Topic based Sentiment Analysis for COVID-19 Tweets

Manal Abdulaziz[1], Mashail Alsolamy[3]
The Department of Information Systems
King Abdulaziz University Jeddah,
Saudi Arabia

Abeer Alabbas[4]
Technical and Vocational Training Corporation
Najran College of Technology Najran,
Saudi Arabia

Alanoud Alotaibi[2]
The Department of Information Systems
Imam Mohammad Ibn Saud Islamic University
Riyadh,Saudi Arabia

***Abstract*—The incessant Coronavirus pandemic has had a detrimental impact on nations across the globe. The essence of this research is to demystify the social media's sentiments regarding Coronavirus. The paper specifically focuses on twitter and extracts the most discussed topics during and after the first wave of the Coronavirus pandemic. The extraction was based on a dataset of English tweets pertinent to COVID-19. The research study focuses on two main periods with the first period starting from March 01,2020 to April 30, 2020 and the second period starting from September 01,2020 to October 31, 2020. The Latent Dirichlet Allocation (LDA) was adopted for topics extraction whereas a lexicon based approach was adopted for sentiment analysis. In regards to implementation, the paper utilized spark platform with Python to enhance speed and efficiency of analyzing and processing large-scale social data. The research findings revealed the appearance of conflicting topics throughout the two Coronavirus pandemic periods. Besides, the expectations and interests of all individuals regarding the various topics were well represented.**

***Keywords*—Social media analysis; COVID-19; topics extraction; sentiment analysis; LDA; spark; twitter**

I. Introduction

The Coronavirus outbreak has been a severe disruption to the global economy and this has affected most, if not all, nations. In fact, ever since the World Health organization (WHO) declared the pandemic a Public Emergency of International Concern (PEIC) [1], the subsequent restrictions to curb the spread of the virus continue to do more damage than good. These restrictions include but are not limited to travelling restrictions and closure of non-essential businesses[2]. In light of these measures, most people have resorted to social media to express their views regarding everything that is going on in the world [3]. The impact of social media platforms is becoming more noticeable than ever before. Social networking sites are considered the global big data center because people use their applications and invest much time in these media outlets [4].

The usefulness of these social media platforms is attributable to their ability to highlight valuable insights on varying perceptions of issues that are happening in real-time [5]. Furthermore, social networking has a remarkable impact and is one of the most increasingly growing social information structures.

A predominant social media platform is one of the most common social networking sites. According to research, an analysis of Twitter data, especially people's emotions, can be useful in many areas such as the stock market, election vote, management of disaster, and crime [5]. Analyzing tweets during and after Coronavirus could be worthy as the condition and people's reactions are changing every instant during this critical period. This study motivated by examining how human emotions and concerns changing with respect to COVID-19 from the start of the pandemic and till now. We have attempted to used PySpark for improving the sentiment of topic modeling analysis and relies on a lexicon-based algorithm that is applied using big data and Machine Learning techniques. Researchers contributed to an opensource of textual datasets. To achieve this aim, a dataset of tweets about COVID-19 created by Christian et al. [6] is selected for this work since it is the only dataset covering the period from January to October, and is considered big data. The contribution of this work is to analyze the COVID-19 tweets dataset from January to October and compare the changes in people's feelings by applying machine learning methods and answering the following two research questions:

Question 1

What are the most traded topics in the COVID-19 pandemic in the course of two different periods, the first period from March to April (first pandemic wave), and second period from September to October (second pandemic wave)?

Question 2

How have people concerns change during the COVID-19 from the start of the pandemic and till now?

The rest of the paper is structured as follows: The second section presents the related works for sentiment analysis and topic modeling. Section three presents the content analysis methods. Section four shows the proposed model of this research while the data analytic and implementation of the model present in section five and six respectively. Section

seven discusses the work's results. Finally, this paper ends with conclusion and future work.

II. Related Work

The vast majority of research studies that cover tweets' sentiment analysis are more inclined towards machine learning algorithms [7]. An apt example is the analysis of the negative connotations that revolve around the Coronavirus related tweets. The researchers often utilize exploratory and descriptive methodology as well as the visual and textual data to get valuable insights based on Naïve Bayes Classifier (NBC) and logistics regression (LR) classification method of machine learning.

The findings showed 91% accuracy of tweets with Naïve Bayes while 74% with the logistic regression classification method [8] made prediction for the feelings of people on Twitter through building a model to explore the real sentiment of people about COVID-19. They made a comparison between five classifiers, which are logistic regression, multinomial Naive Bayes, Decision Tree, Random Forest, XGBoost, and Support Vector Machine (SVM) over n-gram for feature extraction along with bi-class and multi-class setting. The results showed that SVM and Decision Tree outperform the other classifier. However, the SVM classifier is stable and reliable in all tests. Nonetheless, most classifiers were better done with unigram and bigram in the bi-class setting. In addition, the maximum accuracy of the proposed model was 93%, indicating that the model has the ability to analyze the emotion of people within COVID-19 tweets. K- Chakraborty [9] presented all tweets relevant to COVID-19 and WHO were unsuccessful in guiding people across this pandemic outbreak. They analyzed twenty three thousand re-tweeted tweets over the timeframe from 1 January 2019 to 23 March 2020. The results showed that the highest number of tweets indicated neutral or negative emotions. On the other hand, a dataset comprising 226,668 tweets gathered between December 2019 and May 2020 were analyzed and revealed that there were a maximum number of positive and neutral tweets. Analysis reveals that while people tweeted mainly positively about COVID-19, Internet users were busy re-tweeting negative tweets and that no useful terms could be found in WordCloud. The accuracy reached up to 81% when using deep learning classifiers while 79% when using the formulated model based on a fuzzy rule to identify sentiments from tweets.

Regarding to topic modeling, B.Dahal [10] analyzed large datasets of geotagged tweets containing several keywords related to climate change using topic modeling and sentiment analysis. LDA was used for topic modeling to extract the topics in the text, and Valence Aware Dictionary and sentiment Reasoner (VADER) were applied to assess the general emotions and behaviors. Analysis of sentiments indicated that the general discussion is negative, especially when users respond to political or extreme weather events while Topic modeling reveals that the numerous subjects of climate change discussion are diverse, but that some topics are more prominent than others. The debate on climate change in the United States in particular is less focused on political topics than other nations.Kaila and Prasad [11] focused on the flow of Twitter's information during the spread of the Coronavirus. Tweets associated with Coronavirus are analyzed using sentiment analysis and topic modeling by using LDA.

The study concluded that the flow of information is correct and consistent with minimal misinformation in relation to corona virus outbreak. The LDA identified the most important and reliable topics relating to the epidemic of Coronaviruses while sentiment analysis verified the dissemination of negative sentiments such as fear along with positive sentiments such as confidence [12] A developed topic modeling LDA methodology, to classify the interesting topics from large-scale tweets connected to two well-known Indian county officials. Spark performed the topic modeling method with R language to improve the speed and performance for large-scale real time social data processing and analysis. In addition, tweet sentiment analysis is conducted in this study by using a lexicon-based approach to classify people's sentiment against these two leaders.

III. Content Analysis

The essence of content analysis is to define trends, themes or ideas within such qualitative data (i.e., text). Using content analysis allows researchers to find out about the aims, messages and impacts of communication content.

A. Sentiment Analysis

Sentiment analysis, one of the most promising methods for content analysis in social media, known as emotion AI or opinion mining, leads to natural language processing(NLP) and text analysis to systematically, quantify, extract, identify, and study effective states and personal information [9].

Sentiment analysis is widely applied in the voice of the customer materials such as survey responses and reviews. Analysis of sentimental peoples can be achieved by millions of likes and retweets, but this vast interaction with such a post does not reflect the importance of the feelings toward those posts [9]. This is because there are several factors, such as happiness, irony, satisfaction, sadness and anger among others. The aforementioned factors can have an impact on the nature of posts. However, broad extractions of human feelings from social media networks are important and strongly affect international public trends, market decisions as well as policy development [13].

This gets to show the importance of sentiment analysis in the interpretation of human feelings. An analysis of peoples' sentiments can be classified in different ways. The first one is text classification and this is also referred to as text categorization or text tagging which is the main process in sentiment analysis. It entails the classification of texts into organized groups and the calculation of sentiment analysis depending on the number of occurrences of positive and negative words in each document. Each document has both negative and positive scores. When calculating the sentiment document score, each negative word is denoted by -1, each positive word as +1, and neutral word is denoted by Zero as neutral word [14].

The three main classification levels are: Sentence, document, and aspect levels [9]. The last of the most common sentiment analysis classifications are based on the rating level. There are three main steps to show how sentiment analysis works [13]:

- Data Collection: This entails the use of certain keywords or hashtags to access the information users want depending on their interests. This information has various forms (e.g. tweets, posts, news, texts).
- Prepossessing: The collected information is processed during this step in order to prepare the data for the next phase. This phase includes three main stages. First, the cleaning stage contains the removal of repeated letters, text correction, normalization, stop word removal, and language detection. Next, the Tokenization method focuses on converting text into tokens until it becomes vectors. Lastly, the extraction of features such as grammatical structures and mining characteristics.
- Data analysis: In this stage, all data should be processed and then identified based on the main purpose of research, such as polarity identification, sentiment analysis, or frequency analysis. Processed identified the main purpose of research, such as polarity identification, sentiment analysis, or frequency analysis.

B. Topic Modeling

Topic modeling analyzes "bags" or groups of words together—instead of counting them individually–in order to capture how the meaning of words is dependent upon the broader context in which they are used in natural language. Topic modeling is not the only method that does this– cluster analysis, latent semantic analysis, and other techniques have also been used to identify clustering within texts. There are many techniques to implement topic model such as Latent Dirichlet Allocation (LDA), Latent Semantic Allocation (LSA), and Non Negative Matrix Factorization (NNME). One of the most popular topic models is LDA, which generates latent topics in whole corpus [10]. It is a probability distribution of the topics over every word found in the corpus. So, the number of topics and the words in each topics must be determined before LDA is run. The main assumptions of the model are that each document in the corpus is a probabilistic mixture of topics, and each topic is a probabilistic mixture of terms. Topic model requires a word topic matrix and a dictionary as the main inputs. Before running the LDA model, it should create a dictionary, filtered the extremes, and create a corpus object, which is the document matrix LDA model needs as the main input [10]. The most important parameter of LDA is the number of topics the model should infer from the corpus, k. It is not clear how many topics the dataset should be divided into. Too few topics could lead to an incomplete analysis while multiple distinct topics could theoretically be mixed together. Actually, too many topics could lead to multiple topics reflecting a coherent subject together but becoming individually confusing [10]. To address this problem, authors [10], [12] suggested running LDA with different number of K or such as, 5 topics, 20 topics, and 80 topics, then comparing the content of the inferred topics to decide the optimum number of topics. This method was preferred by many researchers to check the suitable number of K that can used to generate effective result.

The basic premise of topic modeling features entities: words, documents, and Corpora. Word is regarded as the primary unit of discrete data in a document, described as vocabulary items indexed for each uncommon word. The document is an organization of N words. Corpus is a set of M documents, and corpora are the plural shape of the corpus [15]. Today, one of the most common themes of modeling and analysis is LDA. As text data may have a mix of topics and insights, LDA tries to find a probability of hidden distributions in the input data. LDA is a "bag of words" words order is not essential. However, it assumes that documents that have related words customarily have the same topic. Moreover, documents that have collections of words frequently happening together customarily have the same topic. LDA contains two parts; the first part relates to a document, and this is already known. The second part is unknown because it entails words that relate to a topic or the probability of terms relating to a topic that the paper need to determine. Observations in LDA are pointed to as tweets content, the feature collection is meant to as words/vocabulary, and the resulting classifications are referred to as topics.

IV. Proposed Model

This section sums up the research methodology that has been implemented in this study. The methodology depends on two algorithms: the LDA to extract the most of the 'K' Topics from a tweet's text and lexicon-based approach to identify tweet's sentiment. LDA needs to train several times to adjust its parameters while lexicon-based approach uses directly. Therefor, there are training phase and runtime phase as shown in Fig. 1.

In training phase, the model starts with the preprocessing step, which is a crucial step to give the model valuable results and success. Preprocessing consists of four primary processes: text cleaning, tokenization, removing stop words, lemmatization and stemming. All these steps will be described in detail in the next section. Then, extract specific features from the tweet's text that should be in a well-defined format to be used directly as an input to the classification algorithms. There are several methods that can be applied in the extraction of features. In this study, TF-IDF will be used to construct a bag of words where tweets can be divided into words and generate a vector[8]. If one word is introduced, the method is called "unigram," "bigram" for two words, and "trigram" for three words, respectively [9]. So, this model extracts the feature vector based on the "unigram" method. Some parameters should be determined before running the LDA model. The number of topics (k) and the number of iterations, which controls how many times LDA will execute over each document, are the most critical values that should be selected carefully. Besides, selecting the number of topics depending on the chosen dataset is also essential. Therefore, executing the model and changing the values of K until generating the best result will perform this process. This model works based on NLP and will train with 80% of the dataset and test with 20%. After getting the topics, the paper will identify each sentiment. Since the dataset is not labeled by sentiment, a Vader lexicon model, which is a lexicon-based approach that considers a pre-trained model [16], will be used to identify tweet's sentiment. It takes a word and classifies its sentiment to positive, negative, or neutral. This can be performed with polarity calculation of sentiment words; there are three main steps to fulfill this analysis [8].

