Kurdistan Journal of Applied Research (KJAR)
Print-ISSN: 2411-7684 | Electronic-ISSN: 2411-7706

Website: Kjar.spu.edu.iq | Email: kjar@spu.edu.iq

Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks

Kamaran H. Manguri
Department of Computer Science
College of Basic Education
University of Raparin
Rania, Iraq
kamaran@uor.edu.krd

Rebaz N. Ramadhan
Software Engineering Department
Faculty of Engineering
Koya University
Koya, Iraq
rebaz.najeeb@koyauniversity.org

Pshko R. Mohammed Amin
Department of Computer Science
College of Basic Education
University of Raparin
Rania, Iraq
pshko.amin@uor.edu.krd

Article Info

Special Issue on Coronavirus (COVID-19)

DOI:
10.24017/covid.8

Article history:

Received: 22 April 2020
Accepted: 15 May 2020

Keywords:

Sentiment analysis, COVID-19, Coronavirus, Social Media, Twitter, Python, Text Blob.

ABSTRACT

In the past two decades, the growth of social data on the web has rapidly increased. This leads to researchers to access the data and information for many academic research and commercial uses. Social data on the web contains many real life events that occurred in daily life, today the global COVID-19 disease is spread worldwide. Many individuals including media organizations and government agencies are presenting the latest news and opinions regarding the coronavirus. In this study, the twitter data has been pulled out from Twitter social media, through python programming language, using Tweepy library, then by using TextBlob library in python the sentiment analysis operation has been done. After the measuring sentiment analysis, the graphical representation has been provided on the data. The data we have collected on twitter are based on two specified hashtag keywords, which are ("COVID-19, coronavirus"). The date of searching data is seven days from 09-04-2020 to 15-04-2020. In the end a visualized presentation regarding the results and further explanation are provided.

1. INTRODUCTION

Sentiment analysis nowadays can be considered as one of the most popular research topics in the field of natural language processing. The uses of sentiment analysis are covered by some interesting scientific and commercial areas, such as opinion mining, recommender systems,

and event detection[1]. Nowadays the social media platforms such as Twitter, Facebook and YouTube, are a great source of information known as social data [2]. The events occurring in normal daily life are discussed on social media and any individuals are free to discuss and express their opinion about these events. Coronavirus known as COVID-19 which began to appear at the end of last year in Wuhan, China has been one of the most discussed and one of the most spreading diseases worldwide.

According to World Health Organization (WHO) until we are writing this study, more than 20000000 individuals have contracted the disease and more than 157000 individuals have died worldwide (WHO | World Health Organization, 2020/4/20). From this data, we can see that this is one of the most biological virus outbreaks in the last two decades in the century. From many researchers and health professionals analyzing and getting new insights from these data on the social media platforms could help health staff and government organizations to get benefit from those data, and understand from people's reaction and express their feelings [3].

In this study, tweets have been taken from Tweeter's social data, known as tweets according two specified search keyword which are #coronavirus and #COVID-19 to extract tweets and perform sentiment analysis on the datasets and Python programming language has been chosen. Python provides many easy to use libraries to access Twitter social media platforms. Python can access these tweets from Twitter's search API and tweepy library. In summary the sentiment analysis approach has been applied to these data we have collected, and a detailed explanation has been conducted.

2. RELATED WORK

Many researchers are working on Sentiment Analysis on Twitter social media in literature, some key contributions which are providing support for finding user behaviors and situations in the different cases while happening around the world, for this some of the essential papers are included in this section.

Kaur and Sharma[4] analyze the sentiments regarding coronavirus disease (COVID-19), so analyze the sentiments of different people's opinion for this disease. For this purpose, twitter API used for collecting related tweets to the coronavirus, then positive, negative and neutral emotion analyzed by using machine learning approaches and tools. In addition, for pre-processing of fetched tweets NLTK library is used and Textblob dataset for analyzing tweets is used, after that show the interesting results in positive, negative, neutral sentiments through different visualizations.

