2021 9th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC) | 978-1-6654-8292-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/JAC-ECC54461.2021.9691451

An Arabic Egyptian Dialect COVID-19 Twitter Dataset (ArECTD)

Ahmed El-Sayed
Computer and Systems Department
Faculty of Engineering
Alexandria, Egypt
ahmed_elsayed@alexu.edu.eg

Shaimaa Lazem
City of Scientific Research
and Technological Applications
New Borg El-Arab, Alexandria, Egypt
slazem@srtacity.sci.eg

Mohamed Abougabal
Computer and Systems Department
Faculty of Engineering
Alexandria, Egypt
mohamed.abougabal@alexu.edu.eg

***Abstract*—Citizens are increasingly expressing their ideas and feelings on social media platforms such as Twitter. During the coronavirus crisis, numerous emotions are exposed, including sadness, anger, fear, sympathy, surprise, etc. The Arabic Egyptian Dialect COVID-19 Twitter Dataset (ArECTD), comprised of 78K tweets, was collected in the period from the 1st of January 2020 till the 30th of May 2021 focusing on the Egyptian dialect. It was annotated using a combination of manual and a semi-supervised self-learning technique. The tweets of ArECTD were categorized into 10 emotions (sarcasm, sadness, anger, fear, sympathy, joy, hope, surprise, love, and none). Emotion analysis of this dataset could help decision makers understand and respond to the public reactions during the pandemic.**

***Index Terms*—COVID-19; Twitter; Arabic Social Media.**

I. Introduction

Early 2020 has witnessed a rapid increase in the number of the confirmed COVID-19 cases causing a pandemic in March 2020. The first COVID-19 virus cases were confirmed in Egypt on 14 February 2020 [1]. Ever since, Egypt has taken strong precautionary and preventive steps to tackle the COVID outbreak and restrict its spread among people and residents across the country. During the pandemic, social media platforms like Twitter have become a source of information for many people on a variety of topics related to COVID-19. Twitter is used by many Egyptians to express and share opinions on COVID-19 decisions such as quarantine, curfew, wearing masks, emergency plans, and isolation hospitals. The total number of Egyptian twitter active users reached 5.25 million by October 2021 [2].

There is a growing interest in using emotion analysis to understand public reactions during the pandemic with few efforts focusing on the Arabic language. One of the key contributions of this research is to collect and annotate a large Arabic dataset related to COVID-19 focusing on the Egyptian dialect. The overarching goal is to analyze the public reactions of people in Egypt to certain decisions made by the government in response to COVID-19. A good understanding of the public reactions will help in planning future decisions and events. The Arabic Egyptian Dialect COVID-19 Twitter Dataset (ArECTD) consists of 78K tweets related to COVID-19 were collected in the period of the 1st of January 2020 to the 30th of May 2021. The dataset was labeled using a semi-supervised self-learning approach into ten emotions: sarcasm, hope, joy, surprise, sympathy, love, sadness, anger, fear and none.

The rest of this paper is structured in the following manner. The related work is reviewed in Section II. The research approach and the methods used are explained in Section III. In Section IV, a preliminary analysis of ArECTD presented and the insights are discussed. Finally, conclusion and possible future directions are discussed in Section V.

II. Related Work

Much of the recent works on COVID-19 focuses on analyzing social media data to track misinformation and fake news about COVID-19, understand the sentiments in COVID-19 tweets, and explore the mental health consequences of the pandemic. Redha and Al-Laith [3] proposed a system for detecting fake news surrounding COVID-19. They collected over 7 million Arabic tweets from January 2020 to August 2020 using the trending hashtags. A sample of 2,500 tweets was manually annotated into fake or genuine labels. Six machine learning classifiers were trained on the manually annotated dataset: Naive Bayes, Logistic Regression, Support Vector Machine, Multilayer Perceptron, Random Forest Bagging, and eXtreme Gradient Boosting. The best classifier was used to automatically predict the fake news classes of remaining unlabeled tweets.

Qazi at al. [4] presented GeoCoV19, a large-scale Twitter dataset related to the ongoing COVID-19 pandemic. The dataset has been collected in the period from February 1 to May 1, 2020 and consists of more than 524 million multilingual tweets with their location information. The dataset covers 62 international languages with 5.5 million Arabic tweets. Geographical coverage, users, and multilingualism were analyzed and an approach was employed to extract geolocation information from user location and tweet content at different granularity levels.