- Step (1): This step involves categorizing words as

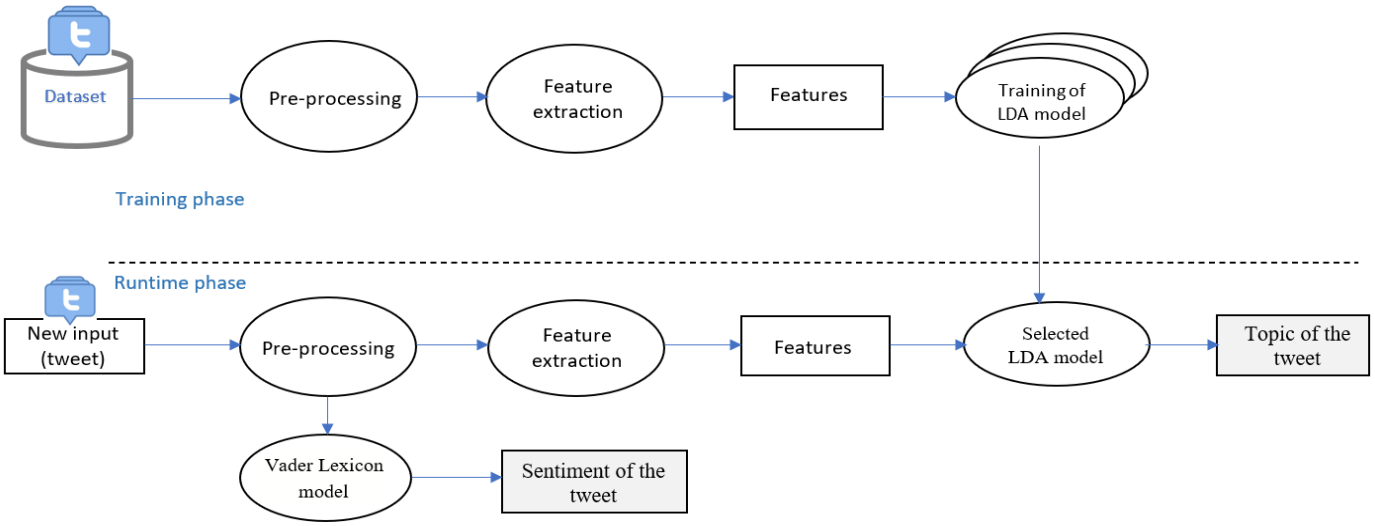

Fig. 1. The Proposed Model.

negative or positive in two groups, based on the frequency of negative or positive words. Then the polarity score for each tweet can be conveniently determined.

- Step (2): The tweet polarity is determined by summing the tweet value of each selected feature. Each tweet is scored using the following formula:

$$PolarityScore(tweet) = \sum featurevalue(tweet_i) \quad (1)$$

where n is the number of features in $tweet_i$.

- Step (3): To conclude a tweet's feeling, the following rules are established based on its polarity [8].
Polarity Score < 0, Negative
Polarity Score $= 0$, Neutral
Polarity Score > 0, Positive

In the runtime phase, we will follow the same process of preprocessing step and feature extraction. After that we will execute the trained model with this feature to get the topic of a new tweet. However, considering the Vader lexicon method only deals with words, it will be necessary to first send the processed tweet to obtain its sentiment.

V. Data Analytics

This section presents information about the dataset using in this research including the description of it and sampling phase.

A. Dataset Description

In order to examine human emotions and concerns with respect to COVID-19, researchers contributed to an opensource of textual datasets. To achieve our objective, a dataset of tweets about COVID-19 created by Christian et al. [6] is selected for this work since it is the only dataset covering the period from March to October, and it considers big data. The selected dataset is an open-source dataset available for research purposes. It has been collected continuously since 22 January 2020 using the Standard Twitter API. The total of obtained tweets were 942,149,169 tweets on 25 October 2020 for many languages. The English language represented 67.59% of the total tweets. The authors have dedicated more computational resources to pandemic-related tweets as the influence of COVID-19 rises around the world. This is one of the reasons why the number of tweets at some times has risen considerably. They used trending topics on Twitter and some keywords such as coronavirus, ncov19, ncov20, stay-at-home, covid, and virus to collect tweets. According to Twitter Service Terms and Conditions, this dataset contains Tweet ID only. Therefore, "rehydrate" the tweets is applied to obtain all information of tweets using the code available by the authors. Because the dataset's size is enormous, the paper presents its explanation and preprocessing (incoming section) on the file of 1September. As shown in Fig. 2 using Jupyter notebook, there are 33 attributes having different data types (Boolean, float, integer, object). In this study, four attributes were used and these are presented in Table I.

TABLE I. Description of Selected Attributes

Attribute Name	Description
Text	Unique identifier of the tweet as integer
Lang	The language of the Tweet text
Retweet_count	Indicates how many times the tweet was retweeted
Created_at	The date and time in (UTC) of creating a tweet

B. Dataset Sampling

This research worked on English tweets only, so there were 636,798,623 tweets. Since this work runs on personal computer with limited processing capacity, the dataset was decreased to approximately 600,000 tweets. The assumption is that highly retweeted tweets are more critical. For that assumption, we will select tweets having the highest retweet_count per month. Since the number of tweets varies per month, as shown in Fig. 2, a specific percent will apply to each month to track people's

