Correlation Analysis of Spatio-temporal Arabic COVID-19 Tweets

Tarek Elsaka
telsaka@sharjah.ac.ae
Agricultural Research Center
Cairo, Egypt

Imad Afyouni
Ibrahim Hashem
Zaher Al Aghbari
Computer Science Department, University of Sharjah
Sharjah, UAE

ABSTRACT

Since the recent COVID-19 outbreak, several researchers have begun to focus on various difficulties to data mining of social data to study people's reactions to the outbreak. Recent approaches have mostly concentrated on the analysis of social data in the English language. In this study, we present an in-depth social data mining approach to extract Spatio-temporal and semantic insights about the COVID-19 pandemic from the Arabic social data. We developed sentiment-based categorization methods to extract major topics at various location granularities (regions/cities). Besides, we used topic abstraction levels (subtopics and main topics). A correlation-based analysis of Arabic tweets and official health provider data will also be presented. Furthermore, we used occurrence-based and statistical correlation methodologies to create many topic-based analysis mechanisms. Our findings demonstrate a positive association between top subjects (for example, lockdown and vaccine) and the increasing number of COVID-19 new cases, but unfavorable attitudes among Arab Twitter users were generally heightened during this pandemic, on issues such as lockdown, closure, and law enforcement.

CCS CONCEPTS

• **Information systems → Spatial-temporal systems**.

KEYWORDS

Spatio-temporal, Arabic Tweets, COVID-19 Pandemic, Correlation Analysis, Sentiment Analysis

ACM Reference Format:
Tarek Elsaka, Imad Afyouni, Ibrahim Hashem, and Zaher Al Aghbari. 2021. Correlation Analysis of Spatio-temporal Arabic COVID-19 Tweets. In *2nd ACM SIGSPATIAL International Workshop on Spatial Computing for Epidemiology (SpatialEpi 2021) (SpatialEpi'21), November 2, 2021, Beijing, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3486633.3491092

1 INTRODUCTION

Twitter and other social media platforms have become home to a variety of real-life events. The recent outbreak of COVID-19 has prompted immediate action from scientists to address many challenges from many research perspectives. Among these challenges, studies associated with the processing and analysis of social data have been intensified, by utilizing Sentiment Analysis (SA) [20] to extract knowledge from people's thoughts, opinions, and feelings. As a result, public opinion on COVID-19-related topics can be tracked and monitored in location and time [21]. Users and posts can also be identified by their location, time, and/or language, although some users' information may be restricted or unavailable. This allows researchers to focus their efforts on researching and investigating a specific trend of interest, utilizing massive datasets from social media as a valuable resource to detect people's reactions to the pandemic.

According to the most recent Internet global data[14], more than 250 million Arabic Internet users [14] in Arab countries, most of them use social media to communicate and contribute daily Arabic material, particularly on Twitter. As a result, researchers looking to learn more about people's reactions to the COVID-19 outbreak might use the Arabic information available on Twitter [23]. Unfortunately, few previous research studies were done on Arabic social data and few of them take the spatial-temporal element of sentiment analysis into account. Besides, few studies have been applied on the relationship between official health statistics and social media content.

In this study, we analyze Arabic social data related to the COVID-19 pandemic collected from January to November 2020 (about 5.5M tweets) to identify people's sentiments and correlations between COVID-19 related themes and subtopics at various Spatio-temporal granularities. Furthermore, we intend to emphasize correlations between social data and official health data records, as well as investigate the impact of the global pandemic on several elements at various Spatio-temporal granularities. This paper is organized as follows: Section 2 outlines a review of some related work. The description and implementation of the proposed methodology are presented in Section 3. Section 4 illustrates how data were collected and classified. The results and findings of the proposed methodology are discussed in Section 4. Section 5 presents concluding remarks and future research directions.

2 RELATED WORK

Researchers began investigating COVID-19-related social media data in early 2020 [7]. Previous analyses of social media content have primarily focused on English tweets or other Latin languages while few academics have investigated Arabic social content, particularly on Twitter. Many researchers have been interested in word embedding models because they can capture both semantic and

SpatialEpi'21, November 2, 2021, Beijing, China

ACM ISBN 978-1-4503-9119-1/21/11…$15.00
https://doi.org/10.1145/3486633.3491092

syntactic characteristics of words from a huge unlabeled corpus, with many of them producing cutting-edge results in many NLP tasks [16]. As a result, several studies encouraged topic analysis to highlight the hot subjects talked on social media utilising word embedding, word frequency, location frequency, language frequency, and character and word n-gram feature weighted by TF-IDF. Other researchers, on the other hand, used feature extraction in feature-based sentiment analysis to detect sentiment polarity and forecast sentiment in social data [6]. Many of them employed ML classifiers to validate semantic analysis results.

