Analysis of sentiments conveyed through Twitter concerning COVID-19

Mohamed Chiny[1,*], *Marouane* Chihab[1], *Omar* Bencharef[2], *Younes* Chihab[1]

[1]Laboratory of Computer Sciences, Ibn Tofail University, Kenitra, Morocco
[2]Department of Computer Sciences, Cadi Ayyad University, Marrakesh, Morocco

Abstract. Due to the social and economic fallout from the COVID-19 pandemic, we sought to gauge the attitudes of social network users, in this case, Twitter, towards the topic using a sentiment analysis approach. We collected 178,683 tweets using the Twitter API based on queries for the high-frequency hashtag #covid19. After the preprocessing step, we classified them in a binary way (positive and negative) and according to their intensity (valence) using the VADER model and then the NRCLex dictionary, which allows us to classify feelings according to their affective class. The results suggest that overall, the feelings detected through the tweets are positive. In addition, users seem to be interestedin the pandemic as a trend rather than as a topic related to other social or economic aspects.

1 Introduction

The COVID-19 disease, which began in late 2019 [1] and was declared by the World Health Organization as a global pandemic on March 11, 2020 [2], has affected millions of people worldwide. It has affected people's way of life, especially after several countries have taken a series of measures that may impact their travel. They have sometimes instituted total containment for several weeks to months. The pandemic continues to be the focus of debate on all media and especially on social networks.

Social networks have become important channels of communication over the past decades. They allow millions of people to share content, publish their opinions and follow their friends. Social networks are characterized by velocity, volume, value, variety and veracity; the 5 V's of big data [5] and the data they contain can reveal relationship structures between individuals and observe or predict user attributes and interests from their individual behaviour [3,4]. In addition, the rapid growth of online social networks of relationships (Facebook, Myspace, etc.), media sharing networks (YouTube, Instagram, etc.), microblogging (Tumbler, Twitter, etc.) and community platforms (Airbnb, CouchSurfing, etc.), have encouraged researchers to investigate published content and analyze user behaviour [6-8].

Twitter is a virtual social network where people share their posts and opinions about the current situation, such as the COVID-19 pandemic. It is considered the most important data source for machine learning research in terms of analysis, prediction and extraction of knowledge and opinions since many people have used it to express their views and attitudes

towards COVID-19 and share their experiences about this pandemic [9]. Therefore, it is important to analyze and understand the feelings of users towards this scourge.

Sentiment analysis is a subfield of Natural Language Processing (NLP) that attempts to identify and extract opinions from the text. The goal of sentiment analysis is gaugin the feelings and emotions of individuals based on the computer processing of subjectivity in text. It allows companies to make sense of data by being able to automate the process [10]. However, most of these texts are unstructured and come from various sources, in this case, social networks. Micro-blogging content from social networks like Twitter and Facebook poses serious challenges, not only because of the amount of data involved but also because of the type of language used to express sentiments, i.e., abbreviations, slang and emoticons.

In our study, we attempted to understand the attitude of Twitter users towards the COVID-19 pandemic through tweets and retweets posted between July 25 and August 30, 2020. We adopted a sentiment analysis approach using VADER, which is a lexicon and rule-based sentiment analysis tool specifically tailored for social media texts. VADER not only gives a binary score of positivity or negativity but also indicates how positive or negative a sentiment is. We also used the NRCLex dictionary, which classifies sentiments according to their affective class.

2 Literature Review

Because of the complexity of natural language, many specialized applications for sentiment analysis have emerged. Among them are those that classify words in a binary way (i.e. positive or negative) according to their semantic orientation in context, and those in which words are associated with valence scores for sentiment intensity. For example, LIWC is a text analysis tool designed to study the various emotional components present in text samples [14,15]. It uses a dictionary of nearly 4,500 words organized in 76 categories related to sentiment analysis. However, it does not consider lexical elements carrying sentiment such as acronyms or emoticons widely used in social network texts [11].

General Inquirer (GI) is a text analysis application with a manually constructed lexicon that contains more than 11,000 words classified in 183 categories [16]. However, as with LIWC, GI suffers from a lack of coverage of lexical features related to the sentiment conveyed through social text.