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 415748 entries, 0 to 415747
Data columns (total 34 columns):
 #   Column                       Non-Null Count   Dtype
---  ------                       --------------   -----
 0   coordinates                  166 non-null     object
 1   created_at                   415748 non-null  object
 2   hashtags                     85184 non-null   object
 3   media                        31361 non-null   object
 4   urls                         121246 non-null  object
 5   favorite_count               415748 non-null  int64
 6   id                           415748 non-null  int64
 7   in_reply_to_screen_name      17158 non-null   object
 8   in_reply_to_status_id        15963 non-null   float64
 9   in_reply_to_user_id          17170 non-null   float64
 10  lang                         415748 non-null  object
 11  place                        1697 non-null    object
 12  possibly_sensitive           134811 non-null  object
 13  retweet_count                415748 non-null  int64
 14  retweet_id                   320323 non-null  float64
 15  retweet_screen_name          320323 non-null  object
 16  source                       415659 non-null  object
 17  text                         415748 non-null  object
 18  tweet_url                    415748 non-null  object
 19  user_created_at              415748 non-null  object
 20  user_screen_name             415748 non-null  object
 21  user_default_profile_image   415748 non-null  bool
 22  user_description             337442 non-null  object
 23  user_favourites_count        415748 non-null  int64
 24  user_followers_count         415748 non-null  int64
 25  user_friends_count           415748 non-null  int64
 26  user_listed_count            415748 non-null  int64
 27  user_location                273131 non-null  object
 28  user_name                    415726 non-null  object
 29  user_screen_name.1           415748 non-null  object
 30  user_statuses_count          415748 non-null  int64
 31  user_time_zone               0 non-null       float64
 32  user_urls                    114107 non-null  object
 33  user_verified                415748 non-null  bool
dtypes: bool(2), float64(4), int64(8), object(20)
memory usage: 102.3+ MB
```

Fig. 2. Description of Dataset's Attributes

Twitter activity about COIVD19. At the end, the processed dataset will save in a new dataset file of type CSV.

VI. Implementation

The implementation was done on two separated parts: preparation the dataset to be ready to use for the proposed model and implementation the model.

A. Dataset Preparation

Each month has 30 files, file per day, so we have 120 files. For each file, the paper applies the following steps using Python Jupyter notebook, tweet here means record or row of data: (file of September 1).

1) Remove non-English tweets according to "lang" attribute.
2) Remove duplicated tweets according to "text" attribute. As seen in Fig. 3, most tweets are retweeted tweets. And this explains why we do not determine a specific number of tweets from each day.

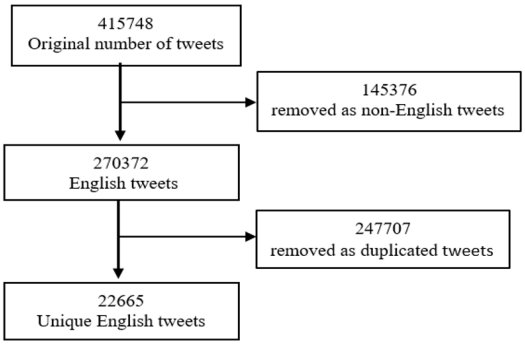

Fig. 3. Changing in the Total Number of Tweets over Preparation.

3) Keep the attributes "created_at", "text", and "tweet_count" and remove the rest.
4) Convert the date format from (Tue Jan 21 22:45:27 +0000 2020) to this format (1/21/2020) .
5) Sort tweets in descending order according to "retwee_count" attribute.
6) Save the result in a CSV file.

We got 567064 tweets after applying previous processes and the distribution of these tweets over each day is shown in Fig. 4 while Fig. 5 show them per months.

Finally, all prepared files are collected together in one CSV file, the sample from this dataset is presented in Fig. 6. As in text attribute, the text includes hashtag, mention, punctuation, symbols, etc. which need to be removed before entering the text to the model- the next section focuses on cleaning tweets.

B. Model Implementation

The model was executed on Google Colaboratory that provides a web based interactive computing platform. And PySpark programming language was used in writing the code. The implementation was done through several steps starting from preprocessing to getting the optimal model.

1) Data preprocessing: In this phase, the raw text of the tweet goes through several stages of cleaning it. This phase is significant in NLP, good cleaning leads to good results. Natural Language Toolkit (NLTK) is a package building in Python to process language. It contains a lot of libraries for text processing. implemented this cleaning was done in five steps using NLTK, as presented in Fig. 7, that gave excellent result.

Step 1: eliminating punctuations, URLs, numbers, hashtags, mention, and symbols from the text; if a symbol associates with a word, the word will also remove which achieves removing of hashtag and mention. Then, converting tweets into lower text. The result of this stage is shown below

Step 2: Applying tokenization that split a text into a list of words using the tokenization methods available in the NLP library. This step is essential to remove unwanted words, as done in the next step.

Step 3: eliminating stopwords located in the stopword library in NLTK. This library includes 179 stopwords of the English language as listed below.

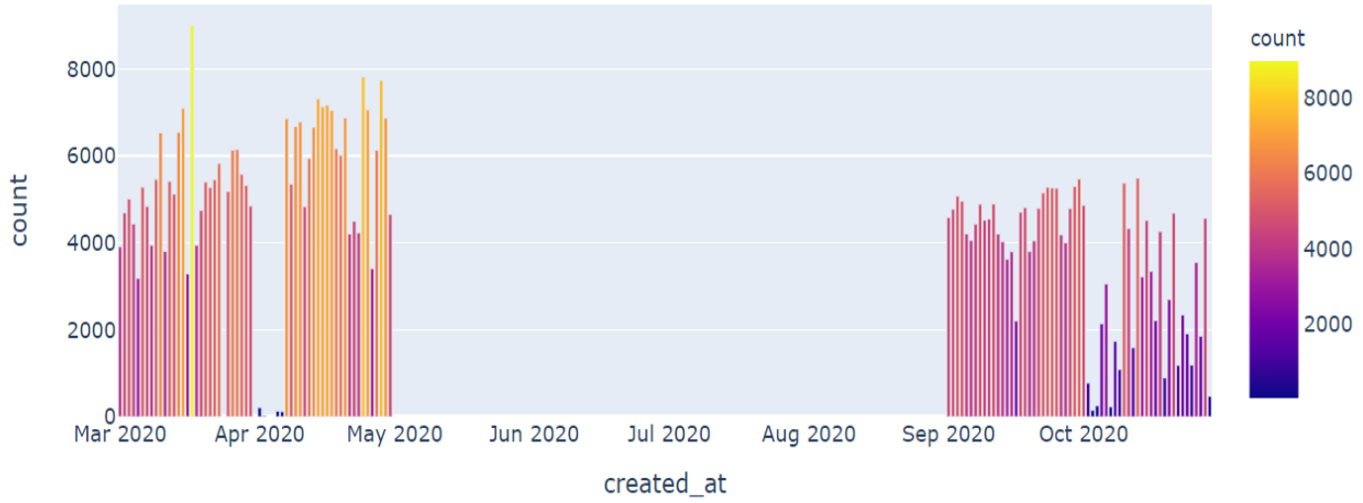

Fig. 4. Distribution of Tweets per Day over 4 Months.

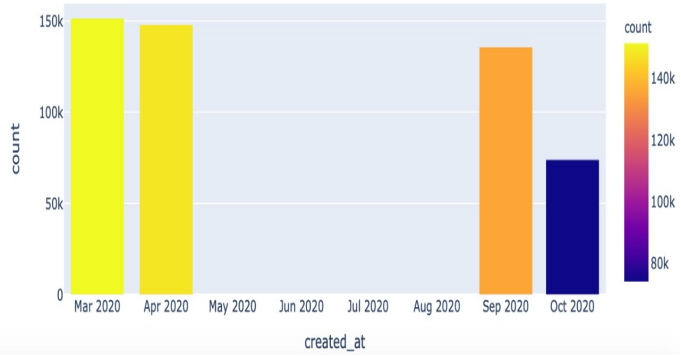

Fig. 5. Distribution of Tweets per Month.

	created_at	retweet_count	text
250480	2020-09-01	188715	RT @techinsider: What coronavirus symptoms loo...
23823	2020-09-01	152401	RT @Pog_llins: 900 people get Coronavirus and ...
21983	2020-09-01	150231	RT @NoahhbyNature: WWE airing their shows w/o ...
255924	2020-09-01	66385	RT @FantasyTweeting: When I found out the Coro...
256634	2020-09-01	40751	RT @williscj_: this was the start of the Coron...

Fig. 6. Sample of the Prepared Dataset.

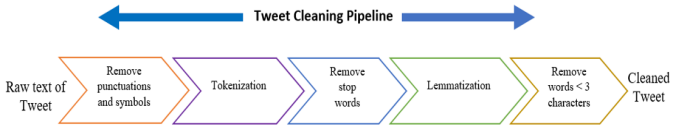

Fig. 7. Preprocessing of Tweet.

As seen, these words do not give meaning when we need to extract information from a text. Therefore, we removed from the tweet to become the text as follow:

```
0    rt what coronavirus symptoms look like day by day
1    rt llins 900 people get coronavirus and the wh...
2    rt wwe airing their shows w o a crowd because ...
3    rt when i found out the coronavirus could last...
4     rt this was the start of the coronavirus idc idc
Name: text, dtype: object
```

```
'nor',      'i',        'its',      'do',       'below',    'our',
'not',      'me',       'itself',   'does',     'to',       'ours',
'only',     'my',       'they',     'did',      'from',     'wasn',
'own',      'myself',   'them',     'doing',    'up',       "wasn't",
'same',     'ourselves','their',    'a',        'down',     'weren',
'so',       "should've",'theirs',   'an',       'in',       "weren't",
'than',     'themselves','what',    'the',      'out',      'have',
'too',      'you',      'which',    'and',      'on',       'has',
'very',     "you're",   'who',      'but',      'off',      'had',
's',        "you've",   'whom',     'if',       'over',     'didn',
't',        "you'll",   'this',     'or',       'under',    "didn't",
'can',      "you'd",    'that',     'because',  'again',    'doesn',
'will',     'your',     "that'll",  'as',       'further',  "doesn't",
'just',     'yours',    'these',    'until',    'then',     'shouldn',
'don',      'yourself', 'those',    'while',    'once',     "shouldn't",
"don't",    'yourselves','am',      'of',       'here',     'mightn',
'should',   'he',       'is',       'at',       'there',    "mightn't",
'now',      'him',      'are',      'by',       'when',     "isn't",
'd',        'his',      'was',      'for',      'where',    'ma',
'll',       'himself',  'were',     'with',     'why',      "needn't",
'm',        'she',      'being',    'about',    'how',      'aren',
'o',        "she's",    'having',   'against',  'all',      "aren't",
're',       'her',      'hadn',     'between',  'any',      'couldn',
've',       'hers',     "hadn't",   'into',     'both',     "couldn't",
'y',        'herself',  'hasn',     'through',  'each',     'wouldn',
'shan',     'it',       "hasn't",   'during',   'few',      "wouldn't",
"shan't",   "it's",     'haven',    'before',   'more',     'rt',
'we',       'ain',      "haven't",  'after',    'most',     'rts',
'be',       'mustn',    'isn',      'above',    'other',    'retweet']
'been',     'needn',    'no',       'won',      'some',     "mustn't",
                                    "won't",    'such',
```

Step 4: The target of this step is to convert words to their root to get distinguishing words. This process is called Stemming and Lemmatization. There are many methods in the NLTK package to implement them. We applied all of them to decide which one is better. Stemming methods are

```
0    coronavirus symptoms look like day
1    llins people get coronavirus whole world w...
2    wwe airing shows w crowd coronavirus gave us q...
3    found coronavirus could last july august
4    start coronavirus idc
Name: text, dtype: object
```

not preferred to use because they remove the letter 's' from the end of the original word; for example, "previous" become "previous" as explained in Table II, row 4.

On the other hand, using the general lemmatization method did not solve the problem because it does not convert verbs to their root; for example, "running" became "running" without change. To overcome these issues, we proposed approch which is apply the lemmatization method three times with change 'Postag' parameter each time in a certain order. The order of execution and effectiveness of this approach are explained in Table II, from 1 to 3. In row 1, all plural nouns converted to their root, while adjective words converted bypass 'a' to Postag parameter in row 2. And all verbs converted to their root by set Postag ='v' in row 3.

Step 5: The last step is to remove words of length 1 and 2, because they do not imply any meaning, especially 'rt' from retweet words that appear in each retweeted text. The final processed data are shown below.

```
0    coronavirus symptom look like day day
1    llins people get coronavirus whole world w...
2    wwe air show crowd coronavirus give quite po...
3    find coronavirus could last july august
4    start coronavirus idc idc
5    people hometown siena sing popular song house ...
6    light discovery count infant suffocation towar...
Name: clean_text, dtype: object
```

TABLE II. THE RESULT OF LEMMATIZATION APPROACH

Used Method		Result
Original Text		['running', 'presents', 'wives','better', 'paid', 'previous', 'cats']
1	Lemmatize method without set pos parameter	['running', 'presents', 'wives','better', 'paid', 'previous', 'cats']
2	Lemmatize method with set pos parameter to 'a'	['running', 'present', 'wife', 'good', 'paid', 'previous', 'cat']
3	Lemmatize method with set pos parameter to 'v'	['run', 'present', 'wife', 'good', 'pay', 'previous', 'cat']
4	Stemme method	['run', 'present', 'wive', 'better', 'paid', 'previou', 'cat']

2) Feature extraction: The next step of implementation is extracting the suitable features from the processed text. The paper applied TF-IDF method to generate a bag of words, then applied "unigram" approach to get the feature vector [8]. TF.IDF consists of two mathematically multiplied parts: TF and IDF. TF (Tearm Frequency)= frequency of word in the tweet / total words in the tweet. IDF(Inverse Document frequency)=log(total number of tweets/number of tweets that the word belongs to)

$$TF - IDF = TF * IDF \tag{2}$$

So, it produces a feature matrix of size number of tweets * number of unique words in all tweets, which then converts to one vector.

```
Here is example for one of the tweets:

processed_text =['drop', 'combat', 'look'], It calculate the frequency of
features=SparseVector(5000, {49: 4.5975, 249: 5.5715, 254: 5.5934}))
```

This method also is a built in method in machine learning library of PySpark that used by importing it as:

```
from pyspark.ml.feature import IDF
```

3) Training model: The LDA model was used to generate the topics. It is built model located in machine learning library under PySpark that can be used by import it as: from pyspark.ml.clustering import LDA The quality of the model's result connected with two important parameters: k and the number of iterations. To adjust these parameters, this executed the LDA model many times with change these values each time. The result of each time was evaluated to decide which values are the best, evaluation strategies of these parameters were presented in the next section. Then, the model was built depending on the best parameters, which can use with a new tweet.

4) Sentiment model: This paper used Vader-lexicon as a model for sentiment. It is a pre-trained model that can use immediately by importing it from NLTK library as: from nltk.sentiment.vader import SentimentIntensityAnalyzer After obtaining the final topics we sent them to Vader-lexicon to get the sentiment of each topic. Also, the paper used this model to pick up the sentiment of a new tweet after processing this tweet, as mentioned in data preprocessing section.

C. Evaluating the Topics

This section presents the evaluation strategies for selecting the appropriate topics, which depend on two parameters: number of topics and number of iterations.

1) Deciding the number of topics: As previously mentioned, large number of topics can lead to wasted topics and in contrast, small number of topics can hide useful topics. Therefore, deciding the number of topics depends on the size and the general subject of the documents. For the documents, the paper implemented LDA with K =5-10 and 15 and evaluated the quality of the outputs manually. The paper focused on the weight of words and the meaning of words in each topic. This paper found that an increase the words more than 10 words to each topics causes sharing some words between the topics as well as words with low weight. In addition, this paper found that 8 topics for our model gave satisfied result.

2) Deciding the number of iterations: Giving good topics depends also on the number of iterations. A the number of iterations is affected by the size of the documents and experiment, so LDA was executed with iteration = 50, 100, 150, 200, 300 , 1000 and evaluated the results manually in each time as done in selecting topics. The paper found that there is no significant difference in range 100 to 200 iterations. Therefore, the paper picked 150 iteration for our model.

VII. RESULT AND DISCUSSION

This work divided the dataset described in previous section into two periods. The first period covers from March 01,2020

TABLE III. TOP WORDS IN TOPICS USING LDA

Topic 1(Drug Research)	Topic 2(news)	Topic 3(Losses)	Topic 4(Economy)	Topic 5(Lockdown)	Topic 6(Updated Cases)	Topic 7(School Closures)	Topic 8(Rules)
vaccine	records	business	economy	minister	hospital	school	hand
study	everyone	worker	nation	question	number	student	thing
community	hour	fund	anyone	travel	confirm	stay	johnson
administration	member	look	rate	update	country	force	child
control	vote	service	return	cover	increase	march	rule
bill	fight	Italy	border	face	person	cancel	information
symptom	India	employee	system	watch	issue	press	post
talk	staff	market	supply	lockdown	patient	distance	result
drug	America	support	advice	message	action	event	change
research	measure	relief	read	warn	official	brief	medium

till April 30,2020 while the second period covers from September 01,2020 till October 31,2020. To analyze and identify the most topics traded in the COVID-19 pandemic during the two periods, it used LDA with K=8 (number of topics) and performed in 150 iterations. Table III and Fig. 8 answer to the research question 1 and shows the most traded topics in the two periods and the top words in each topic.

Fig. 8 shows the distribution of topics among documents with high weight for the third topic, Losses, than other topics. COVID-19 is affecting the world and causing a loss in the economy and humanity. Stop most of the markets and services were affected on employment. Most of the employees and other workers stayed without any work. There was a need for financial support to perform their business. These workers don't sense relief and need more support by providing a fund to eliminate the effect of Coronavirus. Italy is one of the countries affected by this epidemic. The number of registered infections with Coronavirus has increased. In addition, there is an increased number of deaths in the first pandemic. The drug researches and updated cases are two topics were presented with little difference in the weights. In drug researches, the Ministry of health and all its administrations were talked about the latest researches that tried to find vaccines and drugs. They studied the current symptom and how can control this virus to avoid dissemination. When this vaccine is ready, it will bill it to all countries in the world to face this virus. Updated cases is topic discusses how the official websites in each country issued updated number of the confirmed cases that effect with COVID-19. Several actions were performed in each hospital to be ready to receive persons who confirmed affected with COVID-19.

In addition, school closure was the most topic present due to increase the number of active cases. Many schools and universities in the world have already had to isolate classrooms or close entire schools due to the outbreak of Corona infection. Most schools cancel students' attendance and allow them to stay at home and study through distance learning to complete their courses. Topic 8 discuss the Rules published by the official website and in social media. These websites posted the updated information and current results what is the most changes that happen. In March, British Prime Minister Boris Johnson conducted a test for the Coronavirus, and the result of this test was that he was infected with the Coronavirus, and he has published this information across the sites. In addition, these rules were focused on the children and it necessary to clean hands and follow the necessary information. One of the ways to reduce the number of Coronavirus infections is a lockdown. Most of the world's countries have banned travel to another country. During the outbreak of the Coronavirus, a meeting of health ministers was held virtually by watching through a video conference to answer the most important questions and issues. The Ministry of Health confirmed to stay at home, and published warning messages that include not going out or traveling to another country, and not leaving your face uncovered. These messages have been updated based on new events.

