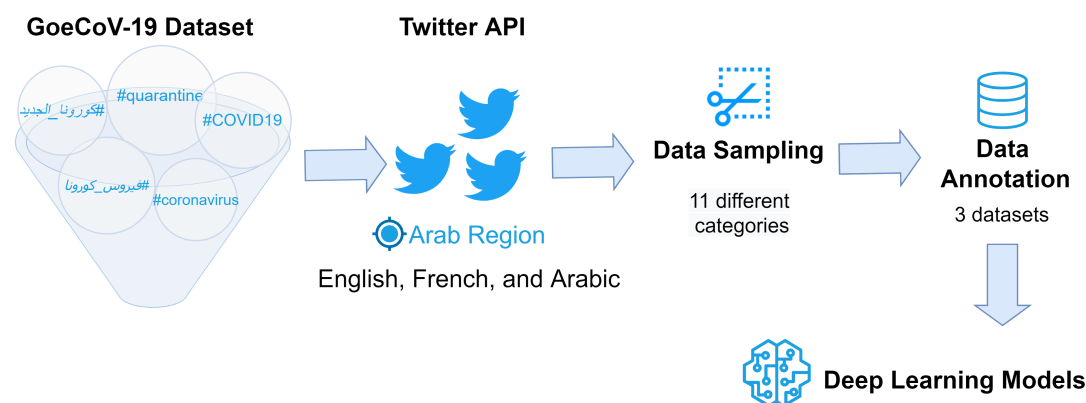


Figure 3.1: The methodology followed in this study

3.1 Datasets

Our datasets are all retrieved from the GeoCoV-19 dataset[2]. It contains more than 524 million multilingual tweets about COVID-19 collected from February 1, 2020, until April 30, 2020, geolocated 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 (Figure 3.3)

Adhering to Twitter data redistribution policies, GeoCoV-19 doesn't share full tweets content. 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 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 Figure 3.2 among the three prominently spoken languages in the region (i.e., Arabic, English, and French).

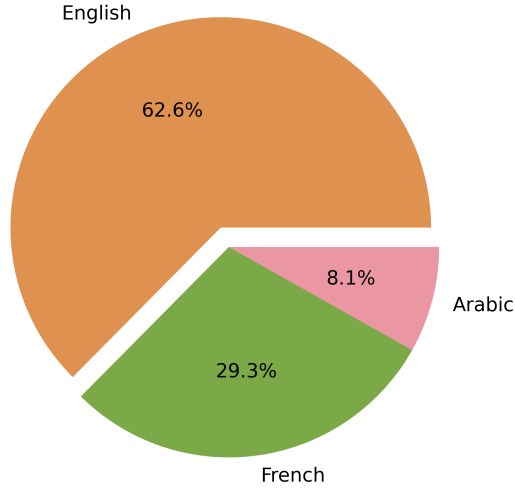


Figure 3.2: Sampled dataset languages distribution

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 relevant literature survey [5] [6] [7] [8]. These classes are the pandemic effect on the economy, panic buying due to lockdowns, views, and information about the vaccine and cure, COVID-19 Statistics, COVID-19 related news and information, politics, public health measures, governmental measures, fake treatment, and conspiracy theories. Next, we extracted a random sample from each of the three datasets and annotated them using the crowdsourcing platform LabelBox [3]. The sample size was determined using the the Central Limit Theorem [15] as follows:

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

where:

$N = 4, 214, 256$ English, $N = 1, 952, 784$ French, $N = 543, 558$ Arabic

$e = 1.32\%$ the margin of error: the percentage of deviation in result in the sample



Figure 3.3: The study area of the work: arab region countries per population

size compared with the total tweets.

$z = 1.96\%$ with confidence level 95%.

$p = 50\%$

Our sample size rounded up as follows: 5,600 tweets in English, 4,725 tweets in Arabic, and 5,496 tweets in French. Each tweet in each of the samples above was

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

Table 3.1: Annotators benchmark score

then annotated using three different people using the LabelBox platform. More specifically, each tweet was displayed to the three annotators independently along with the 11 identified classes, and the annotator was asked to assign one or more class to the tweet based on its content. To ensure high-quality annotations, gold-standard tweets that were annotated by us were interjected among the tweets

that were annotated on LabelBox, without their labels. We then measured the agreement of the annotators with the ground-truth labels provided by us on the gold-standard tweets. The annotators achieved an average of 97%, 95%, 94% accuracy with respect to the gold labels (Table 3.1).

Below, we describe each category in detail along with an example tweet that falls into this category illustrated in Table 3.2.

1. **Uninformative/Unrelated:** Includes any tweet that cannot be classified into any other categories, it may contain non-informative/personal information.
2. **Economics:** Includes any tweets that portrays the economic situation due to COVID-19. Some examples of economics topics:
 - Stocks
 - Bitcoins
 - Fuel and oil
 - Businesses and companies
 - Finances and investments
3. **Stocking Up:** A tweet that mentions or comments on items stockings and panic buying and/or its consequences.
4. **Vaccine/Cure:** Any tweet that contains information about the vaccine or possible cure development should be categorized as vaccine/cure. Example:
 - Vaccine development news
 - Cure – drug news
 - Vitamin C, Zinc

5. **COVID-19 Statistics:** Any tweet that contains statistics on either of the following:
 - First Case
 - New Cases
 - New Deaths

6. **COVID Information:** Any tweet related to COVID news that cannot be classified as first case or new cases and deaths. Example:
 - Studies about the virus
 - Formal news about the crisis
 - COVID transmission and spread news

7. **Politics:** A tweet is classified as politics/news if it mentions or comments on:
 - A political party/individuals (Trump, Biden ... etc.)
 - Political or governmental institutions (Congress ... etc.)
 - Political party actions
 - News article or a reference to a news website

8. **Public Health Measures:** Any tweet that contains safety measures to avoid getting infected or healthcare information. The measures include:
 - Washing hands
 - Social distancing
 - Wearing masks

- Quarantine Healthcare examples:
- Hospitals' status (need of ventilators, oxygen tanks, ...)
- Nurses and doctor's response to the crisis
- World Health Organization (WHO) announcements

9. **Governmental Measures:** Any tweet that contains the government actions/response to the COVID-19 spread should be labeled as governmental measures. Example:

- Government's announcement
- Calls to government actions
- Containment and closure policies (Curfew, school closure, workplace closure, public transport closure, Travel banning ...)
- Economic policies (income support, fiscal measures ...)
- Health system policies (testing policy, emergency investment in health care ...)

10. **Fake Treatment:** Tweets that contain a treatment and contains the below conditions:

- Not verified by World Health Organization (WHO) site
- Not verified by Center of Disease Control and Prevention (CDC) site

Examples of fake treatment:

- Anti-Malaria Drug Hydroxychloroquine
- Azithromycin
- Drinking olive oils

11. **Conspiracy Theory** Any tweet that endorses a conspiracy story should be classified as Conspiracy Theory. Example COVID-19 is:

- Bioweapon
- Resulted from electromagnetic fields and 5G
- Planned by Bill Gates
- Leaked from Wuhan Labs

The distribution of the annotated datasets topics is illustrated in Figure 3.4. As shown in the Figure, the Arabic tweets dataset only contains ten categories, where Stocking Up wasn't found in the sampled data.

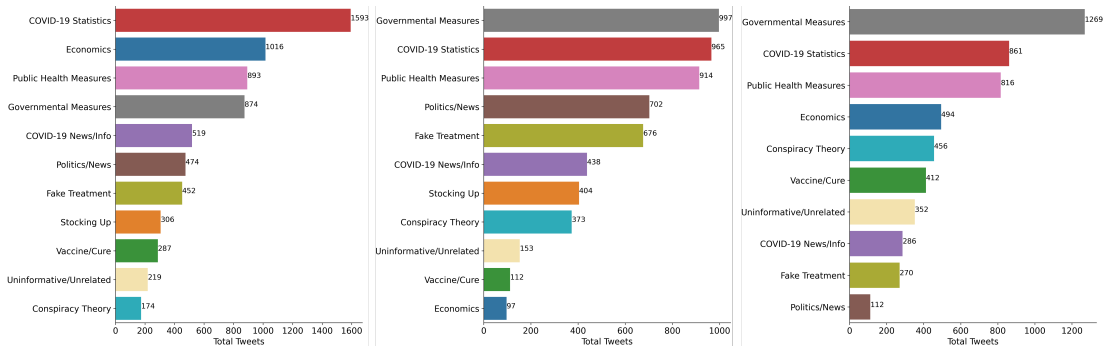


Figure 3.4: Annotated tweets topic distribution: English, French, and Arabic

3.2 Classifiers

Preprocessing the text before feeding it to the deep learning algorithms is essential and can enhance the accuracy of these models. We trained various classifiers to automatically annotate a COVID-19 tweet into one of our 11 predefined classes using the three annotated datasets described in the previous sections. Before training the classifiers, text tweets are preprocessed. The first step of the tweets

Topic	Tweet
Uninformative/Unrelated	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/IFCsc1RLHD 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/IDtXwbMFod #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 Information	Looks like its mutated already! #CoronaVirus is now #coronavirus!
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 @NAH-CONCEO @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

Table 3.2: Tweets topics examples

preprocessing stage is text cleaning, which includes removing URLs, mentions (@username), line breaks, extra white spaces, emojis, and unknown numerical symbols. This step is followed by normalizing text, which involves normalizing the repeated characters to handle the non-standard way of writing some words in the social media and removing diacritics (for Arabic text) and punctuation. After normalizing the tweets, we proceed with the text tokenization step by segmenting the text into tokens and eliminating stopwords. This process maximizes the number of words whose embeddings can be found in the pre-trained word embedding model. Since stop-words play a vital role the same as non-stop-words in BERT-based models [16], we decided to keep them for all the models that rely on contextual embeddings as they utilize in providing context information.