Alqurashi et al. [5] provided a large dataset of over 393 million Arabic tweets containing keywords related to COVID-19 and collected from January 1, 2020 until April 15, 2020, using Twitter streaming API and the Tweepy Python library. Similarly, Banda et al. [6] presented a large-scale curated dataset of over 383 million tweets related to COVID-19 in the period from January 1 to June 7, 2020. It was left open for researchers to conduct analysis on it.

Alhumoud [7] collected a dataset composed of 10 million Arabic tweets in April 2020 using corona keywords, that was reduced to 416,292 tweets only after cleaning and removing noise. The tweets were annotated using positive, negative and neutral sentiments, and classified using SVM, Smart GRU (SGRU), Stacked Bi-GRU (SBGRU), AraBERT as well as an ensemble model. The ensemble model outperformed the classifiers achieving an accuracy of 90.21%. The results showed that the negative sentiment was dominant as opposed to the positive and neutral.

Alhazmi and Alharbi [8] collected over 600K annotated Arabic tweets in Saudi Arabia related to COVID-19 from July 1 to July 31 of 2020. They developed an emotional lexicon-based approach to classify into eight emotion categories, namely: anger, fear, anticipation, trust, surprise, sadness, joy, and disgust. The lexicon contains 9922 words distributed over these emotions. Joy and anticipation were the most dominant among all emotions.

Imran et al. [9] analyzed the people reaction from different backgrounds to understand how different cultures behave and react given a global crisis like COVID-19. Two Twitter datasets were used: one consisted of 27,357 tweets collected by the authors during the period February 2020 to April 2020 using trending hashtags. The second dataset was an available Kaggle dataset that consisted of 460,286 tweets. The dataset had six labels: joy, surprise, sadness, fear, anger and disgust from six countries: Pakistan, India, Norway, Sweden, USA, and Canada. LSTM, BiLSTM, GRU and BERT models were used to identify the emotions of the tweets. The best performing model was LSTM with GloVe word embedding. Emotion and correlation analysis were made for the six countries. There was a high correlation between USA and Canada, and Pakistan and India. Joy dominates the positive tweets while sad and fear are the most dominant negative emotions.

Reviewing the related work revealed the lack of an Egyptian dialect dataset related to COVID-19, which would help to analyze the reactions on social media, especially Twitter, to major COVID-19 decisions. Also, the current datasets are collected over a short period (few months). Hence, there is a need to collect and annotate a large dataset during the outbreak period of the coronavirus focusing on the Egyptian dialect.

III. Approach

Our overarching goal is to use emotions as a basis for analyzing and understanding the opinions expressed by Egyptian citizens during the first year of the pandemic. Given the limitations of the reviewed literature, we aimed at collecting a large dataset which consists of 78K Arabic tweets from Egypt. The details of the methods used are described as following.

A. Collecting Dataset

The Arabic Egyptian Dialect COVID-19 Twitter dataset (ArECTD) was collected as follows. A list of the most common Arabic keywords and hashtags associated with COVID-19 was created as shown in Table I. TwitterScrapper [10] was used to scrape Tweets using the Python packages and parse the retrieved content. The query search date range was from 1-1-2020 until 30-5-2021. A dataset that contains around 1,597,939 tweets was obtained.

TABLE I: The List of hashtags used to collect the dataset.

Keyword	English Translation	Keyword	English Translation
#الكورونا	#Corona	#أزمة_كورونا	#Corona_Crisis
كوفيد٩١	Covid 19	#وباء_كورونا	#Corona_Epidemic
#فيروس_كورونا	#Corona_Virus	#فيروس_كورونا_المستجد	#Novel_Coronavirus
كورونا	Corona	#كورونا_الجديد	#New_Corona
كورونا مصر	Corona Egypt	كورونا فيروس	Corona Virus
#معا_ضد_كورونا	#Together_Against_Corona	إيقاف صلاة الجماعة	Stop prayer groupus
اغلاق المقاهي	Closing of cafes	ايقاف الدوري	Stop the league
تعليق الدراسة	Suspension of study	الوقاية من كورونا	Corona Prevention
اغلاق النوادي الرياضية	Closing Sports Clubs	#خليك_في_البيت	#Stay_Home
اعزل نفسك	Isolate yourself	حظر التجول	Curfew
رفع الحظر	Lift the ban	الحجر المنزلي	Home quarantine
العزل المنزلي	Home isolation	غسل اليدين	Washing hands
#الموجة_الثالثة	#Third_wave	#اللقاح	#The_vaccine
#اللقاح_طريقا_للتعافي	#Vaccines_way_to_recovery	#الموجة_الثانية	#Second_wave