Some academics worked on location-enabled social data features, such as Qazi et al. [22]. They derived their geolocation information using a gazetteer-based approach to extract toponyms from user location and tweet content using Nominatim (Open Street Maps) data at geolocation granularity levels. Similarly, Lamsal [17] released the COV19Tweets Dataset, a large-scale English language tweets dataset with sentiment ratings. They filtered the geotagged tweets to generate the Geo-Dataset that comprised only 141k tweets (0.045 percent of the original dataset). While, Alshalan et al. [5], Alsafari et al. [4], Hamoui et al. [12], and Al-Laith et al. [1] studied Arabic content on Twitter to see which themes were most popular among Arabic usersduring the COVID-19 pandemic. Similarly, Bahja et al. [8], Essam and Abdo [10], and Manguri et al. [19] performed sentiment analysis to recognize the relevance of the feelings/emotions regarding the COVID-19. Additionally, Chakraborty et al. [9] demonstrated that tweets comprising all handles relevant to COVID-19 and World Health Organization (WHO) have failed to guide people effectively on this pandemic epidemic. Finally, some researchers such as Alomari et al. [3] Used distributed machine learning to detect government pandemic measures and public concerns using Twitter Arabic data.

3 DATA COLLECTION AND CLASSIFICATION

Using machine learning models and topic recognition and tracking techniques, we offered a method to automatically detect and process Arabic social datasets related to the COVID-19 epidemic. We studied the Spatio-temporal social data in Arabic tweets related to the COVID-19 epidemic. We created our dataset by merging two publicly available Arabic datasets for COVID-19 tweets: [2] (3,314,859 tweets) and [13] (3,314,859 tweets) (2,111,650 tweets). Unfortunately, both datasets have geo-tweets for only about 2% of total tweets. Then, we inferred locations from non-geotagged tweets to generate a new dataset of location-enabled tweets about 46% (around 2.5M tweets) of the original combined dataset. Most Arabic geotagged tweets come from Arab users around the world, 90% of them were originated from 22 Arabic-speaking countries. Then we developed several pre-processing activities for dataset Feature Extraction, such as {Clean Dataset}, {Filter Fields}, and {Prepare Arabic Text}. These processes remove null values from the tweet's object, removes the unnecessary fields from the tweet's metadata, and prepare the Arabic text in each tweet's object for further text analysis.

Our location-inferring method operated on two levels to extract both Chrononyms and Astionyms from user location [15]. The first level tries to deduce the nation name from the metadata written in the user's location. The second level, on the other hand, only works if the first level fails, and it gets the country name by detecting the city name from user location details. Then, from the COVID-19-related tweets, important features are extracted. The retrieved feature vectors are then subjected to classification techniques in order to find and produce hot topics from the tweet's text. The sentiment analysis technique, on the other hand, is used in conjunction with different viewpoints to infer sentiment and opinion polarity at a variety of spatial (city, nation, region) and temporal (day and month) levels. In the SA processes, we used the lexicon-based approach with Arabic tweets as an opinion resource. We used the Arabic lexicon to extract the polarity feature of the tweet's text. Then we used the Bag of Words (BoW) technique to detect the polarity feature of the tweet's as positive, negative, or neutral. We developed a Sentiment Extraction algorithm to conduct SA processes including loading the Arabic corpus, preparing the Arabic corpus, extracting the tweet's polarity, and evaluating the findings with machine learning classifiers.

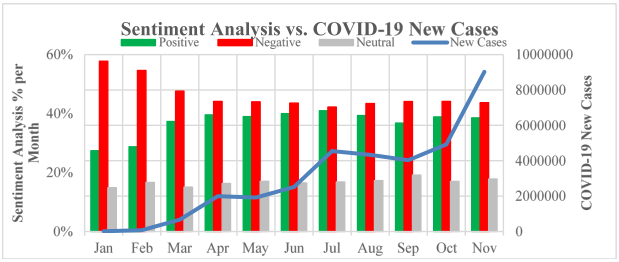

Figure 1: Correlation between Sentiment Analysis and the Official Health Records over the world

Finally, using multiple spatial granularities, such as country and area, a correlation study between the gathered Twitter data and the official health records is performed. Finally, we illustrate visual analytics using the new Arabic COVID-19 dataset and a comprehensive set of experiments.