In the data preparation phase, we replace each word in the tweet with its corresponding word embedding from pre-trained distributed word representations. Recently, pre-trained embeddings have played a vital role in improving the accuracy of text classification models [17]. Word embeddings are used extensively in NLP to capture the semantic relations between a sentence’s words. Embedding techniques are classified into static word embeddings (such as word2vec, Glove, FastText) and contextual embeddings (BERT, ELMo). We used Glove [18] to represent English tweets and FastText [19] to illustrate English, Arabic, and French words. Both Glove and FastText are an extension of the Word2Vec [20] method. The main improvement of Glove is applying word-word co-occurrence probability to build the embedding. In contrast, the primary enhancement of FastText is forming a bag of character n-grams which allows the model to learn weights for words and each of its n-grams.

We also handle a multi-label classification problem. In Multi-label classification, one tweet can belong to multiple classes (labels) that we defined previously.

In other words, we aim to predict tweets into categories that are not mutually exclusive. Before delving into the models' architecture, we must first describe multi-label classification in deep learning models. Since the final score for each class in the output layer should be independent, we used the sigmoid function as an activation function. Each score of the last node is converted between 0 and 1 independent of the other classes scores. Therefore, we used 0.5 as the threshold to classify the tweets. If the score for some category is more than 0.5, the tweet is classified into that class. Also, the tweet can have more than one category with a score greater than 0.5. Since we used a sigmoid activation function, we used binary cross-entropy as a loss function in our models. As for the hidden layers, each layer is followed by a ReLU activation function. This function inputs a real-valued number and thresholds it at zero when less than 0. ReLU is faster to converge and easier to compute with better performance than other activation functions, such as Tanh and Sigmoid [21]. Finally, we used Adaptive Moment Estimation (Adam) as an optimization function.

3.2.1 Convolutional Neural Network

We started by building a CNN model to categorize the text of the tweet into one or more of the defined topics, including English, French, and Arabic tweets. CNN is well-known for its excellent performance in NLP tasks, precisely in the classification problems [22] and it's famed for its ability to extract essential features that contribute to the classification task.

Figure 3.5 shows the architecture of the CNN classifier we created. We used the model with different word embeddings and performed parameter tuning to find the best parameters with minimized binary cross-entropy on the validation data. We used three different variants of CNN, including CNN, CNN with Glove,

and CNN with FastText. Since Glove only supports the English language, we used it only for the English classification. Our proposed CNN model for each language consists of a word embedding layer, followed by the first convolution 1D layer, which accepts the words' vector embeddings resulting from the text representation phase. The input to this layer is $n \times 300$, where n is the number of words and 300 is the vector embedding dimension of each word. In addition, global maximum pooling 1D layer is used to down-sample the features of the convolution layer. Finally, we regularized the network by using dropout layers to overcome the overfitting problem. The final fully-connected Sigmoid classification layer outputs eleven units for English and French or ten units for Arabic, corresponding to the tweets' classification classes.

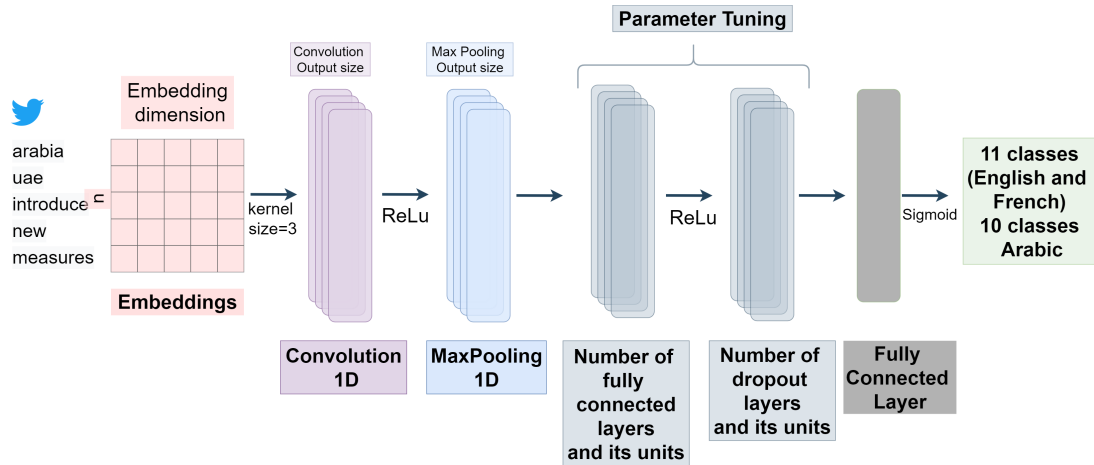


Figure 3.5: The architecture of the proposed CNN model for multi-label classification

3.2.2 Bidirectional Long Short-Term Memory

Recurrent neural networks (RNN) have shown promising solutions on different tasks, along with language models and speech recognition [23] [24]. It predicts the current output conditioned on long-distance features by keeping a memory based

on previous information. A Long Sort Term Memory (LSTM) is a variation of recurrent neural networks proposed by Hochreiter and Schmidhuder (1997) [25] as a solution to vanishing gradient. LSTM networks use purpose-built memory cells to update the hidden layer values. Each cell can be trained to determine which information from the sequence should be kept, transmitted to the output, or discarded. Therefore, they may function better at finding long-range dependencies in the data, unlike a standard RNN. The LSTM is made up of 3 gates (Figure 3.6):

- The forget gate f_t is responsible for discarding non-important information from the cell state.
- The input gate i_t is in charge of adding information to the cell state.
- The output gate o_t is responsible for opting for the valuable information displayed in the current cell state

The equations of the gates of the LSTM are:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (3.1)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (3.2)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (3.3)$$

where:

σ stands for the sigmoid function

W_x weights for the corresponding gates

x_t input at time t

h_{t-1} output of the previous LSTM cell

b_x biases for the relevant gates

The equations for the cell state and the final output are the below:

$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (3.4)$$

$$h_t = o_t \cdot \tanh(c_t) \quad (3.5)$$

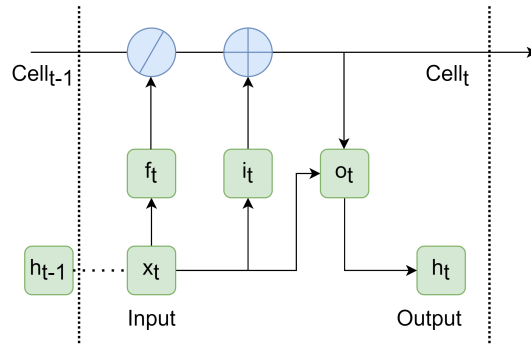


Figure 3.6: LSTM cell structure

To get more information from the sequence of the input text, we make two passes over the sequence, one from left to right (from x_i to x_t) and one from right to left (from x_t to x_i), and finally, we concatenate the forward LSTM \vec{f} with the backward LSTM \overleftarrow{f} .

We proposed a BiLSTM model for each language, as illustrated in Figure 3.7. The model consists of a word embedding layer, followed by the Bidirectional LSTM layer. We applied parameter tuning to identify the number of hidden and dropout layers. We employed three different variants of BiLSTM, including BiLSTM, BiLSTM with Glove, and BiLSTM with FastText.

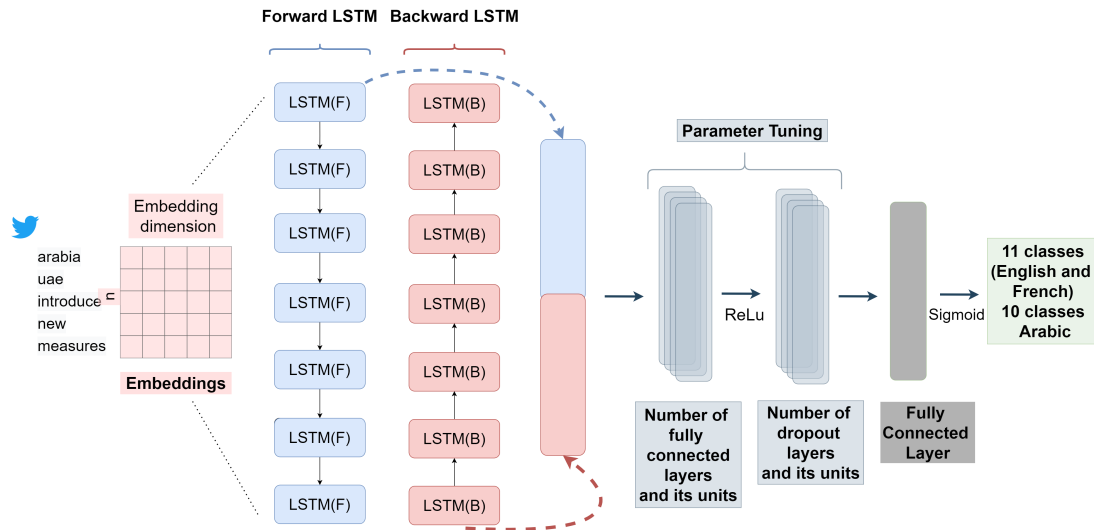


Figure 3.7: The architecture of the proposed BiLSTM model for multi-label classification