Hu-Liu04 consists of a lexicon of nearly 6,800 words. This opinion lexicon was initially constructed through a bootstrapping process using WordNet, where words are grouped into synonym clusters called synsets [17]. Although Hu-Liu04 is suitable for sentiment expressions in social texts and product reviews, it does not capture the sentiments of emoticons or acronyms.

The ANEW lexicon [18] provides a set of emotional ratings for 1,034 words. Unlike LIWC or GI, the words in ANEW were ranked according to their affective class. ANEW words have an associated sentimental valence ranging from 1 to 9. Words with valence scores below five are considered negative, and those with scores above five are considered positive. Nevertheless, as with LIWC and GI, the ANEW lexicon is also insensitive to common lexical features related to sentiment in social text.

The specific nature of social media content poses severe challenges to applications of sentiment analysis due to its vast bias and big data nature [10, 11]. For this reason, the VADER model seems to be more suitable for this situation as it has proven to be very effective when dealing with social network texts, editorials, movie or product reviews, mainly since it not only generates a positivity and negativity score but also tells us how positive or negative a sentiment is.

VADER is based on the wisdom of the crowds (WoC) approach to acquire a valid point estimate of the sentiment valence (intensity) of each lexical feature. The evaluation was

conducted by ten independent human raters (for a total of more than 90,000 ratings), leading to the adoption of 7,500 lexical features with valence scores that indicate the polarity and intensity of the sentiment on a scale of -4 (extremely negative) to +4 (extremely positive) [11].

VADER lexicon revealed its effectiveness in many studies that tried to analyze the feelings through the text of microblogging (in this case, Twitter). For example in [23], the authors attempted to understand the feelings of users towards Bitcoin during the COVID-19 period, or in [24], where the authors conducted a study on the mental health of users in the US during the COVID-19 pandemic by analyzing the feelings conveyed by their tweets.

3 Methodology

3.1 Data collection

The data for this study was collected using the Twitter API and a query-based Python script for the high-frequency hashtag #covid19. 178,683 sample tweets were scraped between July 25 and August 30, 2020. The retrieved data includes user and location information, among other things.

3.2 Preprocessing, filtering and cleaning of data

Data cleaning is the first step in the dataset preparation process. This operation consists of detecting and correcting corrupt, inaccurate or incomplete records in a dataset to produce good quality data for the rest of the processing. We found that a small portion of the records had missing fields such as user description and hashtags. Since this information does not affect the relevance of our study, we retained the 178,683 samples initially collected. We also removed non-alphanumeric characters and stop words from the hashtags for a better presentation of the word cloud.

Concerning the geographical location, in some cases the collected data do not always indicate the country, but sometimes a state or a city, as is the case of the USA or India. If we take into account the records where the name of the country is indicated, then it turns out that the most significant number of tweets comes from the USA (17,113 tweets), followed by India (12,342 tweets) and then the United Kingdom (4,822 tweets).

3.3 Sentiment analysis with VADER

We applied the VADER model to the tweets and retweets collected using the Python language in the Jupyter Notebook environment, integrating it as a module of the NLTK software library that contains pre-trained natural language processing (NLP) models.

Since VADER returns a score that reflects the intensity of feeling between -4 and 4, the results obtained would not give a clear idea of the users' attitude in a precise way. Indeed, it would be more like binary classification if we defined a threshold to delimit the feelings. So, we also used NRCLex, which measures emotion from the text. The NRCLex dictionary contains about 27,000 words and is based on the National Research Council (NRC) lexicon of affect and WordNet synonym sets from the NLTK library.

4 Results

4.1 Exploratory Data Analysis (EDA)

During the period between July 25 and August 30, 2020, when the number of COVID-19 cases is increasing, Twitter users from some countries are more active than others in posting tweets and retweets that mention the virus. However, although the number of active users on Twitter regarding COVID19 is higher in India, the data cleaning and filtering phase revealed that the USA is ranked first regarding the number of tweets and retweets posted (17,113 tweets).