B. Cleaning Dataset

The dataset included noisy data such as tweets written in other languages besides Arabic, inappropriate tweets, etc. Therefore, to have a better-quality dataset for labeling, the dataset was cleaned using the following steps.

- *Filtering out Non-Arabic tweets:* Many Arab users post tweets written in multiple languages. The Non-Arabic words were filtered out.
- *Filtering out tweets from news pages and focus on the tweets produced by individuals rather than organisations:* The dataset was cleaned by dropping tweets belonging to news pages and containing any term of the following: [news, newspaper, urgent, breaking news]
- *Filtering out Non-Egyptian tweets:* Tweets that have geo-location except Egypt were excluded.
- *Filtering out short tweets:* Mining short texts is very important as they are likely to be ambiguous and vague. So, tweets that have a number of words less than three were eliminated.
- *Filtering out duplicates:* The duplicated tweets were also excluded.

Applying the above cleaning steps, resulting in a cleaned version of the dataset that contains around 78,870 tweets.

C. Annotating Dataset

To annotate the ArECTD Dataset, initially a crowd sourcing process similar to [11] was followed. A group of 5 final year undergraduate students was employed. Testing and annotation tasks were designed to label the tweets using 8 labels (sadness, anger, joy, surprise, love, sympathy, fear and none). The initial discussions with the annotators in the early stages of the labeling task resulted in adding two more labels (sarcasm, hope). A total of 11,660 tweets were independently annotated. Majority voting was used to determine the tweet label, whereby 2 annotators had to agree on the same label for each tweet. The tweets breakdown is shown in Table II. The inter-annotator agreement between the annotators was measured using Fleiss's Kappa [12], a statistical measure used to measure the degree of agreement over what would be expected by chance. Fleiss's Kappa degree of agreement between the annotators was 0.746 which is considered substantial. The resulting manually annotated tweets were used by the group in their graduation project presented to Computer and Systems Engineering Department at the Faculty of Engineering, Alexandria, Egypt.

TABLE II: Emotions of the manually annotated 11,660 tweets.

Emotions	Count	Percentage (%)
Sarcasm	2877	24.67
Sadness	2367	20.30
Anger	2041	17.50
None	1163	9.97
Fear	800	6.86
Sympathy	797	6.84
Joy	702	6.02
Hope	478	4.09
Surprise	225	1.93
Love	210	1.80

To annotate the remaining tweets, the semi-supervised self-learning technique described in [13] was employed. Each tweet from the unlabeled data passed through three different classifiers to be classified with the emotion label. If the tweet had a prediction probability greater than or equal to a threshold of 0.8 in all classifiers, it was added to the training set and removed from the unlabeled set. The classification models, Logistic Regression, AraBERT [14] and GRU [15], were used in the self-learning process. Prior to the classification step, a data augmentation step was added to balance the subset of the data that is used to train the three classifiers. AraBERT word embedding was used to replace words with their synonyms that have similar vector representation. This process was repeated until the unlabeled set was empty as shown in Fig 1. The final labels of the dataset after the annotation are shown in Table III. Sadness, hope, and anger, were the three dominant emotions representing 15.02%, 12.30%, and 11.03% of the dataset. The remaining emotion labels including the None label were similarly represented in the dataset.

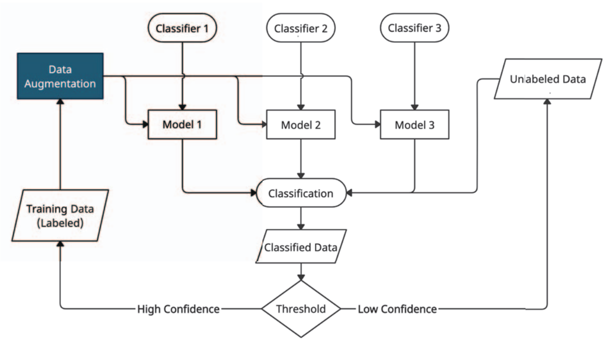

Fig. 1: Semi-supervised self-learning technique.