Prabhakar Kaila et al.[5], informs that the collected data raw was suitable and worthy to be applied to the experiments, regarding to the novel COVID-2019 out breaks. The Latent Dirichlet Allocation (LDA) has been applied to the collected data in the document term matrix from the datasets. The LDA techniques discovered during the operation that the huge related information about the COVID-19 infection paramedic were negative opinions like fear and positive sentiments such as trust.

Medford et al.[6], constructed a list of hashtags related to COVID-19 to search for relevant tweets during a two-week interval from January 14th to 28th, 2020. Tweets are extracted by API and stored as plain text. The frequency associated keywords are identified and analyzed such as infection prevention practices, vaccination, and racial prejudice. After that sentiment analysis is performed to identify each tweet's emotional valence (positive, negative, or neutral) and predominant emotion (anger, disgust, fear, joy, sadness, or surprise). Finally, tweets are identified and analyzed related topics over time by using an unsupervised machine learning method.

Alhajji et al.[7], are performing sentiment analysis runs on Arabic tweets by using Naïve Bayes machine learning model through the Natural Language Toolkit (NLTK) library in Python. Tweets containing hashtags pertaining to seven public health measures imposed by the government were collected and analyzed. Total tweets

were analyzed in this study are 53,127. The results show more positive tweets than negative based on measures, except one.

Cherish Kay Pastor [8] reveal the sentiment of the Filipinos in the effect of extreme community quarantine caused by COVID-19 Pandemic particularly Luzon. The researcher also analyzes the effect of extreme community quarantine and other effects of the Pandemic to personal lifestyle based on the tweets of the users. Natural Language Processing methodology is use to determine the sentiment of users from extracted tweets. Opinions are treated as data for analyzation. A qualitative approach was also used in determining the effects of the extreme community quarantine in the Luzon area.

Rajput, N. K and et al. [9] in their work, not only collected data related to the tweets published during January 2020 but also investigated the tweets corresponding to two main aspects: first comprehending the word occurrence pattern and accordingly the second sentiment recognition. In addition, he stated that the number of ids tweeting regarding coronavirus subject has continuously spiked especially during February as well as March. There were so many repeated words such as COVID-19, Coronavirus, Wuhan city. On the top of that N-grams model like unigram, bigram and trigram repetition were made for the top thousand frequencies. Rajput concluded that almost everybody followed some fluctuation rule which is quite natural distribution. The figures and outcomes were found quite satisfactory. The exponential parameters were influenced for the identical by -1.273 for unigram, -1.375 for bigram and -0.5266 for trigram. The plots for the identical are illustrated. SSE, R2 and RMSE decided on goodness of suited the model.

Ra, M. and et al. [10], are analyzed and visualized the influence of coronavirus (COVID-19) within the earth by executing such algorithms and methods of machine learning in sentiment analysis on the tweet dataset to understand very positive and really negative opinions of the ultimate public round the world. This reveals that Naive Bayes 'machine learning approach has been produced better execution, and it's been thought to be the concept for basic learning. This also brings out another ensemble technique that uses sentiment score because the input function for the classifiers in machine learning, SVM, Max Entropy, Decision Tree, Boosting, and Random Forest. As a result, the LogitBoost, a blended approach, performed better with accuracy of 74%.

AD Dubey, A. D. [11], collected and analyzed tweets from twelve states in this study. The tweets are collected between 11th march to 31st march 2020, Moreover, all tweets are belonging to the novel COVID-19 disease. The aim of this analysis is to know how individuals in those countries reacting to the outbreaks of the disease. There is no doubt that, there some necessary steps required while we are collected and performing the operation, such as pre-processing and removing irrelevant information from the tweets. The outcomes from these experiments shows that most of the people from these societies are thinking positive and they are feeling good that the situation will goes to be better, also it is worthy that there are also signs of fear and sadness. However, four states especially from the Europe continent thinking that, they cannot trust the situation because of the outbreaks and pandemic over large scale of populations.