3.2.3 Bidirectional Encoder Representations

The final model we used to perform multilabel classification on COVID-19 tweets is BERT [26]. Pretrained embeddings methods can capture semantic and syntactic relationships of a language. However, they are unable to catch the contextual information. Meanwhile, a state-of-the-art technique named BERT, which stands for Bidirectional Encoder Representations, was invented by Devlin et al.[17] to capture semantic, syntactic, and contextual relationships. The BERT implementation includes two steps pretraining and fine-tuning. In the first step, the model is trained on an unlabeled dataset in a particular or multiple languages. In the second step, all the initialized parameters are fine-tuned using a labeled dataset. Figure 3.8 shows the architecture of the BERT model. BERT model uses a multilayer bidirectional transformer encoder. The transformers architecture is demonstrated in Figure 3.9. In particular, the encoder represented in the left part of the transformer’s architecture uses an input sequence of the word rep-

resentation and generates a 512-dimensional representation for each word. The decoder in the right part of the transformers contains one more layer than the encoder and masked multi-head attention that handles the output. The main advantage of BERT-based models is that word embedding is trained based on an autoencoder rather than a language model. Furthermore, a bidirectional transformer considers both the previous and next tokens when predicting the token, unlike the N-gram language model, which considers only the previous n words. Thus, the bidirectional transformers can combine contextual information from both directions simultaneously. We used the BERT_{base} uncased model [26] com-

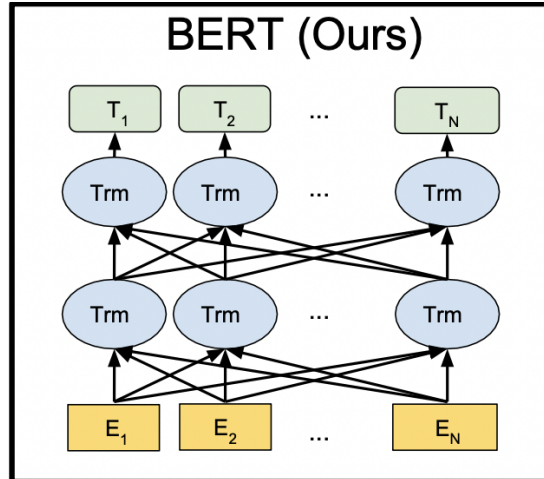


Figure 3.8: The Architecture of BERT Model

posed of 12 attention layers, 12 attention heads, 758 hidden layers dimensions, and a 100 maximum sequence length for the English dataset. For the Arabic and French datasets, we employed Arabic-BERT [27] and CamemBert [28] transformers. The two transformers are inspired by Google’s BERT architecture and are composed of the same architecture as BERT_{base}. Table 3.3 shows the data sources of BERT_{base} uncased, Arabic-BERT, and Camembert. We fine-tuned these three transformers on our labeled datasets related to the COVID-19 con-

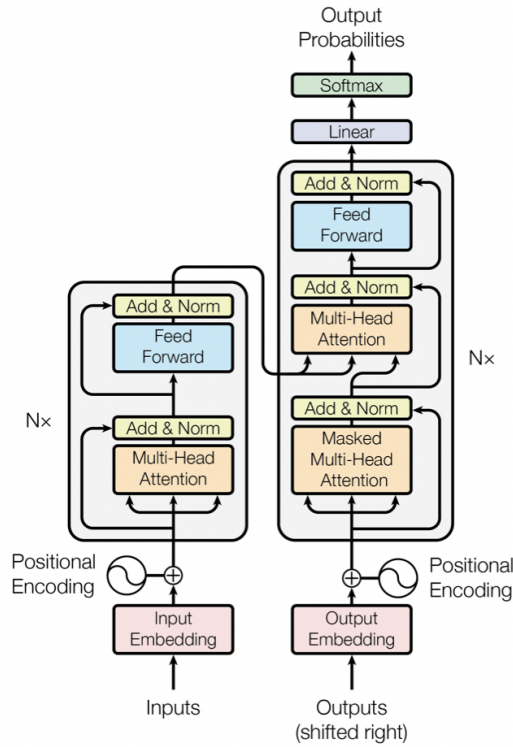


Figure 3.9: The Architecture of Transformers

text. The scheme of the model is illustrated in Figure 3.10. First, the model tokenizes the input tweets to split the word into tokens compatible with BERT-based models. For the $BERT_{base}$ uncased model, we used BERT Tokenizer, and for Arabic-BERT and CamemBert, we used WordPiece tokenizer [29]. Then, we added the special tokens that are composed of the following:

- The [CLS] token marks the start of the sentence, and it is added to the beginning of each text
- The [SEP] token marks the end of the sentence and is appended at the end of each sentence
- The [PAD] token is added to maintain a uniform text length across the entire training dataset

The generated tokens, including special tokens, are converted into IDs, CLS token with ID 101, SEP with ID 102, and PAD with ID 0. Then the formed IDs are fed into the BERT-based model to produce representations of the words in the texts via the multiple transformer layers. Finally, the first head of the final layer, which corresponds to the embedding of the [CLS] token, is fed into the classifier. The classification model is a fully-connected layer passed into a sigmoid function to get the probability distribution over the predicted output classes. We used adaptive moment estimation (AdamW) [30] for the optimization with a learning rate of $2e-5$, and we employed binary Cross-Entropy loss function for the multilabel classification.

Model	Data Source
Bertbased	BookCorpus (11,038 unpublished books) and English Wikipedia Arabic version of OSCAR
Arabic-BERT	Arabic Wikipedia Other Arabic resources (~ 95 GB of text)
CamemBert	Not restricted to MSA, they contain some dialectical Arabic too OSCAR (138 GB of text)

Table 3.3: Comparison between BERT transformers

3.2.4 Experiments and Results

In this study, we considered measuring metrics that deal with multi-label classification problems as proposed in [31]. We evaluated the proposed models using five multi-label performance measures. These metrics are accuracy, Jaccard accuracy, Micro-averaged F1 score, Label ranking average precision score, and Hamming loss.

- **Accuracy** Accuracy measurement calculates the predicted true labels among

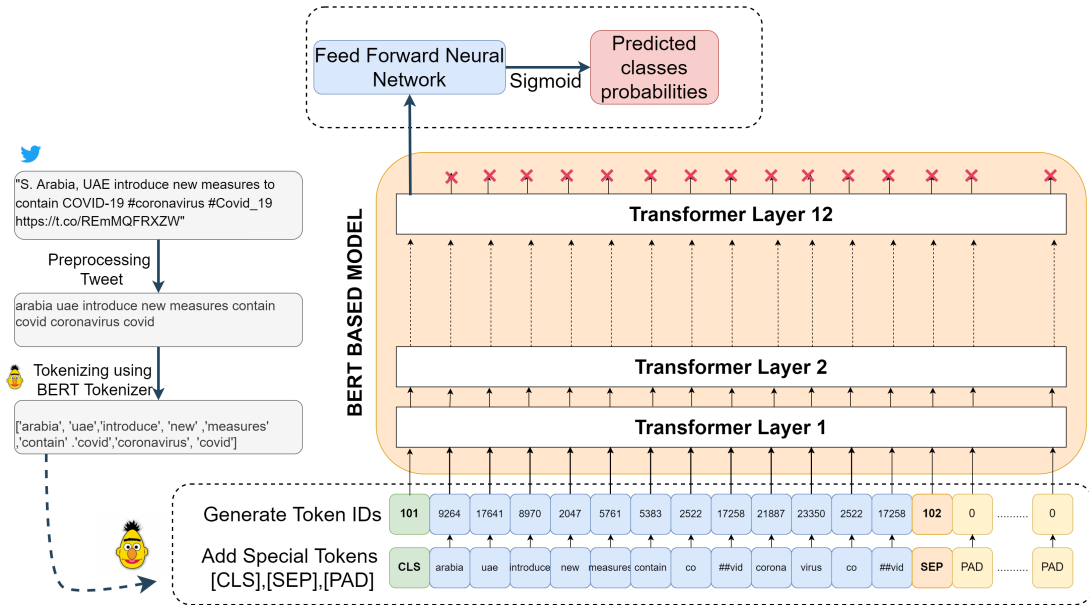


Figure 3.10: The architecture of the proposed BERT model for multi-label classification

the whole predicted labels.

$$\frac{1}{N} \sum_{i=1}^n \sigma(\hat{y}_i == y_i)$$

where $\sigma(\hat{y}_i == y_i)$ equals to 1 when the predicted label \hat{y}_i is equal to the true label y_i , and equals 0 otherwise.

- **Jaccard Accuracy**

$$Jaccard \ Accuracy = \frac{1}{|D|} \sum_{i=1}^D \frac{Y_i \cap \hat{Y}_i}{Y_i \cup \hat{Y}_i}$$

- **Micro-averaged F1 score** F1 score averaged on global calculation.
- **Label ranking average precision score** Average across each ground truth label assigned to each sample. This metric aims to better rank the labels associated with each sample and then measure whether the percent-

age of the higher-ranked labels were true labels. The obtained score is greater than 0, and the optimal value is 1.

- **Hamming loss** It is the fraction of the wrong labels to the total number of labels. Hamming loss evaluation metrics measure the proportion of the wrong labels to the total number of labels. That is the fraction at which the labels are misclassified.

The optimal value of the hamming loss evaluation is zero.

We conducted an extensive experiment to implement the models. The experiments involve training the classifiers CNN and BiLSTM with the text embeddings (Glove[32] and FastText[33]) or without word embeddings and fine-tuning the parameters. We evaluate the performance of the models based on the evaluation metrics described previously. The experiments are done using TensorFlow [34] and Keras Tuner [35]. Table 3.4 shows the final best parameters experimental settings of each model conducted for the classification tasks.