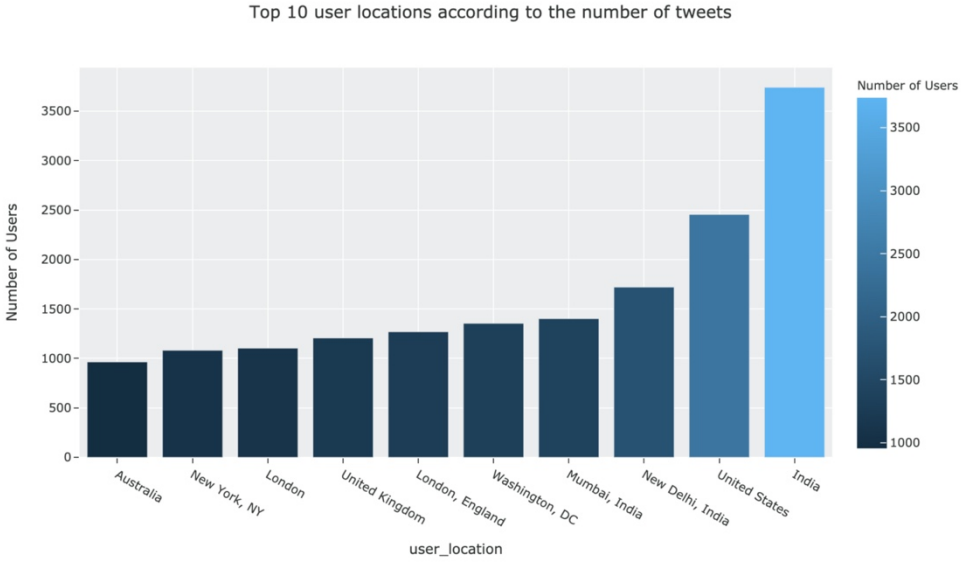

Fig. 1. Top 10 Twitter users who mentioned COVID-19 in their tweets and retweets

What draws attention is the fact that China is absent from this TOP 10 (Fig. 1), despite the fact that it has received a lot of media attention regarding COVID-19 due to the registration of the first positive case of the virus in Wuhan [12], in addition to a population estimated to be around 1.4 billion inhabitants [13] and which is logically supposed to host many active users on social networks. However, this could be partly explained by the deletion operation carried out by Twitter in June 2020, which targeted more than 170,000 accounts coordinating to spread disinformation and propaganda campaigns concerning, among others, the COVID-19 epidemic in China [19].

It would be interesting to get an idea of the most popular pandemic-related hashtags of this period using a word cloud. The word cloud, which is an illustration of the words present in a corpus and whose size changes according to their frequency of appearance. Fig. 2 shows the most used hashtags during the period of the collection of tweets. It clearly indicates that the hashtags that refer to the virus identifier such as 'Covid-19' or 'Coronavirus' are the most dominant, while the hashtags that associate COVID-19 with other words such as 'Health', 'Travel', 'Economy', 'Obesity', 'Technology'... are not as popular.

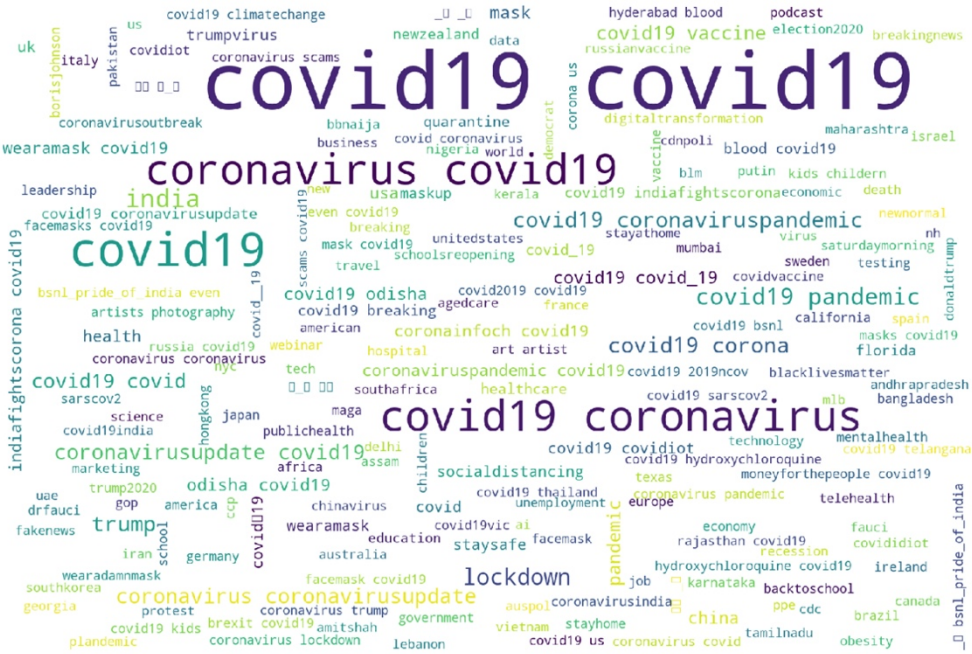

Fig. 2. Word cloud of the most popular Covid19-related hashtags on Twitter.