TABLE III: Emotions of the final dataset (78,870 tweets).

Emotions	Count	Percentage (%)
Sadness	11845	15.02
Hope	9705	12.30
Anger	8706	11.03
None	7402	9.38
Joy	7262	9.20
Love	7255	9.19
Sarcasm	7179	9.10
Fear	6649	8.43
Surprise	6543	8.29
Sympathy	6324	8.01

IV. Results and Discussion

Fig 2 shows the word frequency of the top 15 frequent words relevant to the chosen hashtags. The word "corona" scored the highest with a frequency of 70,208 words. The frequent words all relevant to the pandemic. Words such as God, Lord reflect the Islamic Egyptian culture norm of praying to God to alleviate sickness.

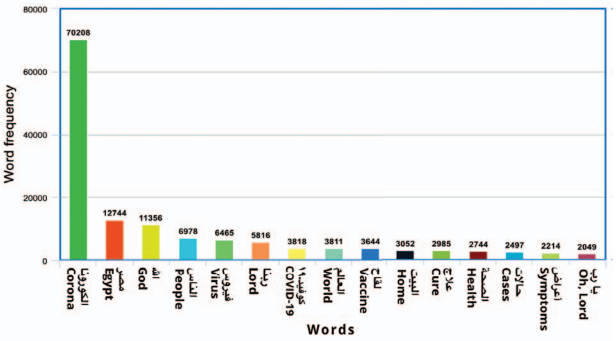

Fig. 2: Word frequency for the top 15.

The collected tweets were authored by 56,176 unique users. Further, Table IV shows the top cities from which the tweets were authored. The top contributors came from the cities of Assiot located in upper Egypt and Cairo the capital. Significantly fewer contributions came from the other seven cities from the North West and the Delta regions. The most dominant emotion for Assiot and Cairo are sadness, hope and anger. This might provide an indication to decision makers, given the negative dominant emotions of the tweets coming from both locations, that the COVID-19 situation in these cities requires extra attention.

The tweeting pattern overtime shows that a peak of communication in April 2020 and decreased overtime as depicted in Figure 3. The April 2020 peak is consistent with the increasing trend of the daily confirmed cases in the first COVID wave as shown in [16].

The emotion labeling of the ArECTD dataset shows that sadness, hope and anger emotions were dominant during the first year of the pandemic. The monthly distribution of the overall tweets is shown in Figure 4 .

TABLE IV: City distribution of ArECTD tweets.

City	Tweets Count	Percentage (%)
Assiot	29124	36.92
Cairo	23464	29.75
Alexandria	7336	9.30
Daqahlia	3427	4.35
Giza	2832	3.59
Gharbia	1668	2.11
Damietta	1612	2.04
Monofia	1515	1.92
Kafr El-Shikh	1036	1.31

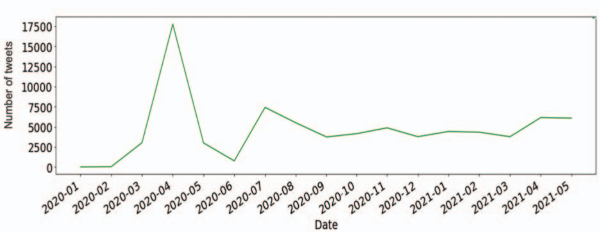

Fig. 3: Covid-19 tweets monthly distribution.

V. Conclusion and Future Work

A large Arabic COVID-19 Twitter dataset, ArECTD, was introduced. The dataset was annotated using semi-supervised self-learning technique and will be made available for free. The dataset is composed of about 78k labeled tweets collected from Egypt and is well-balanced over 10 common emotion labels (sarcasm, sadness, anger, fear, sympathy, joy, hope, surprise, love and none). The dataset was collected during the outbreak year of coronavirus from the 1st of January 2020 till the 30th of May 2021. The annotation process can be applied to any other dataset. To the best of our knowledge, this is the only Twitter dataset focused on the Egyptian dialect that was collected during COVID-19. It was observed that most of the emotions were contributed by Assiot (upper Egypt) and Cairo (the capital) citizens.