	Models	# Hidden Layers	#units	# Dropout Layers	Dropout rate	# LSTM units	Learning rate
English	CNN	None	None	3	0.2	None	0.01
	CNN with Glove	1	64	None	None	None	0.001
	CNN with FastText	1	64	2	0.34	None	0.01
	BiLSTM	None	None	1	0.42	192	0.004
	BiLSTM with Glove	1	64	1	0.7	128	0.0027
	BiLSTM with FastText	None	None	3	0.4	256	4.5e-05
Arabic	CNN	None	None	1	0.1	None	0.001
	CNN with FastText	1	64	2	0.3	None	0.001
	BiLSTM	None	None	2	0.7	224	0.0005
	BiLSTM with FastText	None	None	1	0.77	64	0.0005
French	CNN	None	None	1	0.46	None	0.01
	CNN with FastText	1	32	2	0.41	None	0.01
	BiLSTM	2	32	1	0.67	192	0.007
	BiLSTM with FastText	1	16	2	0.72	64	0.004

Table 3.4: Deep learning models best parameters experimental settings

As shown in Table 3.5, the deep learning models performed poorly compared with transformer-based models, BERT_{base} uncased, Arabic-BERT, and Camembert models with F1 scores 0.84, 0.81, 0.87, respectively. The out performance

of BERT models can be attributed to the contextual representation of the input tweets learned from the enormous texts' corpora used to train these models. Surprisingly, when the deep learning models with the different embeddings were used, the CNN model with an F1 score of 0.78 without any pre-trained embeddings performed better than CNN with Glove or FastText with an F1 score of 0.77 and 0.76 respectively for English text. Contrary, the BiLSTM model has shown a better F1 score (0.81) than CNN models with the Glove embedding. On the other hand, as for the Arabic classification, both CNN and BiLSTM have performed poorly without word embeddings with 0.74 and 0.56 F1 scores. However, the FastText embedding has boosted CNN's F1 score by 0.01 and BiLSTM's by 0.14. Significantly, the Arabic-BERT has overly performed the deep learning models with an F1 score of 0.81 compared to the best deep learning CNN model with Fast text embedding. This improvement can be attributed to the fact that Arabic-BERT utilizes massive amounts of corpus and vocabulary set (8.2 Billion words) that are consisted of both Modern Standard Arabic and dialectical Arabic. Therefore, enabling Arabic-BERT to capture Dialectic Arabic and MSA. Finally, the French text classification deep learning models have shown promising results without embedding a 0.82 F1 score for both CNN and BiLSTM. On the other hand, the Fast text embedding improved the CNN model's performance by 0.02, reducing the BiLSTM model's performance by 0.02.

	Models	Acc	J-Acc	F1-Micro	LRAP	H-Loss
English	CNN	0.78	0.72	0.78	0.88	0.04
	CNN with Glove	0.76	0.73	0.77	0.89	0.05
	CNN with FastText	0.77	0.68	0.76	0.89	0.05
	BiLSTM	0.73	0.67	0.73	0.85	0.05
	BiLSTM with Glove	0.81	0.78	0.81	0.91	0.04
	BiLSTM with FastText	0.79	0.72	0.78	0.90	0.04
	BERTbase	0.80	0.81	0.84	0.92	0.04
Arabic	CNN	0.74	0.67	0.74	0.86	0.05
	CNN with FastText	0.75	0.67	0.75	0.87	0.05
	BiLSTM	0.52	0.47	0.56	0.71	0.07
	BiLSTM with FastText	0.70	0.64	0.70	0.83	0.06
	Arabic-BERT	0.78	0.79	0.81	0.89	0.04
French	CNN	0.82	0.75	0.82	0.92	0.03
	CNN with FastText	0.81	0.78	0.84	0.93	0.03
	BiLSTM	0.81	0.76	0.82	0.92	0.03
	BiLSTM with FastText	0.78	0.75	0.80	0.90	0.04
	CamemBert	0.85	0.85	0.87	0.93	0.03

Table 3.5: Performance of the proposed models

CHAPTER 4

ANALYSIS

In this chapter, we used the TBCOV dataset[4], an extension of the GeoCoV19 dataset we sampled and annotated to train our models. TBCOV, the most prominent Twitter dataset related to COVID-19, is a large-scale Twitter dataset containing two billion multilingual tweets spanning from February 1, 2020, until March 31, 2021. It covers various topics, including social, health, and economic concerns caused by the COVID-19 pandemic. It also covers opinions and perspectives about the government’s measurements and decisions, food shortage, and other topics. Imran et al. (2021) used the XLM-T model, a transformer-based model, to obtain sentiment labels and confidence scores for the TBCOV tweets dataset. Besides, the tweets are also enriched with essential fields, including geolocation information, and gender. Its comprehensive topic coverage and geolocated data prompted us to use it as a testing dataset for our best-performed model. We extracted the Arab region geolocated data from TBCOV and filtered out null values from the gender field and English, Arabic, and French tweets. We have used the Twitter Streaming API to extract the TBCOV tweets’ ids full text and ended up with 10,635,996 geolocated tweets.

4.1 Tweets Statistics and Distribution

Figure 4.1 shows the distribution of the tweets per language. The English language dominates with around 5.7 million tweets, followed by the Arabic language, which covers about 4.7 million tweets, and French tweets with a minor percentage covering about 250,000 tweets. Figure 4.2 shows the distribution of the tweets per country where posts are mainly disseminated from Saudi Arabia (2435206), the United Arab Emirates (2128522), Egypt(900814), Kuwait (869571), Jordan(584391), and Lebanon (457466). For more meaningful comparisons of geotagged tweets across Arab countries, we normalized tweets for each country by population and calculated the number of posts per 100,000 persons

$$\text{Normalized tweets} = \frac{\text{Total number of tweets per country}}{\text{Country's population}} \times 100,000 \quad (4.1)$$

As a result, Figure 4.3 shows the normalized counts of geotagged tweets for each Arab region country with a symbology ranging between 0 and 20k (the dark purple representing the highest value and light yellow representing the lowest value). The results show that UAE had around 21k tweets per 100,000 people, followed by Kuwait, Bahrain, Qatar, Saudi Arabia, Palestine, and Lebanon, which have 20k, 14k, 12k, 7k, 7k, 6.7k tweet per 100,000 persons, respectively. Readers are referred to Table B.2 for more details.

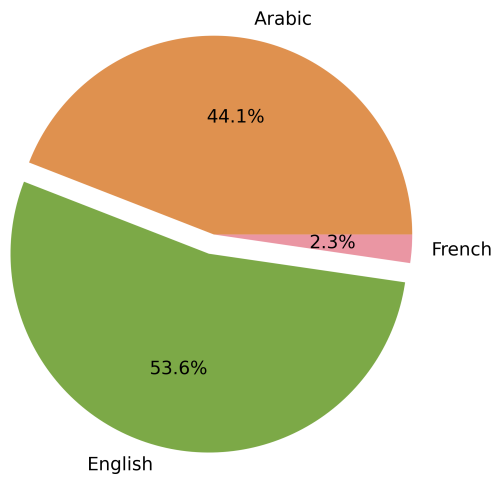


Figure 4.1: Tweets Distribution Per Language

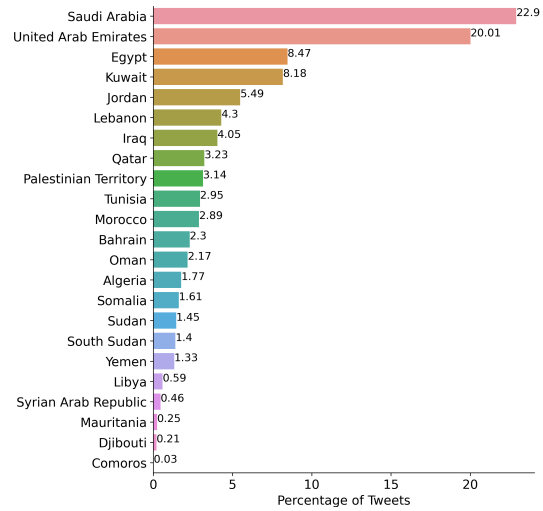


Figure 4.2: Tweets Distribution Per Country

4.2 Predicted Topics Using BERT-based Model Statistics

4.2.1 Topics Distribution

The predicted topics by BERT distribution are depicted in the Figure 4.13. The model showed a dominant prediction for non-informative tweets with a 32% weight and a minimal prediction for conspiracy theory stocking up and fake treatment with 1-2% weight. The high percentage of non-informative tweets indicates that the Arab region population was tweeting about personal topics that were not informative about the pandemic. At the same time, the dataset contains a good representation of other topics, mainly governmental measures, public health measures, COVID information, vaccine, economic, and politics, with a percentage weight ranging between (5-16%). A different representation of the topics distribution is presented in the Figure 4.14 . This Figure gives insights to the topics



Figure 4.3: Geotagged tweets in the Arab Region normalized by country’s population (per 100,000 persons)

dispersal per population in 100,000 person. The highest countries contributing in disseminating in all the identified topics were United Arab Emirates, Qatar, Kuwait, Palestine, Lebanon, Oman, Saudi Arabia, and Jordan. We provide a detailed statistics in Table B.3. We further analyzed the topic distribution per gender. The males, which represents a high percentage in the dataset(68.5%) as illustrated in Figure 4.5, majorly tweeted about all the topics with 73% of conspiracy theory, followed by governmental measures, COVID-19 statistics, and economics with a percentage ranging between 70-72%. In contrast, the females were more oriented in tweeting about politics, stocking up, and fake treatment, with a rate ranging between 33-38% (Figure 4.4).