4 CORRELATION ANALYSIS

We obtained official COVID-19 records about COVID-19 numbers daily from World Health Organization (WHO), the European Centre for Disease Prevention and Control (ECDC), and Johns Hopkins University (JHU) from January to November 2020. We used the occurrence-based model to build several methods for correlation-based analysis to determine the association between sentiment analysis, official health provider data, lockdown information, and topic frequencies. Figure 1 depicts the relationship between sentiment analysis of Arabic tweets linked to COVID-19 and the number of new COVID-19 cases around the world. It shows that some elevated negative feelings at the start of the COVID-19 pandemic until the second quarter of 2020, after which it was in a steady condition while cases rose. Furthermore, Figure 2 demonstrates the correlation between feelings of Arab tweets in Arab countries and Non-Arab countries, as well as the COVID-19, confirmed cases, with some variances in both, particularly in the fourth quarter of 2020.

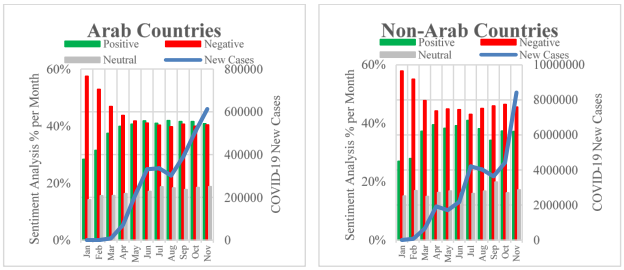

Figure 2: Correlation between Sentiment Analysis and the Official Health Records in Arab and Non-Arab Countries

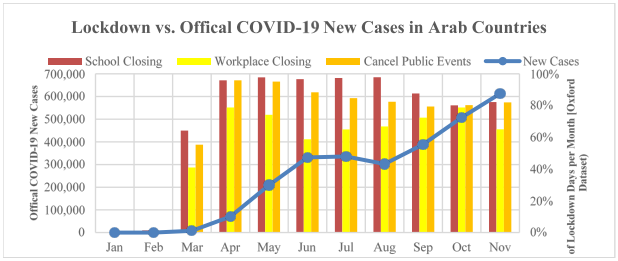

Figure 3: Correlation between Lockdown and Official COVID-19 New Cases in Arab Countries

Figure 2 shows a significant difference in the correlation between feelings indicated in Arabic tweets and the COVID-19 new cases published in some countries throughout the world. While the number of official new cases of COVID-19 increases, most feelings are balanced between positive and negative. Governments are adopting many measures against the COVID-19 pandemic in reaction to the COVID-19 epidemic, such as 'school closures', 'workplace closures', 'cancellation of public events', and 'travel restrictions'. The first COVID-19 pandemic lockdown was launched on January 23, 2020, in Wuhan [18]. Most countries throughout the world have imposed complete or partial lockdowns to prevent the virus from spreading, leaving millions stranded. Furthermore, one-third of the world's population is locked in some form [24]. Several Arab countries began imposing various sorts of lockdowns in the middle of March 2020.

Some institutions, like as Oxford University, collect data on 20 parameters to inform a Risk of Openness Index, which intends to assist governments in determining if it is safe to 'open up' or 'shut down' in their fight against the coronavirus [11]. We obtained information about the global lockdown status from the Oxford dataset. We used data from the indicators 'School shutting', 'Workplace closing', and 'Cancel public activities' as the most relevant indicators to feed our mechanism employed in an experiment intended at determining the association between lockdown and the spatial-temporal data found in our Arabic tweets dataset. Figures 3 illustrates the correlation of the lockdown with COVID-19 new cases in Arab countries, demonstrating the impact of lockdown on the spread of COVID-19 new cases. It demonstrates a negative relationship between the number of new COVID-19 instances and the number

