ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 11 (2021) No. 4, pp. 167-189
DOI:10.33168/JSMS.2021.0409

Cluster Analysis on Covid-19 Outbreak Sentiments from Twitter Data using K-means Algorithm

Adnan Hussein [1], Farzana Kabir Ahmad [2] and Siti Sakira Kamaruddin [2]

[1] Department of Computer Science, College of Computer Science and Engineering, AL-Ahgaff University, Hadhramaut, Yemen

[2] School of Computing, College of Arts and Science, Universiti Utara Malaysia, Sintok, Kedah, Malaysia

taibh_310@hotmail.com, farzana58@uum.edu.my, sakira@uum.edu.my

Abstract. With the remarkable advances on the Internet and recent computer technologies, social media has become prominent platforms to share opinions. Nowadays, COVID-19 is considered as one of the major crises in the world. People use social media to express their thoughts about COVID-19 and actions that have been taken to control it. There is an immense need to discover and understand the public sentiment related to COVID-19 to give better insights for the decision makers and governments in making accurate decisions. In regard to this interest, several researches have explored COVID-19 outbreak sentiment analysis, however most of these studies have used classification approach which require the data to be manually labelled. In analyzing large number of data, labelling process can be an intricate task and expert dependent. This study aims to explore COVID-19 pandemic sentiment by using clustering approach. The data is obtained by crawling COVID-19 related posts from Twitter. The crawled data is pre-processed, and terms are extracted by using Term Frequency-Inverse Document Frequency (TF-IDF) technique. Singular Value Decomposition (SVD) technique is then used to reduce irrelevant features. K-means algorithm is employed to cluster the tweets into k clusters. The results of each cluster are plotted using t-Distributed Stochastic Neighbour Embedding (t-SNE) technique and lexicon-based sentiment analysis has been applied to discover sentiments of these clusters. The results showed relatively 9 clusters were obtained with different topics ranging highest score of 83.25% positivity and 16.75% of negativity are reported. Dominant topics are explored using word cloud and the clustering results have been evaluated with 0.0070 Silhouette coefficient. In future, this study suggests in using other word embedding technique as a data representation to deal with sparsity and high dimensionality of textual data.

Keywords: Clustering, K-Means algorithm, COVID-19, sentiment analysis, Twitter.

1. Introduction

Nowadays social networks can be considered as the most popular platform to gain and spread information among users worldwide. According to the Global Digital reports in the year 2021, there are more than 4.66 billion users around the globe who are using the Internet, and apparently 4.20 billion of them are using social media. This denotes that nearly 60% of the world's population is already online (Kemp, 2021). Social networks have become such preferable tool due to its fascinating characteristics in which it allows users to express and share information rapidly. The emergence of social networks has allowed people to stay connected regardless their geographical location and modernized the way people are communicating. Today, people use social networks such as Facebook, Twitter, Instagram, and LinkedIn to search for information, posting/sharing comments, upload photos, reading news. In addition, they also use this platform to express their feeling on certain events.

Recently, Coronavirus – also known as COVID-19 is global health crisis and has caused large scale of mortality. COVID-19 began to appear at Wuhan, China at the end of 2019 and quickly spread over worldwide and has been declared as a global pandemic (Liu et al., 2020; El-Din et al., 2020). According to the World Health Organization (WHO), the total number of confirmed cases of COVID-19 disease reach to more than 66.5 million cases in the time of writing this article including more than 1.5 million deaths and this number is still increasing. This fast spread of COVID-19 makes it the most discussed topic in social networks. There are massive posts in social networks like Twitter, Facebook and YouTube about COVID-19 that carry many conflicted opinions with different sentiments like fear, anger, bewilderment, exaggerations, lying and so on. Analyzing these posts can give a clear idea about the real reaction of people towards this pandemic, which could be beneficial for government to monitor and take precise decision.

Data clustering is a powerful tool to discover the insights and patterns, which are hidden in data. Clustering approach does not require any pre-knowledge of data labels, and this makes it free from human interpretation. Lately data clustering has been used in conjunction with sentiment analysis to explore human opinions. Sentiment analysis is an active research topic in the field of Natural Language Processing (NLP) (Manguri et al., 2020). It has been used in some interesting commercial and scientific areas, such as event detection (Moutidis & Williams, 2020), social tension detection (Shchoholiev et al., 2021) and recommender systems (Contratres et al., 2018).

This study aims to present cluster analysis on Covid-19 outbreak using K-means algorithm. Data has been crawled from Twitter, cleaned, and pre-processed prior to data clustering process to discover distinctive clusters from feature space. The sentiment analysis is applied to explore the sentiments included in these clusters. The paper is organized as follows: Section 2 provides related works in COVID-19 sentiment analysis, meanwhile Section 3 explains the research framework of this study. Section 4 on the other hands presents the experimental results and discussion. Section 5 concludes the study by offering potential future works.

2. Related Works

Experimental analysis of social media data has been conducted by many studies in the literature (Manguri et al., 2020; Orkphol & Yang, 2019; Ruz et al., 2020). These studies provided support to reveal and explore the different user behaviors and opinions towards several events that are happening around the world. COVID-19 is considered as one of the biggest global event and it requires in-depth investigation, as the published studies on this event are still very limited (Narasamma et al., 2020). Several of COVID-19 aspects and phases were tackled by these studies, such as the sentiment and emotion analysis toward the initial phase of COVID-19 lockdown (Imran et al., 2020), the geographical distribution of emotions toward COVID-19 (Venigalla et al., 2020), the analysis of the public opinion toward the reopening of lockdown (Ahmed et al., 2020) and analysing of the COVID-19 discourses (Xue et al., 2020). Furthermore, some studies collected and published datasets of COVID-19, for instance in the study by Gupta et al. (2020).

COVID-19 sentiment analysis has been studied by using different social networks (Ahmed et al., 2020; Boon-Itt & Skunkan, 2020; Cruickshank & Carley, 2020; Gupta et al., 2020; Kabir & Madria, 2020; Manguri et al., 2020; Medford et al., 2020; Narasamma et al., 2020; Nemes & Kiss, 2020; Pokharel, 2020; Samuel et al., 2020; Xue et al., 2020; Yin et al., 2020) and Wiebo (Wang et al., 2020). It is notable that there are many studies used Twitter data in their works, this due to the ease of getting data by using Twitter API and the reputation of Twitter as a prominent source of news, where it is spreading information more easily and quickly than traditional news media (Gupta & Arora, 2016). In addition, researchers have taken different approaches to conduct sentiment analysis of COVID-19 posts that include classification approach and clustering approach. Table 1 shows a summary of related studies for COVID-19 sentiment analysis on social media data.

The classification approach studies (Ou et al., 2015; Samuel et al., 2020; Manguri et al., 2020; Pokharel, 2020; Nemes & Kiss, 2020; Wang et al., 2020) typically requires a pre-labelled data in order to perform the classification process. The whole dataset is divided into training and testing sub-datasets and classifiers are used to classify these data points. Although the classification approach can achieve high accuracies, it required pre-labelled training data. Labelling process is a complicated task and need a lot of effort (AL-Sharuee et al., 2018; Orkphol & Yang, 2019). Moreover, the classification approach usually is domain or expert dependent which make this model difficult to be generalized (AL-Sharuee et al., 2018).

Another machine learning type used in sentiment analysis is the clustering approach. