Sentiment Analysis and Emotion Detection of Egyptians Tweets during COVID-19 Pandemic

Dr. Wael K. Hanna

Sadat Academy for Management Sciences,
Information System Dept.,
Cairo, Egypt
E-mail: Wael_karam1@yahoo.com

Abstract

Social media considered as effective platform allowed people to express their opinion and feeling during the COVID-19 crisis. The method of determining whether a block of text is good, negative or neutral is known as sentiment analysis. Sentiment analysis aims to analyze people›s opinions as its main objective. There are many challenges for Arabic sentiment analysis such as the Arabic language›s complexity, the dataset related to Arabic Sentiment Analysis is small; and difficult representations of emotion. In this paper, we investigate public's emotional responses associated with this pandemic using Twitter as platform to perform our analysis in Egypt. While people express positive emotions, there are tons of fear, anger, and sadness revealed. First, we collect dataset of Egyptians tweets during Covid-19. Then, we apply Sentiment analysis method to classify the Egyptian tweets. In addition, we develop an emotion detection method to classify tweets into standard eight emotions. Moreover, this research might help to better understand public behaviors to gain insight and make the proper decisions.

Keywords: Sentiment Analysis, COVID-19, Twitter, Egypt.

1.Introduction

Corona virus spreads across over the world, it causes panic to all people, for Geographic distribution of COVID-19 cases over the world. It causes serious symptoms related to respiratory system especially for older people, and those suffering from chronic diseases. It's too hard to diagnose the Corona virus disease because the need to conduct tests for COVID-19 based on your signs and symptoms and it takes long time to get results. The Corona crisis has a major impact in different aspect of life economically and socially.

One of the easiest and available sources for analyzing and detecting human mood is social media sites [1]. During the Corona issue, social media sites like twitter were a primary source of information for many individuals on a variety of topics relating to the COVID-19. This prompted researchers to investigate and analyze people›s responses to the Covid 19 pandemic and its consequences on Twitter. [1][2],[3],[4].

Positive or negative expression of words is referred to as sentiment. Sentiment analysis is a useful tool for assessing the tone of spoken or written words to determine their degree of positivity, negativity, or neutrality. According to [1], there are many challenges for Arabic sentiment analysis such as;

1) The Arabic language›s complexity causes grammatical, semantic, and metaphorical confusion due to its spellings, lexicon, phonetics, and grammar;

2) The research related to Arabic Sentiment Analysis is small;

3) Sentiment analysis in live time, spam identification, grammatical errors, spelling mistakes, unstructured information, and hidden meanings;

4) Symbolic representations of emotion; recognizing similes, metonymy, hyperbole, and ambiguity is difficult for humans, and considerably more difficult for machines.

In [2], the authors created a huge Arabic tweets dataset on COVID-19 that collected since January, 2020. In [5] Saudi Arabic Tweets were classified into typical eight emotions using a motion detection algorithm. Their findings demonstrated that among all feelings, joy and anticipation are the most prevalent. While people display good feelings, fear, anger, and grief are also present. In [6], the author build a Naïve Bayes model to implement Arabic sentiment analysis of Saudi Arabic twitter posts. Saudi People on twitter expressed their support and delight for the COVID-19 precautions.

Most of previous research focused on Arabic sentiment analysis of COVID-19 in Saudi Arabic. In this research, we present textual analysis of Arabic tweets to detect public emotions in Egypt after two years of the Covid 19 Pandemic. Our model classifies tweets into positive and negative tweets. One of key contributions of this research is developing a system that can label and score Arabic text according to the standard emotions. Another key is analyzing the perception of Egyptians people towards COVID 19, and giving insight into their feeling and reactions.

The rest of this paper is organized as follows. Section 2 reviews related work and Section 3 introduces the research methodology. In Section 4, we state sentiment analysis and emotion detection method. Results presented in Section 5, followed by conclusion in Section 6.

2. Related Work

People use social media platforms such as twitter to express their thoughts, ideas, and emotions. Twitter is a fertile ground for researchers to examine and analyze people›s beliefs and behaviors. For Arabic sentiment analysis, many researchers provide an Arabic twitter dataset. In [2] the authors created a huge Arabic tweets dataset on COVID-19 that collected since January 2020 (containing 3,934,610 million tweets). We used this dataset to extract the Arabic tweets from Egypt, because it includes almost 4 million tweets and include many tweets belong to Egyptians or related to Egypt. Also, [3] the authors presented an ArabicCOVID-19 twitter dataset that includes the months of January through March 2020. This dataset Covers the COVID-19 epidemic, with roughly 748k frequent tweets (based on twitter search criteria). In [4] 95K tweets were used to construct a big dataset with three types of sentiment labels (positive, negative, and neutral). This is a general dataset not related to Covid 19.

Table1: Summary of Arabic Datasets

Work	Duration	Description	Covid 19	Egyptian tweets
Sarah et al.,2020 [2]	January, 2020	3,934,610 million tweets	Yes	Yes
Fatima et. al., 2020 [3]	Months of January through March 2020.	748k tweets	Yes	No
Basma et. al., 2021 [4]	May 2012 and April 2020	95K tweets	No	Yes

Because the importance and effects of Covid 19 on the world, many researchers are working on Covid 19 and its impacts. Specially to analyze the people ideas, emotions, and reaction for the Covid 19 pandemic.

In [7], this study uses machine learning (ML) and natural language processing (NLP) techniques to extract subjective data, establish polarity, and identify feelings. then several classification methods were examined.

In [8], the authors use Facebook–a platform that is seldom ever used–to track the development of COVID-19-related phenomena. The suggested analytics method includes a data collection stage, pre-processing, topic modelling based on Latent Dirichlet Allocation (LDA), and a presentation module employing graph structure. The big limitation of this study is integrating semantic technologies to examine the semantic dimension using embedding models in the assessment of the multilingual and cross-lingual semantic similarity/relatedness.

In [9] the twelve weeks since the COVID-19 pandemic›s onset, this study looked at how Arab communities responded to it on twitter. Including coronavirus tweets of the spread of the pandemic, metaphysical responses, symptoms and signs in confirmed instances, and conspiracism.

In [10], the Arabic COVID-19 dataset was processed using the N-gram feature extraction method, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Naive Bayes (NB) machine learning predictive models. The big limitation of this study is missing feature selection and parameter tuning to build the ML algorithms.

This essay [11] investigates Saudi Arabians› perceptions about online education. The proposed sentiment analysis revealed that respondents› attitudes towards online learning remained neutral. Researchers and decision-makers will be able to comprehend the emotional effects of online learning on communities thanks to this study. With sentiment analysis, it is necessary to comprehend how individuals feel about various online learning platforms.

3. Methodology

Figure 1 presents the followed steps to apply sentiment analysis on dataset extracted from twitter.

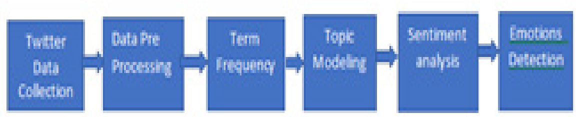

Figure1: Research Methodology

3.1 Data collection

Because people frequently utilized twitter to express how they felt about a situation, the twitter dataset was chosen as the data source.

3.2 Data Preprocessing

After we collected the data, we apply preprocessing steps to clean the tweets as follows:

1. Remove unmeaningful tweets
2. Remove accents
3. Remove unwanted symbols
4. All emojis, URLs, and special characters were removed
5. Stop words are removed

3.3 Term Frequency

Term Frequency is referred to as an Inverse Document Frequency (IDF). records with Inverse Document Frequency. It can be summed up as determining how pertinent a word is to a corpus or series of words in a text. The frequency of a term in the corpus offsets the meaning increase that occurs when a word appears more frequently in the text. Figure 2 shows the term Frequency in our dataset.

Figure 2: Examples of Cloud words in Twitter Data

3.4 Topic Modeling

Topic modelling is different from text classification and clustering tasks because it uses unsupervised data. Topic Modeling does not seek to identify commonalities between documents, unlike text classification or clustering, which tries to simplify information retrieval and create clusters of texts. Typically, there are several themes and a variety of texts in topic modelling.

A type of statistical modelling tool called topic modelling is used to determine which abstract themes in a collection of documents are being discussed at all. By design, topic modelling addresses the challenge of unsupervised topic generation. The statistical approach is typically employed by taking into account that each document discusses a variety of themes, and that each topic is typically indicated by a distribution of words. It is assumed that the documents have a two-step format, such as document = [subject 1, topic 2, topic 3,...,topic N] and then topic 1 = [w 1,w 2,...,w N]. There is no doubt that counting words, their proportions, and associated indicators is how topic modelling is often carried out.

• Latent Dirichlet Allocation (LDA) represents an endless variety of themes probabilities that are reflected in a document and is based on the Bayesian approach to representing all sorts of statistical uncertainties in probabilities.

• Latent Semantic Analysis: This approach aids in maintaining texts and words in a semantic space for categorization by using Singular Value Decomposition as a technique.

• The probabilities of a word in topic and a topic in a text are used in Probabilistic Latent Semantic Analysis (PLSA), which can be trained using an expectation-maximization approach. The multinomial distribution of words is the foundation of this methodology.

3.5 Sentiment Analysis

The method of determining whether a block of text is good, negative, or neutral is known as sentiment analysis. Sentiment analysis is the contextual mining of words that reveals the social sentiment of a brand and aids businesses in determining if the product they are producing will find a market. Sentiment analysis›s objective is to examine public sentiment in a way that will support corporate growth. It emphasizes emotions as well as polarity (positive, negative, and neutral) (happy, sad, angry, etc.). It makes use of a variety of Natural Language Processing techniques, including Automatic, Hybrid, and Rule-based.

• Sentiment analysis with finer resolution: This is based on polarity. A very positive, positive, neutral, negative, or extremely negative category could be created for this. Ratings are given on a scale of 1 to 5. If the rating is 5, it is highly positive; if it is 2, it is negative; and if it is 3, it is neutral.

• Emotion detection refers to the recognition of feelings such as happiness, sadness, rage, sorrow, merriment, and so forth. The lexical approach of sentiment analysis is another name for it.

• Aspect-based sentiment study focused on a certain aspect. For example, if a person wants to evaluate a cell phone›s feature, they would check the aspect such as battery, screen, and camera quality.

• Multilingual sentiment analysis: When analyzing sentiment in many languages, it is necessary to categorize the sentiment as positive, negative, or neutral. This is quite demanding and challenging in comparison.

Three methods are employed:

• Rule-based approach: In this case, tokenization, parsing, and the lexicon method are rule-based. The strategy counts how many positive and negative terms are present in the sample. If there are more positive words than negative words, the emotion is positive; otherwise, it is the opposite.

• Automatic Approach: This strategy relies on machine learning. Predictive analysis is first performed once the datasets have been trained. Word extraction from the text is the subsequent procedure. Different methods, including Naive Bayes, Linear Regression, Support Vector, and Deep Learning, can be employed to extract text, just like these machine learning techniques.

• Hybrid Approach: This method combines the two methods mentioned previously.

3.6 Emotion Detection

Emotions are an essential part of human existence. These feelings affect how people make decisions and improve our ability to communicate with others. Emotion recognition, which is another name for emotion detection, is the recognition of a person›s many sentiments or emotions (such as happiness, sadness, or anger). Emotion identification from text is also made more difficult by the numerous ambiguities and new terminology or terminologies that are being introduced daily. In addition, emotion detection seems to go up to a 6-scale or 8-scale based on the emotion model, going beyond merely identifying the basic psychological states (glad, sad, and angry). [12]

4. Experiment and Results

This section presents the experiment details and the results.

4.1 Experiment Setup

Covid-19 related tweets are used in the experiment. Using TWINT, a twitter scraping tool, the dataset was obtained from twitter. We focused on Egyptian tweets and hashtags about how they feeling toward COVID 19 such as (#Covid #Corona) only tweets from the Egyptians in quarantine period, which began on March 23, 2020, were included. In addition, we used collected dataset from [2] to extract the Arabic tweets from Egypt, because it includes over 4 million tweets and include many tweets belong to Egyptians or related to Egypt.

4.2 Results of Term Frequency

Word	TF-IDF
rt	0.063
كورونا	0.050
mohpegypt	0.044
مصر	0.043
من	0.037
فيروس	0.035
في	0.031
كوفيد١٩	0.029
المستجد	0.027

Figure 3: Top Term Frequency in Our Dataset.

The top term frequency are RT with 0.063, with 0.050 كورونا, mohpegypt with 0.044, with 0.043مصر, and فيرس with 0.035 as shown in figure 3. Then we count the words of characters of each tweet. For example in first tweet in figure, the count of words are 8, and the count of characters are 38 as shown in figure 4.

Figure 4: Count the Words of Characters of Each Tweet

Topic	Topic keywords
1	في, مصر, علي, ..., التباعد, كورونا, المصري, مصري, الناس, rt
2	مصر, الصحة, ..., حالات, الى, mohpegypt, من, كورونا, فيروس, rt
3	كوفيد١٩, المستجد, مصر, من, ؟, كيف, mohpegypt, كورونا, فيروس, rt
4	كورونا, مصري, في, عن, EG, مشاريعمصر, egyprojects, و, rt
5	في, مصر, الى, كورونا, من, ..., الصين, علي, مساعدات, rt
6	في, كورونا, مصر, وزارة, ا, علي, بعد, ..., للاستفسارات, rt
7	بعض, ahmedlsaadany, من, ..., لهم, وباء, ما, و, كل, rt
8	مصر, كورونا, و, ..., من, تحيا, بفيروس, آخر, مليون, rt
9	نصائح, المناعة, و, mohpegypt, كورونا, مصر, في, ..., من, rt
10	مصر, ف, rt, ع, كورونا, ان, ال, امريكا, اجراءات, كل

Figure 5: Top Topic Keywords

As shown in figure 5, Results showed top topics Keywords from the dataset for example; topic 1 top keywords are: في مصر - التباعد- كورونا - المصرى والناس .

Then each tweet was selected to classify the tweet to negative or positive and belong to each topic as the figure below with specific weight. For example tweet6 in the figure, belong topic1 with 0.871995, topic2 with 0.334417, topic3 with 0.365785, and topic4 with -0.129198, topic5 with -0.725574, topic6 with 0.0285206, topic7 with 0.481769, topic8 with 0.199736, topic9 with 0.148989, and topic10 with 0.00662068. Therefore, the tweet6 belong to topic1 and topic7.

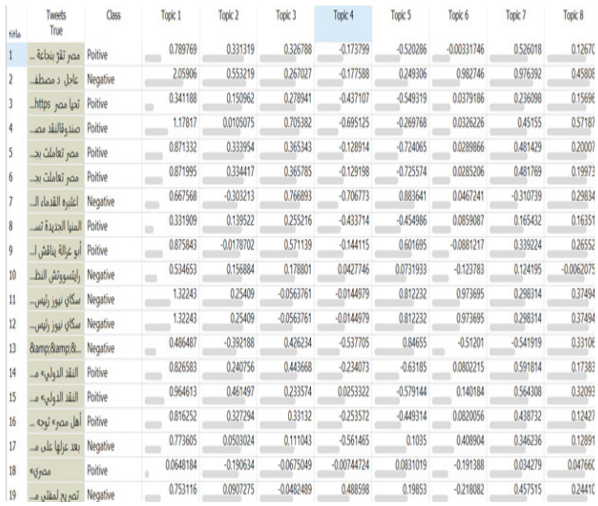

Figure 6: Classifying the Tweet to Negative or Positive Belonging to Each Topic

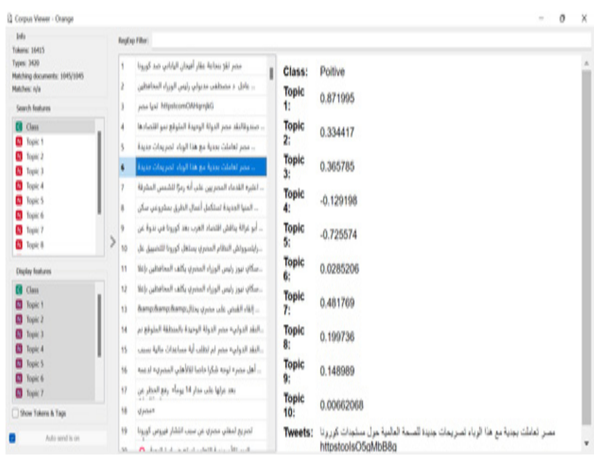

Figure 7: Example of Classifying One Tweet to Negative or Positive Belonging to Each Topic

Figure 7 shows Example of Classifying one tweet to negative or positive belonging to each topic as tweet 6 is belong to class "positive" and belong to topic1 with 0.871995, topic2 with 0.334417, topic3 with 0.365785, and topic4 with -0.129198, topic5 with -0.725574, topic6 with 0.0285206, topic7 with 0.481769, topic8 with 0.199736, topic9 with 0.148989, and topic10 with 0.00662068. Therefore, the tweet6 belong to topic1 and topic7.

4.4 Results of Sentiment Analysis

From the experiment, we classify the emotion on a tweet into one of following emotions - anger, fear, disgust, happiness, sadness, joy and surprise as shown in Figure 8.

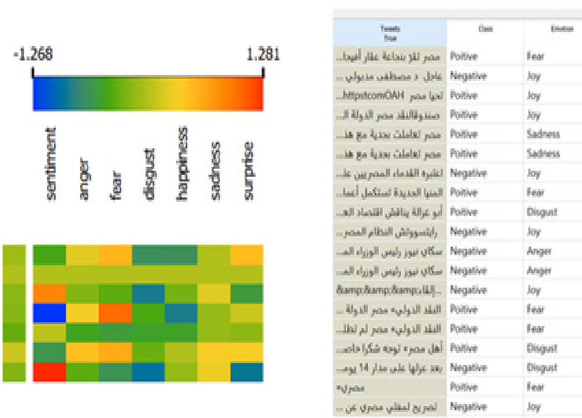

Figure 8: Classifying the Emotion on a Tweet

4.5 Results of Tweets Classification

Based on the tweet text, the sentiment model classify tweets as 1038 negative tweets and 1045 positive tweets as shown in figure 9.

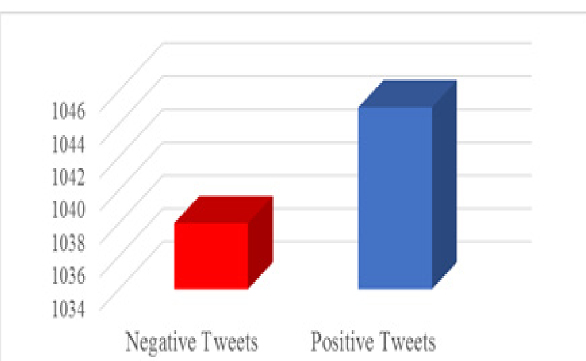

Figure 9: Classifying Tweets as Negative Tweets and Positive Tweets

4.6 Results of Emotions Detection

Then distribute the negative or positive tweet on different emotions as shown in figure 10. For example, the tweet below has negative class with 0.227 anger, 0.52 fear, -0.675 disgust, -0.662 happiness, -0.005 sadness and 0.465 surprise

Class:	Negative
sentiment:	-0.444
anger:	0.227
fear:	0.52
disgust:	-0.675
happiness:	-0.662
sadness:	-0.005
surprise:	0.465
Tweets:	RT 20fourLive عاجل| مصر رئيس الوزراء يكلف المحافظين بإغلاق الحدائق والشواطئ وأماكن التجمعات للحد من انتشار كورونا httpstcoF4s60a...

Figure 10: Distributing the Negative or Positive Tweet on Different Emotions

5. Conclusion

In this research, we present textual analysis of Arabic tweets to detect public emotions in Egypt after two years of the Covid 19 Pandemic. Our model classifies tweets into positive and negative tweets. One of key contributions of this research is developing a method that can label and score Arabic text according to the standard emotions. Another key is analyzing the perception of Egyptians people towards COVID 19, and giving insight into their feeling and reactions. The results showed top terms frequency are - كورونا مصر - فيرس. Sentiment analysis showed number of positive tweets is almost equal to number of negative tweets. In addition, most tweets distributed over different emotions includes anger tweets, fear tweets, happiness tweets, sadness tweets, joy, and surprise tweets. Future research can focus on the availability of lexicons, the usage of Dialect Arabic (DA), the lack of dataset, compound words, and idioms.

6. References

[1]. Ahmed Alsayat, Nouh Elmitwally. (2020). A comprehensive Study for Arabic Sentiment Analysis (Challenges and Applications), Egyptian Informatics Journal, V.21(1), pp. 712-, ISSN 11108665-, https://doi.org/10.1016/j.eij.2019.06.001.

[2]. Alqurashi, Sarah & Alhindi, Ahmad & Alanazi, Eisa. (2020). Large Arabic Twitter Dataset on COVID-19.

[3]. Haouari, Fatima & Hasanain, Maram & Suwaileh, Reem & Elsayed, Tamer. (2020). ArCOV-19: The First Arabic COVID-19 Twitter Dataset with Propagation Networks.

[4]. Alharbi, Basma & Alamro, Hind & Alshehri, Manal & Khayyat, Zuhair & Kalkatawi, Manal & Jaber, Inji & Zhang, Xiangliang. (2020). ASAD: A Twitter-based Benchmark Arabic Sentiment Analysis Dataset.

[5]. Huda Alhazmi and Manal Alharbi. (2020). Emotion Analysis of Arabic Tweets During COVID-19 Pandemic in Saudi Arabia" International Journal of Advanced Computer Science and Applications(IJACSA), V. 11(10). http://dx.doi.org/10.14569/IJACSA.2020.0111077.

[6]. Alhajji, Mohammed & Khalifah, Abdullah & Aljubran, Mohammed & Alkhalifah, Mohammed. (2020). Sentiment Analysis of Tweets in Saudi Arabia Regarding Governmental Preventive Measures to Contain COVID-19. 10.20944/preprints202004.0031.v1.

[7]. Ali, Manal. (2021). Arabic Sentiment Analysis about Online Learning to Mitigate Covid-19. Journal of Intelligent Systems. 30. 524.540-10.1515/jisys-20200115-.

[8]. Amara, Amina & Hadj Taieb, Mohamed Ali & Ben Aouicha, Mohamed. (2021). Multilingual Topic Modeling for Tracking COVID-19 Trends Based on Facebook Data Analysis. Applied Intelligence. 51. 110.1007 .22-/s104893-02033-020-.

[9]. Essam, Ba & Abdo, Muhammad S.(2021). How Do Arab Tweeters Perceive the COVID-19 Pandemic?. Journal of Psycholinguistic Research. 50. 10.1007/s109366-09715-020-.

[10]. Aljameel, Sumayh & Alabbad, Dina & Alzahrani, Norah & Alqarni, Shouq & Alamoudi, Fatimah & Babili, Lana & Aljaafary, Somiah & Alshamrani, Fatima. (2020). A Sentiment Analysis Approach to Predict an Individual's Awareness of the Precautionary Procedures to Prevent COVID-19 Outbreaks in Saudi Arabia. International Journal of Environmental Research and Public Health. 18. 218. 10.3390/ijerph18010218.

[11]. Althagafi, Asma & Althobaiti, Ghofran & Alhakami, Hosam & Alsubait, Tahani. (2021). Arabic Tweets Sentiment Analysis about Online Learning during COVID-19 in Saudi Arabia. International Journal of Advanced Computer Science and Applications. 12. 10.14569/IJACSA.2021.0120373.

[12]. Nandwani, Pansy & Verma, Rupali. (2021). A Review on Sentiment Analysis and Emotion Detection From Text. Social Network Analysis and Mining. 11. 10.1007/s13278--00776.6-00776-021-6.