It should be noted that the word cloud is often a preliminary step in the analysis of a textual corpus using algorithms derived from Natural Language Processing. However, words illustrated in large size do not necessarily mean that they are semantically more important than other words. In fact, algorithms such as TF-IDF suggest that the more frequent a word is, the more its importance is reduced within the same document [20].

4.2 Sentiment analysis about the pandemic on Twitter

First, we wanted to have an overall idea of the common feeling conveyed through the analyzed tweets. We, therefore, applied the VADER algorithm by measuring the intensity of the feeling experienced on a scale from -4 to +4 and compared it to the 0 thresholds (neutral). The result highlights a binary polarity that ranks the tweets according to the positivity or negativity of the feelings.

Fig. 3 shows that 128,370 tweets convey positive feelings against 50,738 that convey negative feelings. It is therefore clear that with regard to this global pandemic, most Twitter users express a positive or an optimistic feeling.

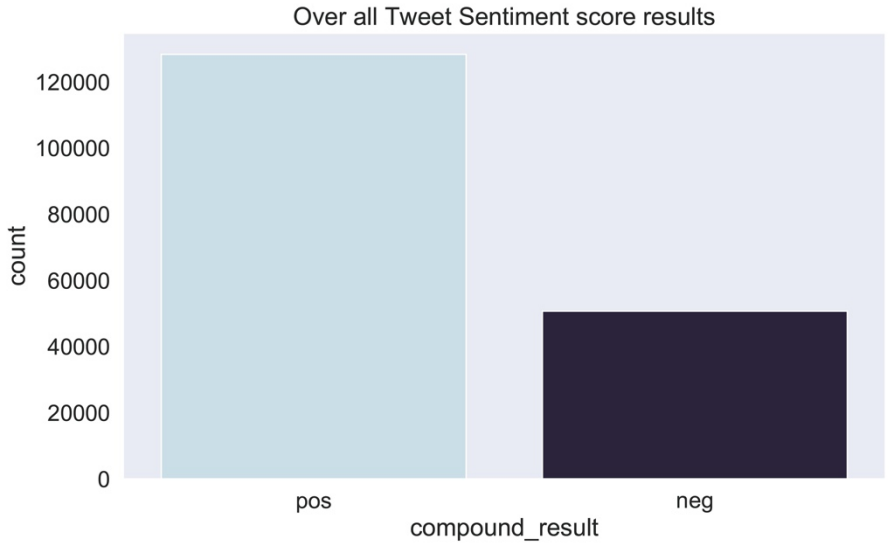

Fig. 3. Ranking of feelings about the pandemic in a binary way using VADER

It should be noted that the power of VADER lies in its ability to detect the intensity of the feeling, i.e. how positive or negative it is. However, to represent such a result, it would be better to split the rating interval between -4 and +4 into several segments. In our case, we split it into two segments to show the overall impression of the pandemic. However, we thought of segmenting the feelings according to their affective class. We, therefore, opted to use the NRCLex dictionary, which divides the texts according to the following feelings: Positive, Negative, Trust, Fear, Anticipation, Sadness, Joy, Anger, Surprise and Disgust. A tweet can be classified into several sentiments at the same time.