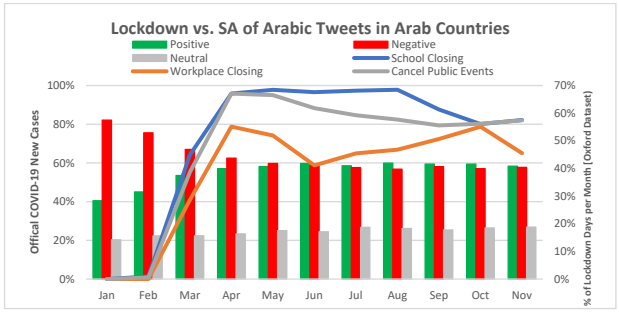

Figure 4: Correlation between Lockdown and Sentiment Analysis in Arab Countries

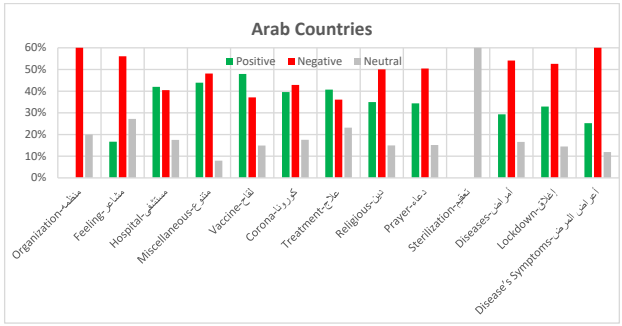

Figure 5: Correlation between Main Topics and Sentiment Analysis in Arab Countries

of lockdown days in the Arab area – particularly for the indicator 'Cancel Public events'. In the Non-Arab region, there is a negative link between the number of new COVID-19 cases and the number of lockdown days – particularly for the indicators 'cancel public activities' and 'school closing'.

Figures 4 show the relationship between lockdown and sentiment analysis of Arabic tweets tweeted by users between January and November 2020, despite the fact that the lockdown began late in Arab countries. The relationship demonstrates different perspectives on the lockdown indicators and the sentiment of Arabic tweets about COVID-19 in Arab and non-Arab countries. For example, they demonstrate a favourable relationship between the number of days a school is closed and the quantity of positive emotions expressed in Arabic tweets.

Another approach was created to determine the spatial-temporal correlation between sentiment analysis and topic-frequencies. Following that, in order to obtain a broad view of the subjects identified in Arabic tweets, we clustered them into main top topics. Figures 5 depicts the association between primary subjects and sentiment analysis of Arabic tweets published in Arab countries. It demonstrates the good and negative emotions connected with each topic in that region.

5 RESULTS AND DISCUSSION:

Researchers can analyse the impact of health preventive measures, environmental conditions, and discussion topics on the spread of the COVID-19 pandemic using spatial-temporal analysis of social media posts. Such spatiotemporal analysis necessitates the geotagging of social media content. However, only a small proportion of the collected stream of postings (4.5 percent at most) gets geotagged. As a result, in this study, we created a novel location inference technique from non-geotagged tweets based on user profiles and textual content, raising the total percentage of geotagged tweets from 2% to 46%. (about 2.5M tweets). This helped us to investigate the correlation between spatiotemporal social data and official health data, which revealed a positive relationship between top themes (such as Treatment and Vaccine) and the increasing number of COVID-19 new cases. Furthermore, we used a sentiment analysis technique at several spatial granularities, such as city and country, as well as independent topic scales. According to official figures, the United States, India, and Brazil have the highest total number of COVID-19 cases. According to certain correlation analyses, the unfavourable attitudes of Arab Twitter users were heightened throughout the pandemic. Furthermore, we demonstrated a link between the themes discussed on Arabic social media, such as government-enforced lockdowns and travel restrictions, and the amount of new COVID-19 instances. Furthermore, the data revealed a strong relationship between Arab users' unpleasant feelings and the number of daily confirmed cases of COVID-19.

6 CONCLUSION:

This research presented a complete social data mining approach for extracting COVID-19-related insights in Arabic, with a focus on the correlation between spatiotemporal social data and health data. Furthermore, a strategy for inferring geo-information from non-geotagged tweets was created, which increased the total percentage of location-enabled tweets from 2% to 46%, outperforming most prior relevant efforts. In addition, many topic-based analytical methods based on occurrence-based and statistical correlation approaches were implemented. A correlation-based study of Arabic tweets and official health provider data was also given in the paper. Our findings demonstrate that combining social data mining with other data sources, such as health data, has a high potential for predicting the evolution of such occurrences. Furthermore, such correlation can later be used to other forms of data, such as contact tracing and GPS data, to provide a comprehensive picture of human behaviour and the relationship between social and physical user interactions.

REFERENCES

[1] Ali Al-Laith and Mamdouh Alenezi. 2021. Monitoring People's Emotions and Symptoms from Arabic Tweets during the COVID-19 Pandemic. *Information* 12, 2 (2021), 86.