4.2.2 Topics Trend Analysis

Figure 4.6 depicts the monthly distribution of the predicted topics debated over the tweets. A similar pattern can be noticed in topics variation over time, where

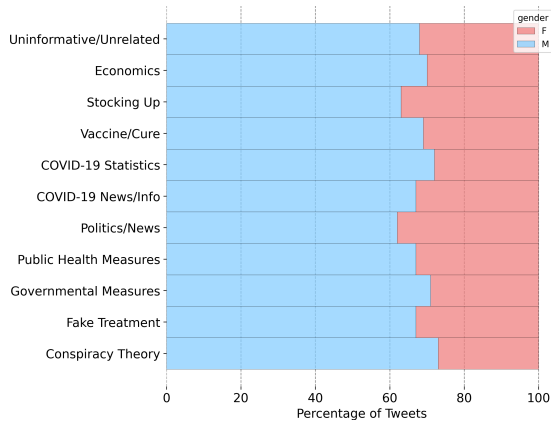


Figure 4.4: Distribution of tweets' sentiment per gender

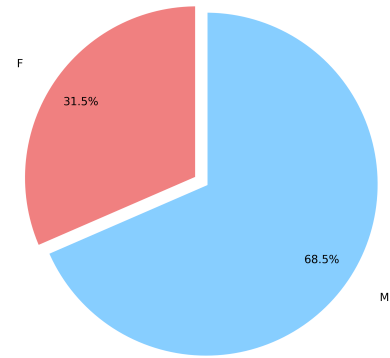


Figure 4.5: Gender Distribution

all topics started to rise during the pandemic, extending from Feb 2020 and peaking in April 2020. Then, the trend began to fall from April 2020 to Oct 2020, where the topics spiked again except for stocking up. Finally, a sudden rise in the vaccine topic can be noticed after Oct 2020, where other topics kept the same pattern.

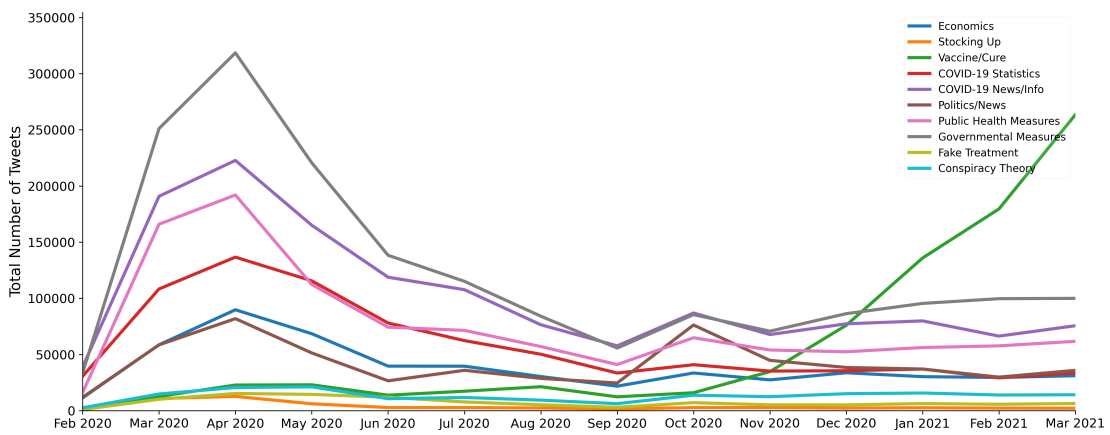


Figure 4.6: Monthly trends of the identified topics

4.3 Sentiment Analysis

The sentiments segregate in the Arab region between positive, negative, and neutral, with negative ones representing the highest proportion (37.8%). Males have dominantly contributed to these sentiments dissemination(Figure 4.9). Fig-

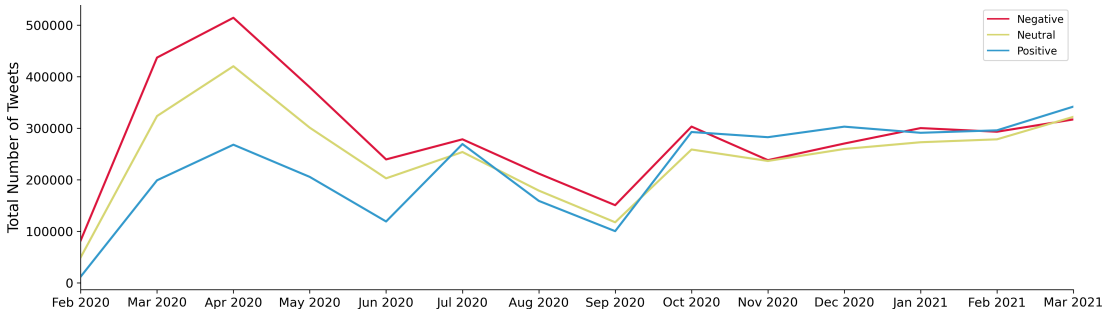


Figure 4.7: Monthly trends of the sentiments

Figure 4.7 presents a monthly aggregation of sentiment labels for all tweets in the three languages. As expected, the negative sentiment dominates until mid Sep 2020. A significant rise of negative sentiment is noticeable at the beginning of March, peaking in April and then averaging down during the later months, where the negative sentiment average is 4 million. The negative sentiment spikes with two peaks after April 2020. However, none reaches as high as tweets surged in this month. The neutral sentiment in the Arab region stays between the negative and positive sentiment trends. It follows a similar pattern until mid-Sep 2020, where it starts to be lower than both trends. Whereas, the positive sentiment remained lower than negative and neutral ones until the mid of Sep 2020. After that, the positive sentiment dominates, coinciding with the vaccine topic’s discussion period, as illustrated in the Figure. Figure 4.8 presents the distributions of sentiment labels for the three languages. Interestingly, the Arabic language shows the dominance of the positive sentiment throughout the 14 months except

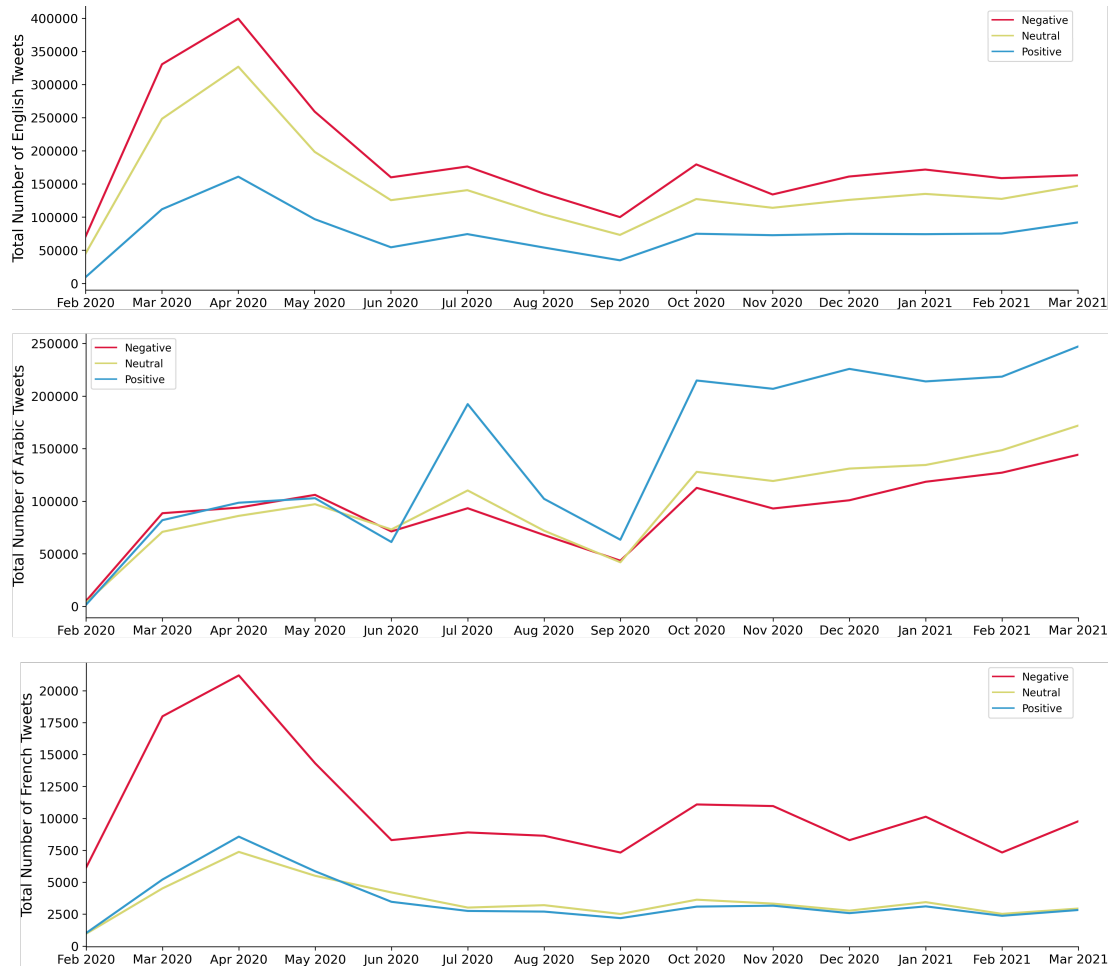


Figure 4.8: Monthly trend of sentiments per language

June 2020 and a few weeks in the middle. On the other hand, the negative sentiment surpasses the other two sentiment classes for English and French languages, showing peaks in April and May 2020. The sentiment analysis by country (Figure 4.12) shows an overwhelming negative sentiment distribution circulated in the tweets across all Arab regions except a few countries. Surprisingly, Saudi Arabia (41.7%) and other countries, including Kuwait (36.5%), Bahrain (36.5%), and Jordan (35.6%), had higher positive sentiment than negative. The rest of the Arab region, including Mauritania, Morocco, South Sudan, and Palestine, show moderate to strong negative sentiment.