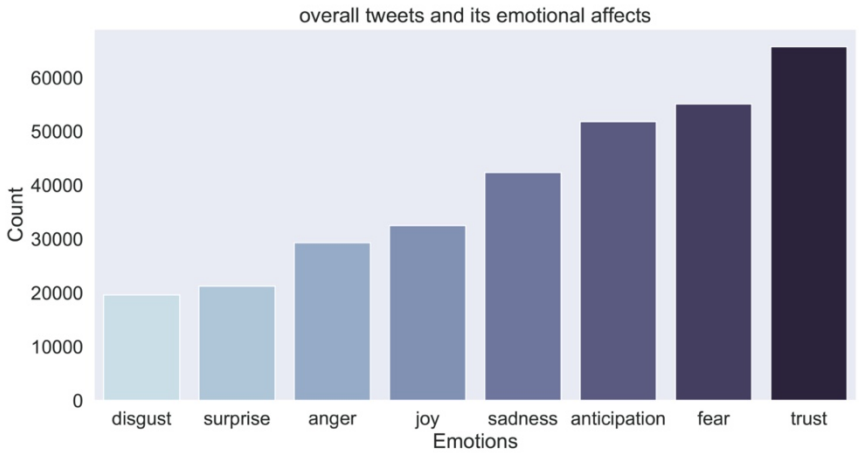

Fig. 4. Distribution of tweets according to their affective class using the NRCLex dictionary

Table 1. Number of tweets according to their affective class using the NRCLex dictionary

Emotions	Count
Positive	106827

Negative	81543
Trust	65753
Fear	55111
Anticipation	51838
Sadness	42377
Joy	32512
Anger	29312
Surprise	21256
Disgust	19643

Table 1 and Fig. 4 illustrate the segmentation of tweets according to their emotional classes. The feeling of trust seems to be the most prevalent in the tweets during the collection period, followed by fear and anticipation. Overall, the feelings classified as positive are the most dominant, as illustrated in Fig. 3 and Fig. 4. Nevertheless, the classification of the feelings according to the affections experienced allows us to have a more precise idea of their intensity and their nature in the tweets.

5 Discussion

The brutal shock produced by the COVID-19 and by the measures to stop certain economic activities that have been instituted by many states to curb it is plunging the world economy into a severe recession. According to World Bank forecasts, which were announced at the same period of data collection of this study, the world GDP will decrease by 5.2% before the end of 2020. These forecasts herald the deepest recession the world has seen since World War II [21].

However, according to the results of our study, it can be estimated that the overall impression of COVID-19 is not as dramatic as one might think. Considering the results of sentiment analysis by calculating their intensity or their segmentation according to affective class, it can be clearly seen that the invocation of the word COVID-19 by Twitter users is seen as a trend rather than a topic of discussion that covers social and economic aspects. Indeed, the Word cloud in Fig. 2 clearly illustrates that the most popular hashtags include the name of the virus alone rather than associating it with other words that highlight its impact on different domains, in this case, social and economic.

In [22], the authors applied the TextBlob library on tweets collected in the first half of April 2020. Their results show that more than 60% of the tweets convey a calm and relieved feeling, compared to less than 5% of tweets that convey a feeling of fear and worry. Overall, these results seem to be consistent with those we found in our study.

6 Conclusion

The COVID-19 pandemic has affected millions of people worldwide and has had a significant impact on many social and economic areas. We, therefore, sought to identify the feelings of users across the social network Twitter towards this scourge by adopting a sentiment analysis approach. We started by collecting 178,683 tweets using the Twitter API. After sentences preprocessing, we classified them in a binary way (positive and negative), then, according to their intensity (valence) using the VADER model. Since the latter scores the sentiment on a scale of -4 to 4, we had difficulties to represent the results in a clear way, so we used the NRCLex dictionary, which allows classifying the sentiments according to their affective class.

The results found show that the overall feeling experienced by the users is relatively positive. Moreover, although the topic is a trend on Twitter (and on other social networks for

that matter), we found that the most popular hashtags do not associate the words 'Covid-19' and 'Coronavirus' with other social or economic aspects. This supports the idea that most users feel positive about the pandemic and are not too concerned about the impact on other parts of society.

It should be noted that the geographical distribution of Twitter users is not precise enough in our study, as the data collected sometimes reports the country and sometimes the state or city. Therefore, It would be helpful to aggregate these data with a dataset that includes all the states and cities of the world. However, this study was able to highlight the global attitude of social network users (in this case Twitter) towards the most publicized phenomenon at the moment, COVID-19. However, since the data of this study goes back to August 2020, it would be interesting to gauge the sentiment of users in early 2021, especially with the implementation of large-scale vaccination campaigns against the Coronavirus around the world.