3.DATA AND METHODOLOGY

In this section, the process of data collection by Tweepy python library has been described. We have also shown the evaluation procedure of sentiment analysis. Finally, a data analysing approach by Textblob python library explained in detail.

3.1. Dataset

In this paper, Tweepy python library has been utilized for data extraction from Twitter API (Application programming interface). Moreover, Tweepy allows appropriate data retrieval by searching via keywords, hashtags, timelines, trends, or geo-location [12]. However, this research has not targeted a particular continent, country, or city for data collection, because coronavirus is an almost ubiquitous health problem. In spite of having numerous restrictions from Twitter API, we have applied successive attempts to access as many posts as possible. All around the world, approximately 500,000 tweets have been fetched regarding COVID-19 and CORONAVIRUS keywords from 09-04-2020 to 15-04-2020 from tweeter. In order to avoid redundancy, a specific timeframe (last 24 hours) has been considered to pull out the tweets during a week. The gathered data has been stored in CSV format, and fed to the Sentiment Analysis library, namely, Textblob. The following tables show the retrieved data from Twitter API.

Table1: Twitter Data About Coronavirus During One Week

Date	No. of tweets for Coronavirus keyword	No. of tweets for Covid-19 keyword	Subtotal tweets
Thu 09-04- 2020	44357	49183	93540
Fri 10-04-2020	71053	80781	151834
Sat 11-04-2020	16399	26231	42630
Sun 12-04-2020	28391	26761	55152
Mon 13-04-2020	31775	26747	58522
Tue 14-04-2020	22389	35987	58376
Wed 15-04-2020	34229	35949	70178
Total tweets per week			530232

3.2 data collection implementation and procedure

In this section the whole data collection and sentiment analysis procedure has been explained in advance. As it mentioned the social media platform that has been chosen to collect data sets was Twitter. To explain the entire work step by step we have to describe and express each stage as follow.

Firstly, the essential step is to establish the connection between Python and Twitter Microblog. Twitter provides public API's throughout URLs to access their data. Python has tweepy library to access Twitter's data regarding Twitter's API. The first thing in implementation is to calling required libraries, such as Tweepy, and Text Blob. TextBlob is one of the Python's library to perform Sentiment Analysis.

The data we have collected from Twitter were the tweets which were mostly alphabetic character, but nowadays users also include emotional signs such as laughing and sad, and even any emoji's to express their feelings. The data collection was performed during one week, and each day's data has been kept into different CSV files. The information that has been targeted was content and also the timestamps of the tweets. The main operation is when the tweets were retrieved from Twitter, the tweets were sent to a method, which perform the sentiment analysis, by using Python's TextBlob library. Moreover, along with each tweet their sentiment state has been recorded. With every single request to the API call, massive number of tweets are being collected. there is also worthy to say that, the searches were based on two keywords which are (#COVID-19, and #coronavirus). This definitely leaded to gaining a larger number of tweets. eventually, a suitable recording format has been chosen to the data.

Finally, after collecting the tweets and conducting the sentiment analysis action, the information's were sent to the next step which were the results and explanation by doing different experiments.

3.3. Sentiment Analysis Procedure and Algorithm

Since Sentiment Analysis (SA) is a hot topic in the Natural Language Processing (NLP) field, there are several models to determine the state of sentiment (Positive or negative emotion) within text, paragraph or the whole document [13]. Sentiment Analysis has a certain procedure that begins with grabbing the collected data, then identifying the data. Later on, the required features will be extracted to the next step which is sentiment classification. Finally, decision will be conducted in the phase of sentiment polarity as well as subjectivity.

Figure 1: Sentiment analysis procedure

In general, Sentiment analysis uses machine learning and lexicon-based approaches for investigating emotion inside a piece of text [14]. According to Textblob documentation [15], Textblob takes advantage of Naïve Bayes (NB) model for classification (look at figure 2). NB classifier has been trained on NLTK (Natural Language ToolKit) to detect valence of aggregated tweets [Ibid].

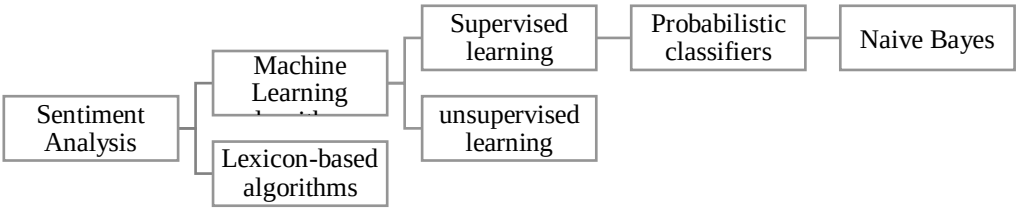

Figure 2: Naïve Bayes Hierarchy

NB is a probabilistic algorithm that uses Bayes theorem to compute sentiment distribution over the data. However, NB dissects any text to a bag of words which means the positions of the words are completely disregarded [14]. The Bayes equation to predict the sentiment probability is:

$$P(label|features) = \frac{P(label) * P(features|label)}{P(features)}$$

Where P(label) is the prior probability of a label, P(features|label) is the prior probability that a given feature set is being classified as a label , and P(features) is the prior probability that a given feature set is occurring [14, 16].

4. RESULTS AND DISCUSSION

In this section, Positive versus Negative versus Neutral for sentiment polarity and objective versus subjective versus Neutral for subjectivity results are shown and discussed.

4.1. Sentiment Polarity

In this study, the bar chart demonstrates sentiment polarity over Twitter for 7 consecutive days starting in 9th April 2020. The data has been shown by two keywords, namely, Coronavirus and Covid-19. Units are measured in percentage. It is clear the total number of tweets are 530K. Overall, more than 36% of people published optimistic views, while only around 14% of the tweets were negative. However, the neutral toll regarding both coronavirus and covid-19 keywords was significantly high (50%).

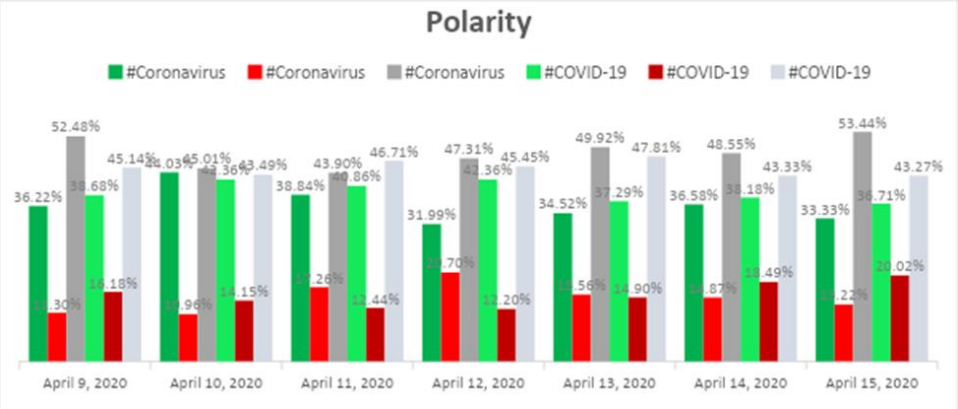

Figure 3: Sentiment Polarity

In 9th April, 52% of #Coronavirus and 45% of #COVID-19 tweets were neutral. Moreover, 11% of #Coronavirus and 16% of #COVID-19 had low emotion. On the other hand, 36% of #Coronavirus and 38% of #COVID-19 were positive towards a coronavirus outbreak. Last day figures (15th April) shows 53% and 43% were neutral for both #Coronavirus and #COVID-19 respectively. There was slight drop in positivity which were 33% for #Coronavirus and 36% of #COVID-19. The figures of different days are relatively similar. Therefore, there is no significant sharp fluctuation in the tolls. Finally, including a large quantity of neutral tweets indicates that most of the corpus were facts rather than opinion. In this regard, the Next section will focus on subjectivity of the data in detail.

4.2. Subjectivity

In terms of the perspective of tweeters, there are subjective and objective viewpoints. There are seven pie charts for each of #Coronavirus as well as #COVID-19. It can be observed that their pie chart portions are very similar, as a result of subtle variance in #Coronavirus and #COVID-19 data. Overall, for both keywords, the large portion of the records were objective which was approximately 64%. Meanwhile it seems that about 22% were being subjective of which expressed their feelings and opinions. Lastly, 14% of the tweets have no clear characteristic to be neither subjective nor objective.

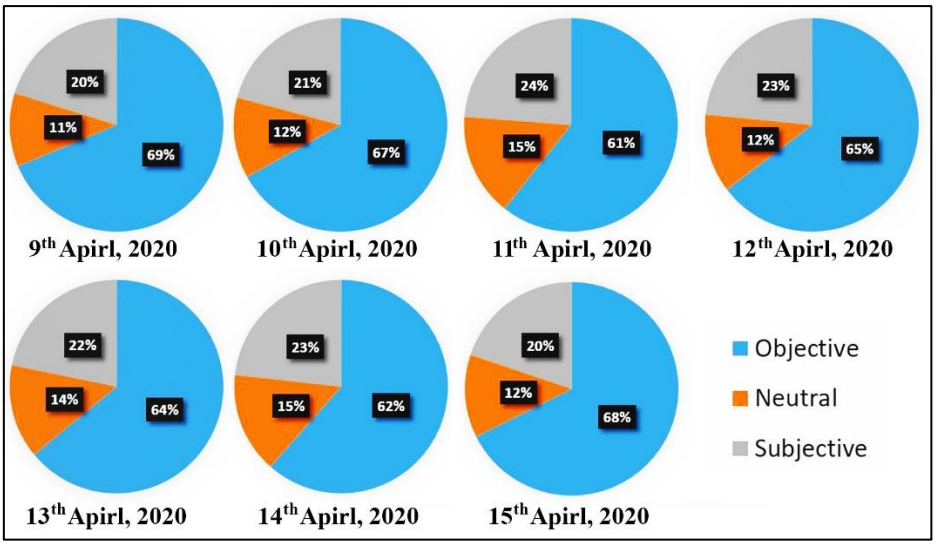

Figure 4: Subjectivity of Coronavirus keyword

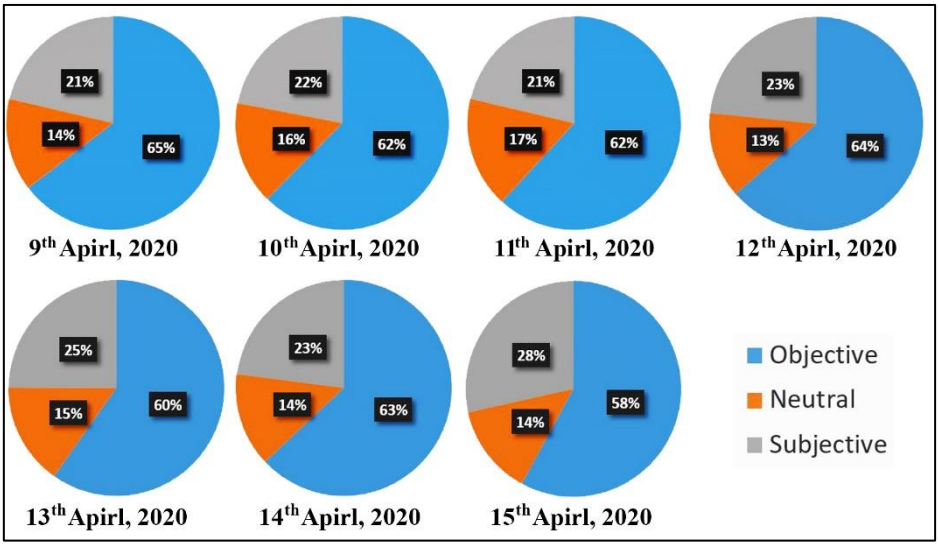

Figure 5: Subjectivity of COVID-19 keyword

4.3. Discussion of sentiments

The aim of this analysis is to identify the emotional state of people about coronavirus. As it can be interpreted in the chart, the majority of the reactions, nearly 60%, toward this health problem were smooth and relaxed as shown in figure 6. In addition, approximately 13% and 7% of the tweeters involved the feeling of content and hopeful respectively. On the contrary, 7% of the contributors experienced relieved mood. Additionally, minority of the collected data expressed their feeling as confident, happy, worried