The dominant monthly emotions were identified. The initial explorations revealed that emotion analysis could be a useful tool to help decision makers understand and respond to the public reactions on social media. Our analysis highlighted the strong presence of positive emotions such as hope and the cultural traits such as sarcasm alongside fear, anger, and sadness reactions.

In the future, we plan to continue exploring the emotion trends in the dataset using quantitative and qualitative methods.

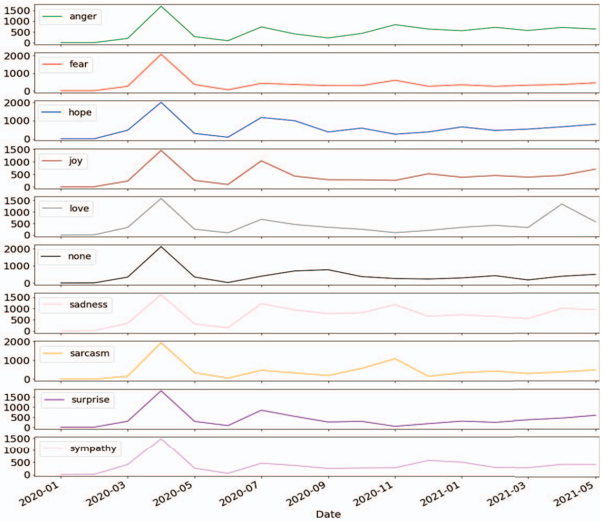

Fig. 4: Monthly emotion distribution of ArECTD.

References

[1] Egypt Today (2021). Egypt announces first Coronavirus infection [Online]. Available: https://www.egypttoday.com/Article/1/81641/Egypt-announces-first-Coronavirus-infection.

[2] Leading countries based on number of Twitter users [Online]. Available: https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries.

[3] A. Redha, A. Al-Laith, "Fake News Detection in Arabic Tweets during the COVID-19 Pandemic". International Journal of Advanced Computer Science and Applications(IJACSA), 12(6), 2021.

[4] U. Qazi, M. Imran, F. Ofli, "GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information", IEEE Dataport 2020.

[5] S. Alqurashi, A. Alhindi, E. Alanazi, "Large arabic twitter dataset on covid-19", arXiv preprint arXiv:2004.04315, 2020.

[6] J. Banda, R. Tekumalla, G. Wang, J. Yu, T. Liu, Y. Ding, G. Chowell, "A large-scale covid-19 twitter chatter dataset for open scientific research – an international collaboration", Epidemiologia, 2(3), 315-324, 2021.

[7] S. Alhumoud, "Arabic Sentiment Analysis using Deep Learning for COVID-19 Twitter Data', IJCSNS International Journal of Computer Science and Network Security, vol. 20, 2020.

[8] H. Alhazmi, M. Alharbi, "Emotion Analysis of Arabic Tweets during COVID-19 Pandemic in Saudi Arabia", International Journal of Advanced Computer Science and Applications (IJACSA), 11(10), 2020.

[9] A.S. Imran, S.M. Daudpota, Z. Kastrati, R. Bhatra, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets", IEEE Access 2020.

[10] Ahmet Taspinar (2020). GitHub: TwitterScraper [Online]. Available: https://github.com/taspinar/twitterscraper.

[11] I. Alsarsour, E. Mohamed, R. Suwaileh, T. Elsayed, "DART: A Large Dataset of Dialectal Arabic Tweets". The International Conference on Language Resources and Evaluation (LREC), 2018.

[12] K. Hallgren, "Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial". Tutorials in Quantitative Methods for Psychology, 8, 23-34, 2012.

[13] A. Al-Laith, M. Shahbaz, H. Alaskar, A. Rehmat, "AraSenCorpus: A Semi-Supervised Approach for Sentiment Annotation of a Large Arabic Text Corpus", Applied Sciences, 2021.

[14] A. Wissam, B. Fady, H. Hazem, "AraBERT: Transformer-based Model for Arabic Language Understanding", Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection 2020.

[15] Sh. Apeksha, D. Nyavanandi, L. Simone, "Performance Evaluation of Deep Neural Networks Applied to Speech Recognition: RNN, LSTM and GRU", Journal of Artificial Intelligence and Soft Computing Research, 2019.

[16] University of Oxford (2021). Our World in Data [Online]. Available: https://ourworldindata.org/coronavirus/country/egypt.