References

1. H. Wang, Z. Wang, Y. Dong et al, Phase-adjusted estimation of the number of coronavirus disease 2019 cases in Wuhan, China, Cell Discovery, **vol. 6** (2020).
2. W. H. Organization, Coronavirus, https://www.who. int/health-topics/coronavirus#tab=tab_1 (2020).
3. Michal Kosinski, David Stillwell, and Thore Graepel, Private traits and attributes are predictable from digital records of human behavior, Proceeding of the National Academy of Sciences of The United States of America, **110** (2013).
4. Kahr, M., Leitner, M., Ruthmair, M., Sinnl, M. Benders decomposition for competitive influence maximization in (social) networks. Omega, 102264. doi:10.1016/j.omega.2020.102264 (2020).
5. Bazzaz Abkenar, S., Haghi Kashani, M., Mahdipour, E., & Mahdi Jameii, S. Big data analytics meets social media: A systematic review of techniques, open issues, and future directions. Telematics and Informatics, 101517. doi:10.1016/j.tele.2020.101517 (2020).
6. Y. Feng, P. Zhou, D. Wu, and Y. Hu, Accurate Content Push for Content-Centric Social Networks: A Big Data Support Online Learning Approach, IEEE Transactions on Emerging Topics in Computational Intelligence, **no. 99** (2018).
7. J. Heidemann, M. Klier, and F. Probst, Online social networks: A survey of a global phenomenon, Computer networks, **vol. 56** (2012).
8. Mohamed Chiny, Omar Bencharef, Moulay Youssef Hadi, Younes Chihab, A Client-Centric Evaluation System to Evaluate Guest's Satisfaction on Airbnb Using Machine Learning and NLP, Applied Computational Intelligence and Soft Computing (2021).
9. Zhang, X., Saleh, H., Younis, E. M. G., Sahal, R., & Ali, A. A. Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System. Complexity, 2020, 1–10. doi:10.1155/2020/6688912 (2020)
10. Parul Pandey, Simplifying Sentiment Analysis using VADER in Python (on Social Media Text), https://medium.com/analytics-vidhya/simplifying-social-media-sentiment-analysis-using-vader-in-python-f9e6ec6fc52f (2018).
11. C.J. Hutto, Eric Gilbert, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text, Conference: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media At: Ann Arbor, MI (2015).

12. World Health Organization, Archived: WHO Timeline - COVID-19, https://www.who.int/ news/item/27-04-2020-who- timeline---covid-19, (April 2020).
13. World Bank, World Development Indicators, https://datatopics.worldbank.org/world-development-indicators/ (2018).
14. Pennebaker, J. W., Francis, M., & Booth, R. Linguistic Inquiry and Word Count: LIWC 2001. Mahwah, NJ: Erlbaum (2001).
15. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. The development and psychometric proper- ties of LIWC2007. Austin, TX: LIWC net (2007).
16. Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. General Inquirer. Cambridge, MA: MIT Press (1966).
17. Hu, M., & Liu, B. Mining and summarizing customer reviews. In Proc. SIGKDD KDM-04 (2004).
18. Bradley, M. M., & Lang, P. J. Affective norms for English words (ANEW): Instruction manual and affective ratings (1999).
19. Le Monde avec Reuters , Twitter supprime 170 000 comptes diffusant des messages favorables à la Chine, https://www.lemonde.fr/pixels/article/2020/06/12/twitter-supprime-170-000-comptes-diffusant-des-messages-favorables-a-la-chine_6042620_4408996.html (June 2020).
20. Shengqi Wu, Huaizhen Kou, Chao Lv, Wanli Huang, Lianyong Qi, Hao Wang, Service Recommendation with High Accuracy and Diversity, Wireless Communications and Mobile Computing (2020).
21. World Bank, COVID-19 to Plunge Global Economy into Worst Recession since World War II, https://www.worldbank.org/en/news/press-release/2020/06/08/covid-19-to-plunge-global-economy-into-worst-recession-since-world-war-ii (June 2020).
22. Kamaran H. Manguri, Rebaz N.Ramadhan, Pshko R. Mohammed Amin, Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks, Kurdistan Journal of Applied Research (May 2020).
23. Toni Pano, asha Kashef, A Complete VADER-Based Sentiment Analysis of Bitcoin (BTC) Tweets during the Era of COVID-19, Big Data and Cognitive Computing, **Vol. 4(4)** (2020).
24. Valdez D, ten Thij M, Bathina K, Rutter LA, Bollen J, Social Media Insights Into US Mental Health During the COVID-19 Pandemic: Longitudinal Analysis of Twitter Data, J Med Internet Res, **Vol 22(12)** (2020).