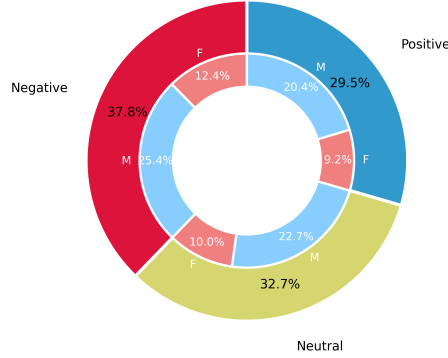


Figure 4.9: Distribution of tweets' sentiment per gender

The weighted average sentiments score across countries illustrated in the map (Figure 4.10) emphasize the findings above. Equation below shows the computation of the weighted average score S_c :

$$S_c = \frac{\sum_{t_i^c \in \{pos, neut\}} \phi_i^c + \sum_{t_i^c \in \{neg\}} \phi_i^c}{N_c} \quad (4.2)$$

where t_i^c represents the sentiment label of tweet i from country c , ϕ_i^c the model's confidence score for t_i^c , and N_c the total number of tweets per country c .

Saudi Arabia, Kuwait, Jordan, and Bahrain normalized sentiments show a significant positive sentiments with weighted average score 0.31, 0.24, 0.22, 0.22, respectively. Readers are referred to Table B.1 for more details.

The sentiments distribution per topic (Figure 4.11) shows an overwhelming negative sentiment in Politics/News (71.65%), Stocking Up (54.9%), Government Measures (52.3%), and COVID News/Info (50.52%) whereas other topics illustrate a neutral and positive sentiments domination.

We further analyzed the topics distributed by sentiment across the countries (Figure 4.15), the sentiments can be grouped into (1) Topics having **neutral, slightly to highly negative sentiments** include economics, stocking up, pol-



Figure 4.10: Arab Region sentiment based on normalized scores of the labeled sentiment in each country

itics, and governmental measures where the highest overall negative sentiment is observed in the political discussions, (2) topics ranging from **slightly positive, neutral to slightly negative sentiments** are observed in public health measures, fake treatment, COVID-19 statistics, COVID-19 news, and conspiracy theory, excluding Mauritania, South Sudan, and Somalia, with around -0.4 z-score only in conspiracy theory topic, (3) the vaccine topic’s sentiment varied between **highly positive and neutral**. For example, Saudi Arabia shows high positive sentiments about the vaccine (0.42), whereas Palestine (0.03), South Sudan (0.05), and Iraq (0.08) tend to discuss the topic neutrally. More details are provided in Table B.4.

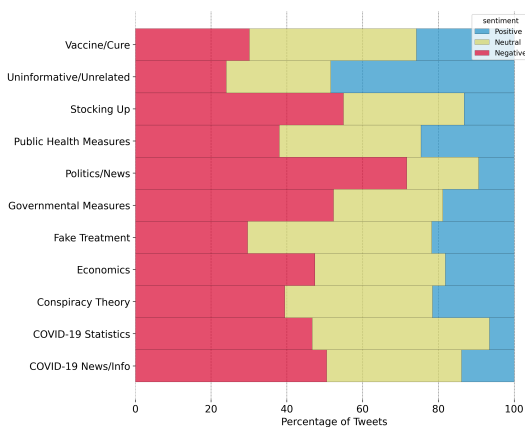


Figure 4.11: Topic Sentiments Distribution

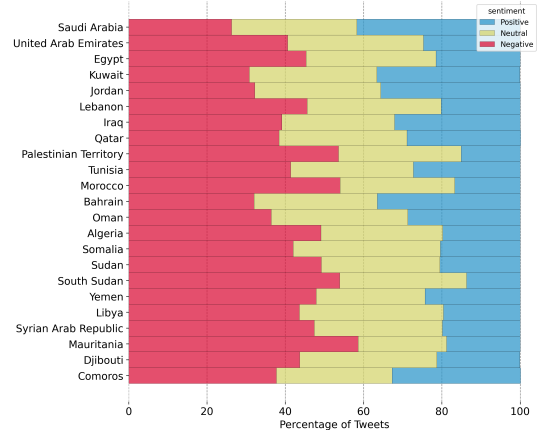


Figure 4.12: Sentiments distribution per country

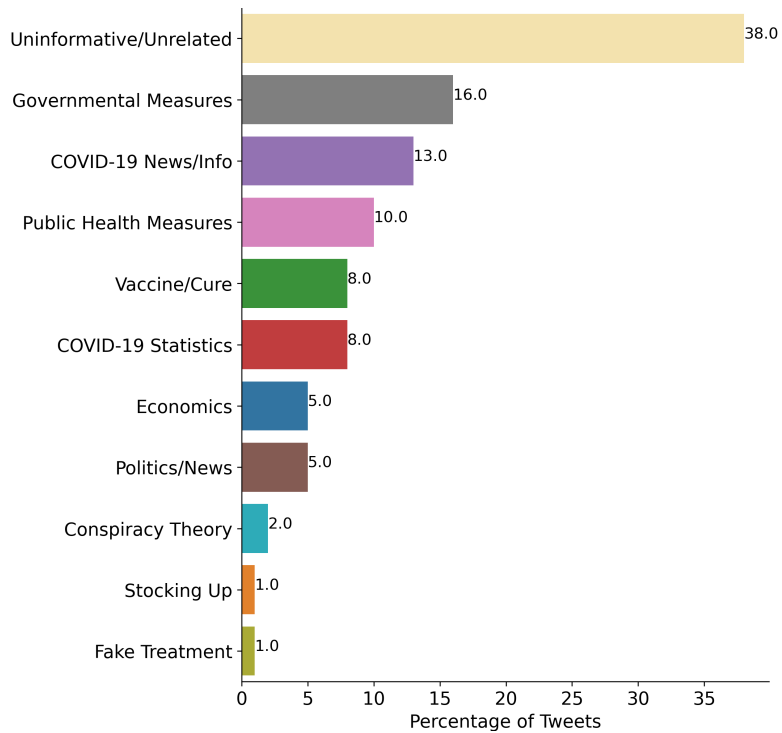


Figure 4.13: Topic Distribution

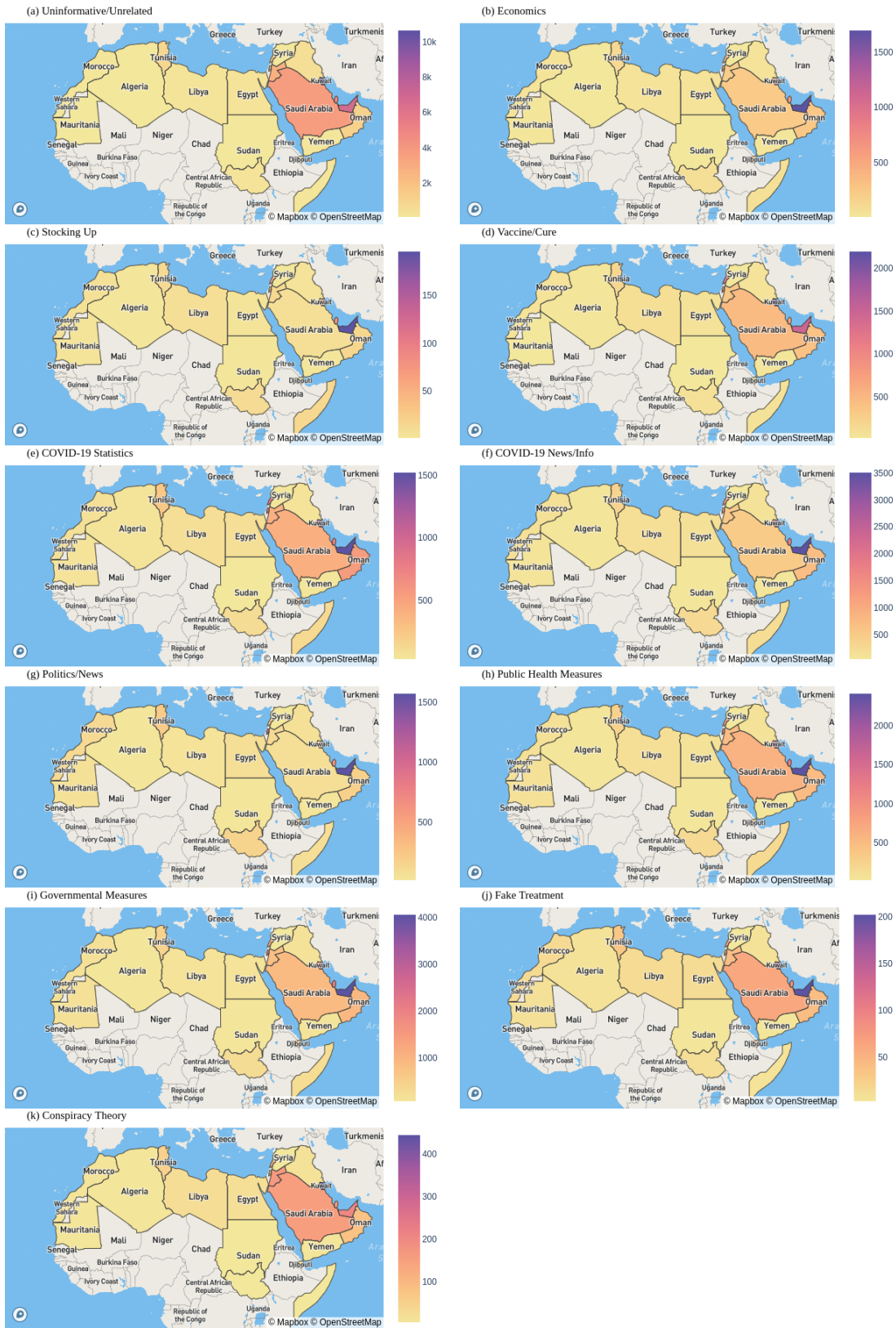


Figure 4.14: Geotagged Topics in the Arab Region normalized by country's population (per 100,000 persons)