or frightened. This minority only comprised 2% of the cluster. It is worth to mention almost no records showed discouraged and difficulty point of view. In last, this study exposed the fact that members of tweeters were more optimistic when they use COVID-19 than Coronavirus word in their tweets.

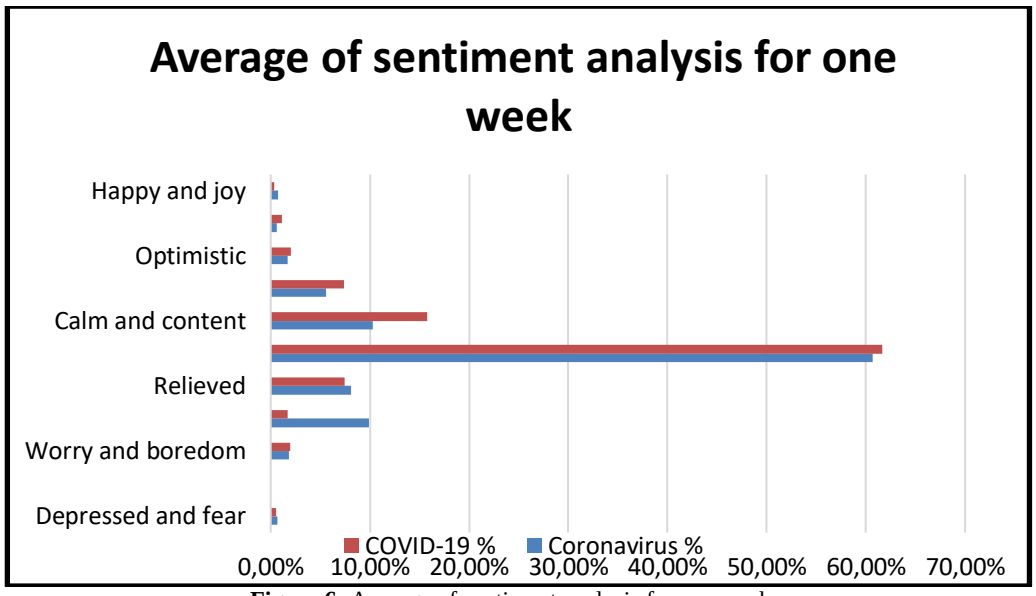

Figure 6: Average of **s**entiment analysis for one week

In order to have a clear perception of the above figures, the following table shows emotion analyzing technique in this research. The bellow table (Table 2) has been applied as the "Emotional Guidance Scale" for polarity evaluation [17]. The higher emotion on the scale which is happy and joy and its value is 1 denotes better-feeling, while the lower on the scale which is and its value, is -1 depressed and fear denotes more negative emotional states, in addition to the middle emotional state with 0 value represents neutral and relaxed state. Also, the rest of the feelings cab be distributed evenly as spectrum between these three mentioned scales.

Table2: Emotional Guidance Scale

Scale	Emotion
1	Happy and joy
0.8	Confident
0.6	Optimistic
0.4	Hopeful
0.2	Calm and content
0	Neutral and relaxed
-0.2	Relieved
-0.4	Pessimistic and impatient
-0.6	Worry and boredom
-0.8	Discouraged and difficulty
-1	Depressed and fear