[2] Eisa Alanazi, Abdulaziz Alashaikh, Sarah Alqurashi, and Aued Alanazi. 2020. Identifying and Ranking Common COVID-19 Symptoms From Tweets in Arabic: Content Analysis. *Journal of medical Internet research* 22, 11 (2020), e21329.

[3] Ebtesam Alomari, Iyad Katib, Aiiad Albeshri, and Rashid Mehmood. 2021. COVID-19: Detecting government pandemic measures and public concerns from Twitter arabic data using distributed machine learning. *International Journal of Environmental Research and Public Health* 18, 1 (2021), 282.

[4] Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. 2020. Hate and offensive speech detection on arabic social media. *Online Social Networks and Media* 19 (2020), 100096.

[5] Raghad Alshalan, Hend Al-Khalifa, Duaa Alsaeed, Heyam Al-Baity, and Shahad Alshalan. 2020. Detection of Hate Speech in COVID-19–Related Tweets in the Arab Region: Deep Learning and Topic Modeling Approach. *Journal of Medical Internet Research* 22, 12 (2020), e22609.

[6] Muhammad Zubair Asghar, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi. 2014. A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research* 4, 3 (2014), 181–186.

[7] Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale Arabic sentiment lexicon for Arabic opinion mining. In *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*. 165–173.

[8] Mohammed Bahja, Rawad Hammad, and Mohammed Amin Kuhail. 2020. Capturing Public Concerns About Coronavirus Using Arabic Tweets: An NLP-Driven Approach. In *2020 IEEE/ACM 13th International Conference on Utility and Cloud Computing (UCC)*. IEEE, 310–315.

[9] Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. 2020. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing* 97 (2020), 106754.

[10] Bacem A Essam and Muhammad S Abdo. 2021. How Do Arab Tweeters Perceive the COVID-19 Pandemic? *Journal of psycholinguistic research* (2021), 507—-521.

[11] Thomas Hale, Anna Petherick, Toby Phillips, and Samuel Webster. 2020. Variation in government responses to COVID-19. *Blavatnik school of government working paper* 31 (2020), 2020–11.

[12] Btool Hamoui, Abdulaziz Alashaikh, and Eisa Alanazi. 2020. COVID-19: What Are Arabic Tweeters Talking About?. In *International Conference on Computational Data and Social Networks*. Springer, 425–436.

[13] Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. Arcov-19: The first arabic covid-19 twitter dataset with propagation networks. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*. 82–91.

[14] Simon Kemp. 2021. Digital 2021: Global Overview Report. https://datareportal.com/reports/digital-2021-global-overview-report

[15] EA Khomutnikova, EV Gunbina, MS Zhurkova, and FV Fetyukov. 2020. SEMANTICS AND ETYMOLOGY OF ENGLISH ASTIONYMS IN THE ASPECT OF LINGUISTIC GEOGRAPHY. In *European Proceedings of Social and Behavioural Sciences EpSBS*. 505–513.

[16] Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. 2016. How to generate a good word embedding. *IEEE Intelligent Systems* 31, 6 (2016), 5–14.

[17] Rabindra Lamsal. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence* 51, 5 (2021), 2790–2804.

[18] Rita Yi Man Li and Xiao-Guang Yue. 2020. Covid-19 in Wuhan: pressing realities and city management. *Frontiers in public health* 8 (2020), 1079.

[19] Kamaran H Manguri, Rebaz N Ramadhan, and Pshko R Mohammed Amin. 2020. Twitter sentiment analysis on worldwide COVID-19 outbreaks. *Kurdistan Journal of Applied Research* 5, 3 (2020), 54–65.

[20] Amr Mansour Mohsen, Hesham Ahmed Hassan, and Amira M Idrees. 2016. A proposed approach for emotion lexicon enrichment. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* 10, 1 (2016), 242–251.

[21] A. M. Mostafa. 2017. Advanced Automatic Lexicon with Sentiment Analysis Algorithms for Arabic Reviews. *American Journal of Applied Sciences* 14 (2017), 754–765.

[22] Umair Qazi, Muhammad Imran, and Ferda Ofli. 2020. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special* 12, 1 (2020), 6–15.

[23] Y. Salameh. 2020. How Many Countries Speak Arabic Around the World? https://www.tarjama.com/how-many-countries-that-speak-arabic-around-the-world/

[24] AKM Ahsan Ullah, Faraha Nawaz, and Diotima Chattoraj. 2021. Locked up under lockdown: The COVID-19 pandemic and the migrant population. *Social Sciences and Humanities Open* 3, 1 (2021), 100126.