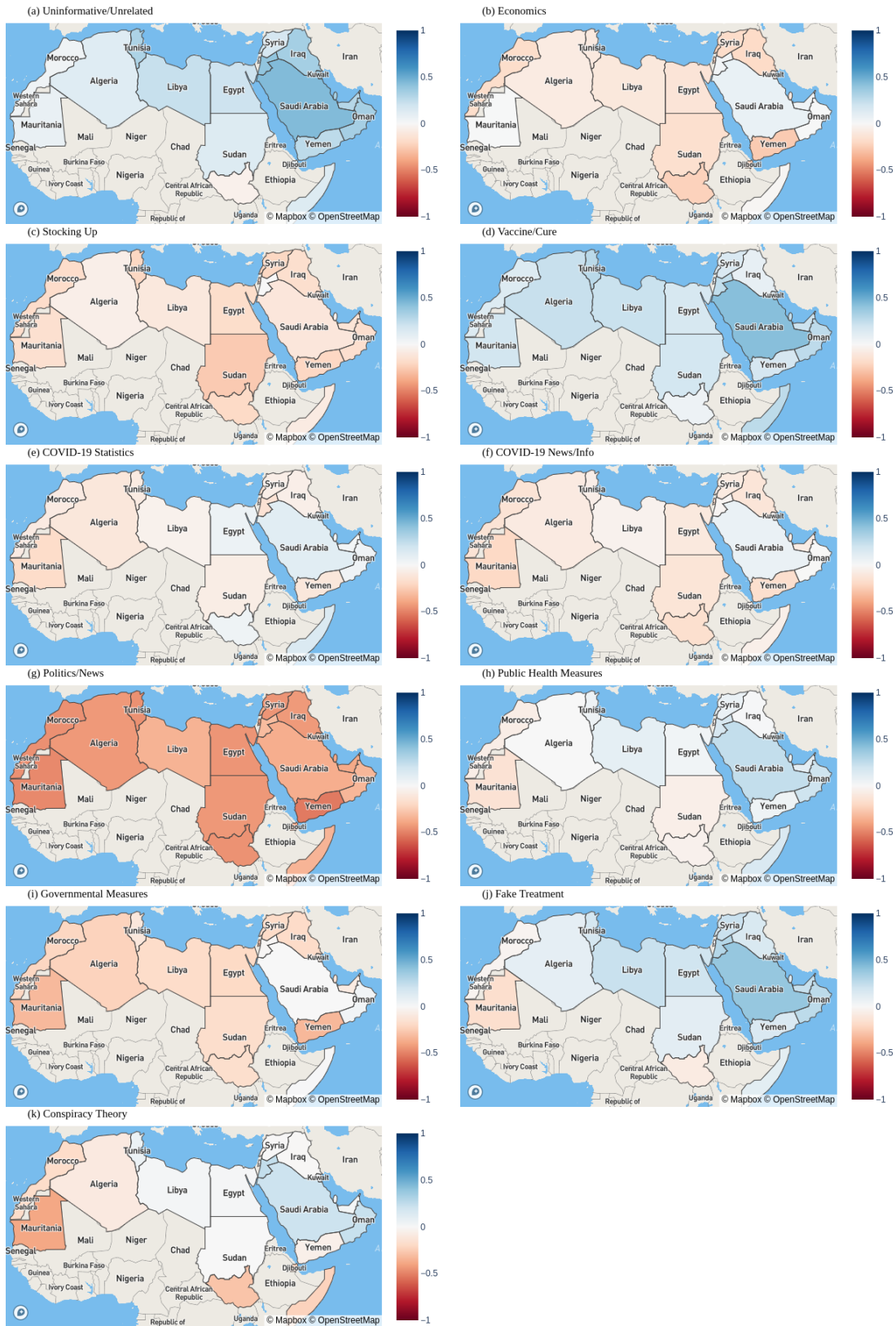


Figure 4.15: Topics sentiment in the Arab Region based on normalized scores of the labeled sentiment in each country 50

CHAPTER 5

CONCLUSION

This thesis is the first to perform a multi-label topic classification for Arabic, English, and French tweets, especially at a finer and large-scale level across the Arab region. In addition, this study investigated to what extent deep learning models can assist in understanding society’s concerns and behavior during the COVID-19 pandemic. We proposed a multi-label topic classifiers by employing deep learning models CNN and BiLSTM with different pre-trained word vector representations and three BERT-based transformers, Bert_{base} uncased, Arabic-BERT, and CamemBert. The BERT-based models outperformed the deep learning models in classifying English, multi-dialect Arabic, and French tweets with F1-Micro scores of 0.84, 0.81, and 0.87, respectively. The topics include 11 classes: Economics, Stocking Up, Vaccine, COVID-19 Statistics, COVID-19 Information, Politics, Public Health Measures, Governmental Measures, Fake Treatment, and Conspiracy Theory and Non-informative.

The analysis shows how public discussions and sentiments evolved for 14 months (between February 1, 2020, and March 31, 2021). The topics followed a similar pattern of the rapid rise and slow decline across the region with a sud-

den rise of the vaccine topic after Sep 2020. Negative sentiments are observed in economics, stocking up, politics, and governmental measures topics, slightly negative sentiments are observed in public health measures, COVID-19 statistics, COVID-19 info, fake treatment, and conspiracy theory. In contrast, the vaccine topic's sentiment varied between high positive and neutral. These findings help us understand how Twitter users described their concerns about the pandemic.

As future work, we will consider using different deep learning architectures mainly Graph Convolutional Networks (GCN) and building more complex ensemble models to investigate the accuracy using other models and utilize other contextual models such as ROBERTA, ALBERT, and ELMo.

APPENDIX A

ABBREVIATIONS

COVID-19	Coronavirus disease
WHO	World Health Organization
UAE	United Arab Emirates
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
BERT	Bidirectional Encoder Representations
SVM	Support Vector Machine
ELMo	Embeddings from Language Model
TF-IDF	Term Frequency–Inverse Document Frequency
Glove	Global Vectors for Word Representation
LDA	Latent Dirichlet Allocation
API	Application Programming Interface

APPENDIX B

LIST OF TABLES

Country	Sentiment Score
Saudi Arabia	0.31
Kuwait	0.24
Jordan	0.22
Bahrain	0.22
Oman	0.14
Comoros	0.12
Qatar	0.12
Iraq	0.11
United Arab Emirates	0.08
Tunisia	0.07
Somalia	0.05
Libya	0.03
Djibouti	0.03
Egypt	0.01
Lebanon	0.00
Syrian Arab Republic	-0.03
Yemen	-0.04
Algeria	-0.05
Sudan	-0.06
Palestinian Territory	-0.12
South Sudan	-0.13
Morocco	-0.13
Mauritania	-0.21

Table B.1: List of geotagged weighted average score sentiments in the Arab Region

Country	Population	Total number of tweets	Normalized tweets
United Arab Emirates	9890400	2128522	21521
Kuwait	4270563	869571	20362
Bahrain	1701583	244298	14357
Qatar	2881060	343050	11907
Saudi Arabia	34813867	2435206	6995
Palestinian Territory	4803269	333672	6947
Lebanon	6825442	457466	6702
Jordan	10203140	584391	5728
Oman	5106622	230354	4511
Tunisia	11818618	313998	2657
Djibouti	988002	22096	2236
South Sudan	11395758	148694	1305
Somalia	15893219	170950	1076
Iraq	40222503	430246	1070
Libya	6871287	62686	912
Egypt	1.02E+08	900814	880
Morocco	36910558	307394	833
Mauritania	4649660	26650	573
Yemen	29825968	141302	474
Algeria	43851043	187762	428
Sudan	43849269	153910	351
Comoros	869595	2820	324
Syrian Arab Republic	17500657	49278	282

Table B.2: List of geotagged tweets in the Arab Region normalized by country's population (per 100,000 persons)

Country	Uninfo*	Econ*	SU*	Vacc*	Stat*	Info*	Pol*	PHM*	GM*	FT*	CT*
Algeria	19.79	3.9	0.66	7.2	12.51	20.3	6.28	12.62	21.03	1.88	1.66
Bahrain	55.07	4.13	0.32	4.72	5.13	10.37	2.71	6.96	13.82	0.78	1.5
Comoros	8.3	4.04	0.35	4.33	8.44	22.98	2.98	16.81	37.8	1.99	0.5
Djibouti	11.51	6.92	0.49	4.63	13.69	21.47	6.63	12.9	32.61	1.17	0.4
Egypt	27.95	5.08	0.43	8.31	9.99	18.15	9.89	9.67	16.31	1.42	2.22
Iraq	48.93	4.08	0.44	5.45	5.43	12.4	5.9	7.25	15.36	0.76	1.39
Jordan	48.43	3.81	0.26	6.22	7.27	9.82	2.15	8.77	14.29	0.76	2.89
Kuwait	52.4	3.09	0.33	10.79	6.2	9.39	1.13	7.58	9.4	0.78	2.18
Lebanon	18.96	5.93	0.82	16.68	11.66	18.12	3.43	12.68	16.78	0.95	1.84
Libya	21.2	5.23	0.89	11.58	10.36	21.27	5.3	12.38	15.87	1.5	2.67
Mauritania	10.99	3.41	0.78	3.34	9.82	18.85	10.72	16.26	31.44	1.77	1.54
Morocco	14.92	5.08	0.87	7.76	9.88	20.33	15.01	12.41	21.15	1.75	1.19
Oman	25.3	5.96	0.35	8.82	12.43	16.32	4.12	11.36	20.96	0.95	1.77
Palestinian Territory	18.05	6.33	0.41	11.86	9.92	10.87	15.37	13.41	27.69	0.57	0.68
Qatar	39.41	5.2	0.64	5.97	6.28	14.86	5.22	10.42	17.69	0.71	1.32
Saudi Arabia	53.8	3.49	0.11	6.73	6.1	7.28	1.22	8.84	12.79	0.91	2.22
Somalia	17.26	6.78	1.38	6.21	11.42	20.39	7.5	16.52	23.61	0.61	0.46
South Sudan	21.53	6	0.89	6.32	8.4	21.43	12.45	12.93	21.31	0.92	0.3
Sudan	23.33	4.87	0.59	6.97	8.85	19.73	9.14	12.81	20.97	1.37	1.29
Syrian Arab Republic	22.93	4.94	0.41	8.48	13.33	18.81	6.97	9.15	22.69	1.14	1.88
Tunisia	32.91	4.64	0.51	7.54	9.36	14.65	8.44	9.86	16.74	1.25	1.87
United Arab Emirates	31.31	7.87	0.91	6.99	7.07	16.29	7.31	11.19	18.89	0.94	0.96
Yemen	30.55	5.03	0.46	5.23	8.79	17.24	5.95	9.11	24.84	1	1.91