5. CONCLUSION

5.1. Conclusion

The present study applied sentiment analysis on twitter data related to worldwide COVID-19 outbreaks. The data has been collected during one of the most spread weeks of coronavirus

which is 09-4-2020 to 15-04-2020 (According to https://coronavirus.jhu.edu/map.html website) by using Twitter API and tweepy library of python. Two keywords are selected for searching tweets which are #coronavirus and #COVID-19 for determining polarity and subjectivity. TextBlob library of python Sentiment Analysis techniques applied for collected tweets which are 530232 tweets. The results shown the neutral toll regarding both coronavirus and covid-19 keywords for polarity was significantly high which is more than 50 percent and the large portion of the records were objective which was approximately 64 percent. From this study we can say that people's reactions vary day to day from posting their feelings on social media specifically Twitter. These data regarding the outbreak of Coronavirus known as COVID-19 shows us how people, government organizations and media agencies broadcast the situations.

5.2. Future work

Due to lack of time, and computational process, many aspects has been left for the future works. It would be interesting to take the following research area into consideration:

1. Since there are so many professional and official people on Twitter, you may find more reliable source of information on Twitter than other social medias such as Facebook, Wechat , Instagram. However, it is very essential to explore other social media with regard to sentiment analysis.
2. In our contribution, we chose Textblob which is using Naïve Bayes model. But there are other models that may provide interesting results such as lexicon-based algorithms.
3. This research application is not only convenient for Coronavirus health issue but it can also be adopted as model to discover sentiment emotion for the future similar cases.

REFERENCE

[1] K. Sailunaz and R. J. J. o. C. S. Alhajj, "Emotion and sentiment analysis from Twitter text," vol. 36, p. 101003, 2019.

[2] P. Tyagi and R. J. A. a. S. Tripathi, "A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data," 2019.

[3] A. Alsaeedi and M. Z. J. I. Khan, "A Study on Sentiment Analysis Techniques of Twitter Data," vol. 10, no. 2, 2019.

[4] C. Kaur and A. Sharma, "Twitter Sentiment Analysis on Coronavirus using Textblob," EasyChair2516-2314, 2020.

[5] D. Prabhakar Kaila, D. A. J. I. J. o. A. R. i. E. Prasad, and Technology, "Informational Flow on Twitter–Corona Virus Outbreak–Topic Modelling Approach," vol. 11, no. 3, 2020.

[6] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. J. m. Lehmann, "An" Infodemic": Leveraging High-Volume Twitter Data to Understand Public Sentiment for the COVID-19 Outbreak," 2020.

[7] M. Alhajji, A. Al Khalifah, M. Aljubran, and M. Alkhalifah, "Sentiment Analysis of Tweets in Saudi Arabia Regarding Governmental Preventive Measures to Contain COVID-19," 2020.

[8] C. K. J. A. a. S. Pastor, "Sentiment Analysis of Filipinos and Effects of Extreme Community Quarantine Due to Coronavirus (COVID-19) Pandemic," 2020.

[9] N. K. Rajput, B. A. Grover, and V. K. J. a. p. a. Rathi, "Word frequency and sentiment analysis of twitter messages during Coronavirus pandemic," 2020.

[10] M. Ra, B. Ab, and S. Kc, "COVID-19 Outbreak: Tweet based Analysis and Visualization towards the Influence of Coronavirus in the World."

[11] A. D. J. A. a. S. Dubey, "Twitter Sentiment Analysis during COVID19 Outbreak," 2020.

[12] (2020). *Tweepy Documentation.* Available: http://docs.tweepy.org/en/latest/index.html
[13] V. Bilyk, "What is Sentiment Analysis: Definition, Key Types and Algorithms."
[14] W. Medhat, A. Hassan, and H. J. A. S. e. j. Korashy, "Sentiment analysis algorithms and applications: A survey," vol. 5, no. 4, pp. 1093-1113, 2014.
[15] *TextBlob: Simplified Text Processing*. Available: https://textblob.readthedocs.io/en/dev/index.html
[16] A. Stuart, S. Arnold, J. K. Ord, A. O'Hagan, and J. Forster, *Kendall's advanced theory of statistics*. Wiley, 1994.
[17] B. Blackwell, "Emotions (Part 1): What They Are, and How to Use Them."