Economy is also another topic present in our research. The nations expected shrink of the world's economy due to Coronavirus and the closure of borders, and their relations with other countries. Public health, people's trust in the instructions they read, and state support for them can contribute to restoring the economy. The least visible topic in the Fig. 8 is News, the news during the COVID-19 pandemic show that India becomes the second country most affected by the epidemic, after America. Additionally, there is a significant rise in casualty numbers in hours, which made India fight this epidemic by implementing a set of measures and enforcing everyone in the country to apply them. Also, other staff and other members help in fighting COVID-19.

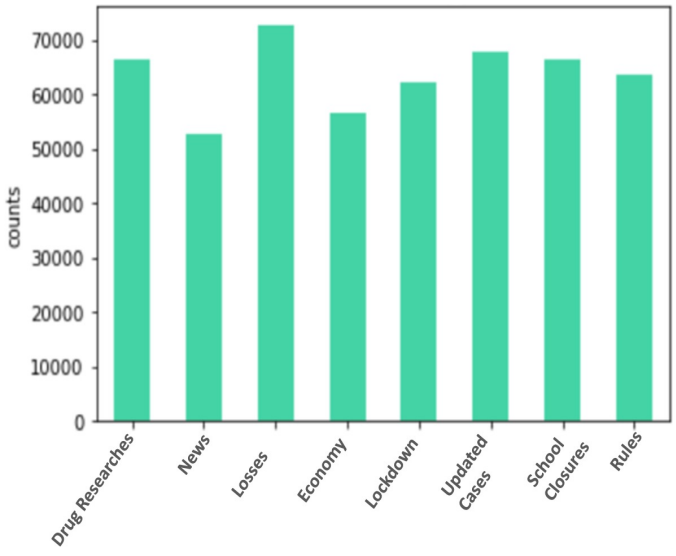

Fig. 8. Topics Distribution among Documents

A. Analysis of COVID-19 Trends

After the discussion of the most topics traded in the pandemic of COVID-19, this section analyzes the topics during

the periods. The study is divided into two main periods according to the outbreak of the Coronavirus around the world. The first period covers March and April(Pandemic revolution), the second period along the months of September and October(Epidemic stabilization). The topics were appeared at different rates in each period according to the strength of the epidemic and its appearance during the period. Fig. 9 and Table IV below show the topic distribution during the two periods.

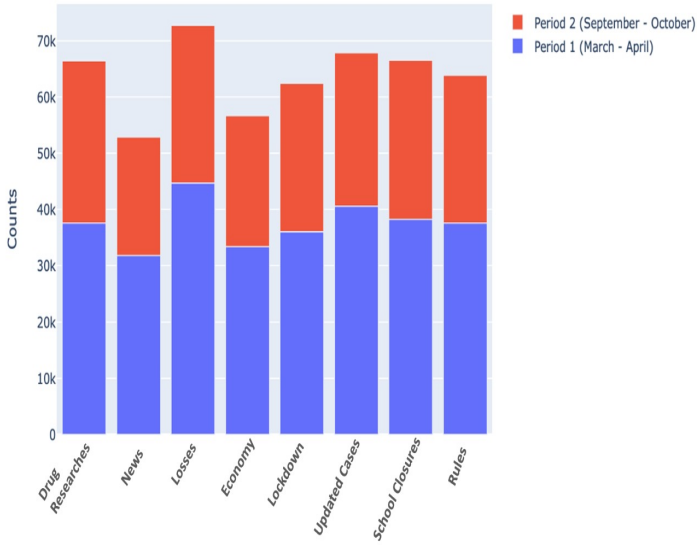

Fig. 9. Topic Distribution during the Months

TABLE IV. DISTRIBUTION OF THE TOPICS THROUGH THE PERIODS

Topic	1st period	2nd period
Drug Research	0.5658	0.4341
news	0.6003	0.3996
Losses	0.6134	0.3865
Economy	0.5883	0.4117
Lockdown	0.5558	0.4242
Updated Cases	0.6073	0.3966
School Closures	0.59001	0.4023
Rules	0.5887	0.4218

In general based on the information in Fig. 9 and Table IV, all of the topics were distributed between the two periods in close proportions. “losses” is the most topic with high percentage than others while “drug research” is the most in the second period. This is a natural state because COVID-19 is affecting the world and causing a loss in the economy and humanity. Stop most of the markets and services were effected on business. “Lockdown” got a high percentage in the first period and then decreased slightly to 42%. Most activities have been curtailed during the Coronavirus outbreak.

Currently, there is an increasing number of cases after a recession in previous months, which could lead to a return to lockdown again. In the same way the 7th topic “School Closures” decreased little bit because until now all schools and universities perform most of educational learning online without student’s attendance. The learning process that performed in the first period is similar in the second one but may much effective now. Furthermore, “updated cases” topic

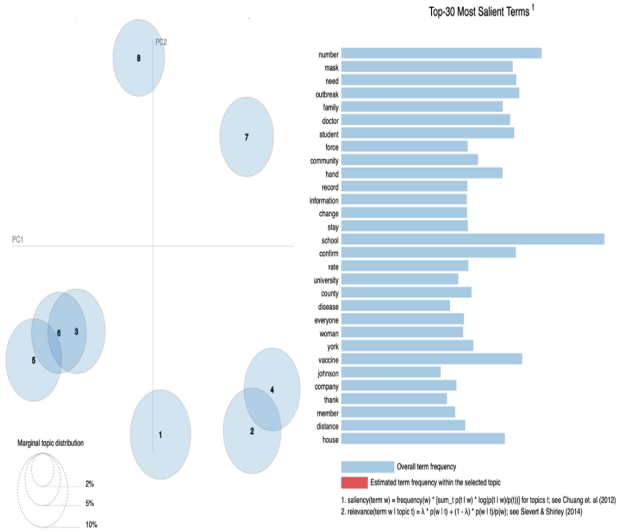

Fig. 10. Top Eight Clusters Representation

was decreased in the second period to reach about 39% . This happen due to some change in the situations in the two periods. Today, people have adapted to the current situation, and the number of cases decreased than it was at the beginning of the period, as the number of confirmed cases reached a high percentage compared to what is happening today. In addition, the “news” topic was also decreased for the same reasons mentioned before in the ”updated case” topic. At the beginning of the pandemic there was high percent of posted news regarding COVID-19 in each minute or second. Most people were good followers of this news with continuous in watching the updated cases and number of infection but it decrees in the month of September and October.

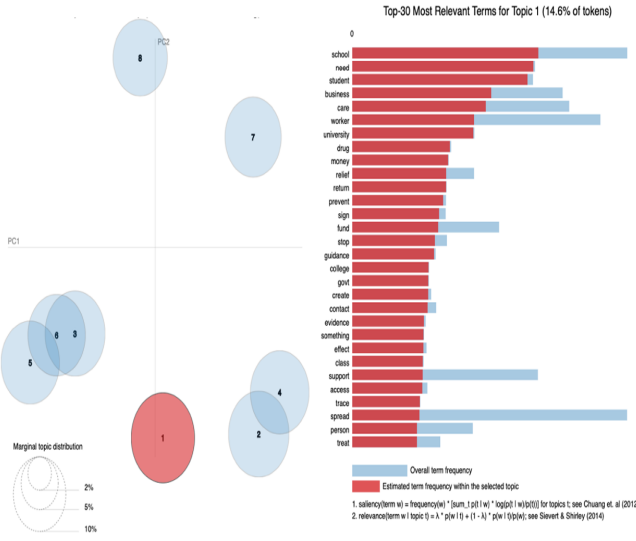

Fig. 11. Term Frequency of First Cluster/Topic Representation

Fig. 10 and 11 display the interactive topics chart by using LDAvis, which is a web based tool to visualize the topics using LDA [17]. It provides an interactive and global view of the most chosen topic. This is executed by Spark and

with the value of K=8. The two figures display LDA plot visualization for the top eight topics discussed in the previous. They demonstrate the general view of the topic model and the interaction between topics. The total term frequency is defined by the light blue color, while the red one reflects the approximate term frequency within the chosen topic. Fig. 10 indicates the average word frequency across all topics while Fig. 11 reflects the cluster of topics match with the frequency of words within the a selected topic. For example in topic1, the most words frequencies were schools, students, need and business. Each circle in the figures matches a topic, the right panel has the horizontal bar plot reflecting the relevance of currently selected topic for words indicating the top 30 words in the topic separately.

B. Textual Data Sentiment Analytics

In this paper, the eight topics were identified from each of the categories: neutral, positive, and negative. Topics are visualized with word cloud shown individually. These tasks are not easy since many pre-processed words have no semantic meaning. However, it can be difficult to understand the relationship between various tokens/words in these subjects, and these meanings can differ significantly from other forms of reviews. Fig. 12, 13, and 14 shows the most positive, neutral, and negative words for each topics. Therefore, the second research question can be answered in this section as it will appear in the following paragraph.

Fig. 12 shows the most positive words for each topics. For the positive cases, “agreement”, “ability”, “advantage”, “appreciate”, “accept”, “authority”, “assure”, “benefit”, “admit”, “asset” and, “allow” are the most frequent words. Those words denote a positive aspect, which is people accepting the COVID-19 pandemic and trying to live with and accept it.

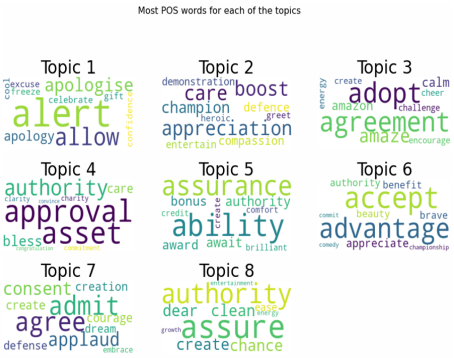

Fig. 12. The Most Positive Words for each Topics

Fig. 13 shows the most neutral words for each topic which are “accord”, “adult”, “adviser”, “access”, “academy”, “absence”, “abundance”. The neutral category is typically more representative and appeared the blend of positive and negative topics which displays the most frequent topics in recent times. Even though they do not mean a specific emotion, they shed light on subjects that are important to users.

Fig. 14 shows the most negative words for each topic which are “cancel”, “attack”, “avoid”, “burden”, “anxiety”, “abuse”, “accident”, ”anger”, “absentee”. These words denote a negative aspect, of people’s anxiety, fears, and surrounding circumstances during the COVID-19 pandemic.

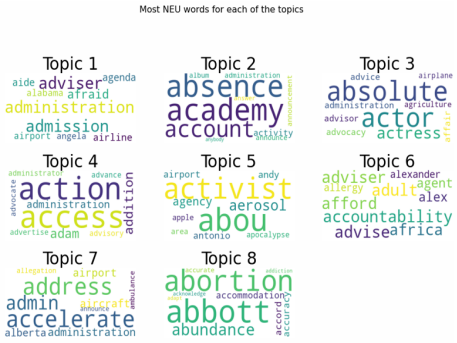

Fig. 13. The Most Neutral Words for each Topics

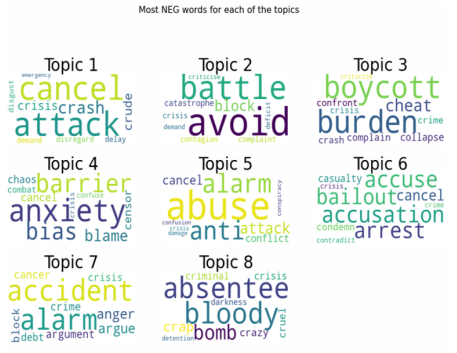

Fig. 14. The Most Negative Words for each Topics

VIII. Conclusion and Future Work

Social media allows people to not only share their opinions but also express their feelings. A widely used social media platform is Twitter. The research conducted in this paper entailed developing a model that can extract the most topics traded in the Coronavirus pandemic and then analyze the sentiment of these topics during two different periods, from March to April and September to October. For this study, the dataset of English tweets about COVID-19 was selected. 567,064 tweets were processed and analyzed. The implementation was done through several steps starting from pre-processing to getting the optimal model. This work has performed lemmatization step three times to get the best result and overcome the problems of converting word to its right root. So, this approach is the best and delivers the best result. Additionally, this research used LDA model for finding the topics and produced the eight most important topics related to the Coronavirus, which are presented and discussed. This model works based on NLP and will train with 80% of the dataset and test with 20%. Furthermore, this paper presented the sentiment analysis of the collected tweets using lexicon-based approaches to classify people’s feelings based on most of the traded topics. The experiment of this research conducted on the Spark platform with Python to enhance the analysis and processing large set of related tweets.

The challenges of this research were in the dataset preparation phase. The dataset contains tweet-ID only, therefore, it got more time to rehydrate it and then extract all information of tweets using the code available by the dataset's authors. This study has some limitation, it appeared when working with a big data volume, the research spent much time to do that comparing to the time given to finish this research. Furthermore, separating tweets of each day in a file caused long time to collecting them and merge in one dataset. In the future, the plan to label the prepared dataset with its sentiment and use other machine learning algorithms like supervised algorithms, and then comparing the result of the labeled dataset with the result of sentiment analysis that performed in this research.

REFERENCES

[1] "A Joint Statement on Tourism and COVID-19 - UNWTO and WHO Call for Responsibility and Coordination." [Online]. Available: https://www.who.int/news/item/27-02-2020-a-joint-statement-on-tourism-and-covid-19—unwto-and-who-call-for-responsibility-and-coordination (accessed Jan. 21, 2021).

[2] S. Al-Saqqa, G. Al-Naymat, and A. Awajan, "A large-scale sentiment data classification for online reviews under apache spark," in Procedia Computer Science, 2018, vol. 141, pp. 183–189, doi: 10.1016/j.procs.2018.10.166.

[3] E. Alomari, I. Katib, and R. Mehmood, "Iktishaf: a Big Data Road-Traffic Event Detection Tool Using Twitter and Spark Machine Learning," Mobile. Networks Appl., 2020, doi: 10.1007/s11036-020-01635-y.

[4] L. DeNardis and A. M. Hackl, "Internet governance by social media platforms," Telecomm. Policy, vol. 39, no. 9, pp. 761–770, 2015, doi: 10.1016/j.telpol.2015.04.003.

[5] M. Y. Kabir and S. Madria, "CoronaVis: A Real-time COVID-19 Tweets Data Analyzer and Data Repository," 2020, [Online]. Available: http://arxiv.org/abs/2004.13932.

[6] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of covid-19 policies by mining a multilanguage twitter dataset," pp. 3–6, 2020.

[7] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," Information, vol. 11, no. 6, pp. 1–22, 2020, doi: 10.3390/info11060314

[8] M. Sethi, S. Pandey, P. Trar, and P. Soni, "Sentiment Identification in COVID-19 Specific Tweets," in Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020, pp. 509–516, doi: 10.1109/ICESC48915.2020.9155674.

[9] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, and R. Bag, "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers — A study to show how popularity is affecting accuracy in social media," Appl. Soft Comput. J., vol. 97, p. 106754, 2020, doi: 10.1016/j.asoc.2020.106754.

[10] B. Dahal, S. A. P. Kumar, and Z. Li, "Topic modeling and sentiment analysis of global climate change tweets," Soc. Netw. Anal. Min., pp. 1–20, 2019, doi: 10.1007/s13278-019-0568-8.

[11] R. P. K. Kaila and A. V. K. P. Prasad, "Informational Flow On Twitter - Corona Virus Outbreak – Topic Modelling Approach," Int. J. Adv. Res. Eng. Technol., vol. 11, no. 3, pp. 128–134, 2020.

[12] P. Monish, S. Kumari, and C. Narendra Babu, "Automated Topic Modeling and Sentiment Analysis of Tweets on SparkR," in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1–7, doi: 10.1109/ICCCNT.2018.8493973.

[13] A. H. Alamoodi et al., "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," Expert Syst. Appl., pp. 1–13, 2020.

[14] A. Assiri, A. Emam, and H. Al-Dossari, "Real-time sentiment analysis of Saudi dialect tweets using SPARK," 2016 IEEE International Conference on Big Data, pp. 3947–3950, 2016, doi: 10.1109/BigData.2016.7841071.

[15] E. S. Negara, D. Triadi and R. Andryani, "Topic Modelling Twitter Data with Latent Dirichlet Allocation Method," 2019 International Conference on Electrical Engineering and Computer Science (ICECOS), Batam Island, Indonesia, 2019, pp. 386-390, doi: 10.1109/ICECOS47637.2019.8984523.

[16] Gilbert, C. H. E., & Hutto, E. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In 8th International Conference on Weblogs and Social Media(ICWSM-14),p.216-225.

[17] C. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," in Proceedings ofthe Workshop on Interactive Language Learning, Visualization, and Interfaces, 2015, pp. 63-70, doi: 10.3115/v1/w14-3110