Table B.3: List of normalized tweets per topic by country’s population (per 100,000 persons)

Uninfo*: Uninformative/Unrelated

Econ*: Economics

SU*: Stocking Up

Vacc*:Vaccine/Cure

Stat*: COVID-19 Statistics

Info*: COVID-19 News/Info

Pol*: Politics

PHM*: Public Health Measures

GM*: Governmental Measures

FT*: Fake Treatment

CT*: Conspiracy Theory

Country	Uninfo*	Econ*	SU*	Vacc*	Stat*	Info*	Pol*	PHM*	GM*	FT*	CT*
Algeria	0.13	-0.09	-0.06	0.23	-0.11	-0.09	-0.44	0.00	-0.23	0.10	-0.10
Bahrain	0.37	-0.01	-0.04	0.24	0.02	-0.06	-0.23	0.12	-0.07	0.28	0.17
Comoros	0.30	-0.05	0.11	0.23	-0.12	0.01	-0.10	0.20	0.15	0.19	0.20
Djibouti	0.08	-0.26	-0.10	0.14	0.15	-0.05	-0.40	0.01	0.06	0.08	-0.25
Egypt	0.21	-0.13	-0.18	0.17	0.06	-0.10	-0.45	0.04	-0.20	0.20	0.02
Iraq	0.34	-0.19	-0.17	0.08	-0.06	-0.13	-0.43	0.01	-0.19	0.16	-0.01
Jordan	0.40	0.03	0.00	0.25	-0.15	-0.02	-0.32	0.16	-0.01	0.29	0.24
Kuwait	0.41	-0.08	-0.11	0.15	0.19	-0.03	-0.31	0.10	-0.13	0.31	0.04
Lebanon	0.19	-0.22	-0.22	0.14	-0.15	-0.06	-0.36	0.08	-0.16	0.16	-0.06
Libya	0.23	-0.11	-0.11	0.22	-0.05	-0.04	-0.35	0.09	-0.19	0.23	0.02
Mauritania	0.07	0.00	-0.16	0.19	-0.19	-0.20	-0.48	-0.15	-0.33	-0.17	-0.39
Morocco	0.04	-0.18	-0.18	0.14	-0.09	-0.14	-0.46	-0.07	-0.21	-0.03	-0.19
Oman	0.34	0.01	-0.17	0.26	0.04	0.00	-0.33	0.21	0.03	0.26	0.23
Palestinian Territory	0.14	-0.21	-0.17	0.03	-0.08	-0.15	-0.40	-0.07	-0.29	0.01	-0.01
Qatar	0.34	-0.10	-0.04	0.25	0.03	-0.07	-0.39	0.09	-0.12	0.16	0.06
Saudi Arabia	0.44	0.07	-0.12	0.42	0.12	0.07	-0.37	0.24	0.00	0.39	0.21
Somalia	0.14	-0.02	-0.09	0.26	0.15	-0.06	-0.34	0.15	0.01	0.10	-0.23
South Sudan	-0.05	-0.23	-0.20	0.05	0.05	-0.17	-0.46	-0.05	-0.17	-0.09	-0.29
Sudan	0.14	-0.17	-0.26	0.18	-0.07	-0.13	-0.44	-0.06	-0.18	0.11	0.00
Syrian Arab Republic	0.18	-0.19	-0.21	0.15	-0.06	-0.09	-0.45	0.07	-0.20	0.22	0.00
Tunisia	0.31	-0.10	-0.21	0.27	-0.09	-0.07	-0.46	0.05	-0.11	0.18	0.06
United Arab Emirates	0.29	0.04	-0.14	0.21	0.03	-0.08	-0.38	0.10	-0.09	0.23	0.02
Yemen	0.26	-0.26	-0.21	0.17	-0.12	-0.17	-0.52	0.05	-0.33	0.15	-0.06

Table B.4: List of geotagged weighted average score sentiments in the Arab Region per topic

Uninfo*: Uninformative/Unrelated

Econ*: Economics

SU*: Stocking Up

Vacc*:Vaccine/Cure

Stat*: COVID-19 Statistics

Info*: COVID-19 News/Info

Pol*: Politics

PHM*: Public Health Measures

GM*: Governmental Measures

FT*: Fake Treatment

CT*: Conspiracy Theory

BIBLIOGRAPHY

- [1] “Arab region total population.” <https://data.worldbank.org/indicator/SP.POP.TOTL?locations=1A>.
- [2] U. Qazi, M. Imran, and F. Ofli, “Geocov19: a dataset of hundreds of millions of multilingual COVID-19 tweets with location information,” *ACM SIGSPATIAL Special*, vol. 12, no. 1, pp. 6–15, 2020.
- [3] “Labelbox.” <https://labelbox.com/>.
- [4] M. Imran, U. Qazi, and F. Ofli, “Tbcov: Two billion multilingual covid-19 tweets with sentiment, entity, geo, and gender labels,” *arXiv preprint arXiv:2110.03664*, 2021.
- [5] S. Kumar, R. R. Pranesh, and K. M. Carley, “A fine-grained analysis of misinformation in covid-19 tweets,” 2021.
- [6] S. A. Memon and K. M. Carley, “Characterizing covid-19 misinformation communities using a novel twitter dataset,” *arXiv preprint arXiv:2008.00791*, 2020.
- [7] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. D. S. Martino, A. Abdelali, H. Sajjad, K. Darwish, *et al.*, “Fighting the covid-19

infodemic in social media: a holistic perspective and a call to arms,” *arXiv preprint arXiv:2007.07996*, 2020.

- [8] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, X. Liu, and T. Zhu, “Twitter discussions and emotions about covid-19 pandemic: a machine learning approach (2020),” *arXiv preprint arXiv:2005.12830*.
- [9] M. Aljabri, S. M. Chrouf, N. A. Alzahrani, L. Alghamdi, R. Alfehaid, R. Alqarawi, J. Alhuthayfi, N. Alduhailan, *et al.*, “Sentiment analysis of arabic tweets regarding distance learning in saudi arabia during the covid-19 pandemic,” *Sensors*, vol. 21, no. 16, p. 5431, 2021.
- [10] S. Alqurashi, B. Hamoui, A. Alashaikh, A. Alhindi, and E. Alanazi, “Eating garlic prevents covid-19 infection: Detecting misinformation on the arabic content of twitter,” *arXiv preprint arXiv:2101.05626*, 2021.
- [11] M. S. H. Ameer and H. Aliane, “Aracovid19-mfh: Arabic covid-19 multi-label fake news and hate speech detection dataset,” *arXiv preprint arXiv:2105.03143*, 2021.
- [12] M. S. H. Ameer and H. Aliane, “Aracovid19-ssd: Arabic covid-19 sentiment and sarcasm detection dataset,” *arXiv preprint arXiv:2110.01948*, 2021.
- [13] L. Alsudias and P. Rayson, “Covid-19 and arabic twitter: How can arab world governments and public health organizations learn from social media?,” 2020.
- [14] “Twitter developer account.” <https://developer.twitter.com/en/apply-for-access>.

- [15] “Calculating the number of respondents.” https://help.surveymonkey.com/articles/en_US/kb/How-many-respondents-do-I-need.
- [16] Y. Qiao, C. Xiong, Z. Liu, and Z. Liu, “Understanding the behaviors of bert in ranking,” *arXiv preprint arXiv:1904.07531*, 2019.
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [18] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.
- [19] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.
- [20] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.
- [21] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, JMLR Workshop and Conference Proceedings, 2011.
- [22] S. Lai, L. Xu, K. Liu, and J. Zhao, “Recurrent convolutional neural networks for text classification,” in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

- [23] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, “Recurrent neural network based language modeling in meeting recognition,” in *Twelfth annual conference of the international speech communication association*, 2011.
- [24] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional lstm and other neural network architectures,” *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
- [25] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [26] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *CoRR*, vol. abs/1810.04805, 2018.
- [27] A. Safaya, M. Abdullatif, and D. Yuret, “KUISAIL at SemEval-2020 task 12: BERT-CNN for offensive speech identification in social media,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, (Barcelona (online)), pp. 2054–2059, International Committee for Computational Linguistics, Dec. 2020.
- [28] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, “Camembert: a tasty french language model,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [29] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.*, “Google’s neural machine translation system: Bridging the gap between human and machine translation,” *arXiv preprint arXiv:1609.08144*, 2016.

- [30] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.
- [31] Q. Yang, H. Alamro, S. Albaradei, A. Salhi, X. Lv, C. Ma, M. Alshehri, I. Jaber, F. Tifratene, W. Wang, *et al.*, “Senwave: monitoring the global sentiments under the covid-19 pandemic,” *arXiv preprint arXiv:2006.10842*, 2020.
- [32] “Global vectors for word representation.” <https://nlp.stanford.edu/projects/glove/>.
- [33] “Fasttext word vectors.” <https://fasttext.cc/docs/en/crawl-vectors.html>.
- [34] “Tensor flow.” <https://www.tensorflow.org/>.
- [35] “Keras tuner.” https://keras.io/keras_tuner/.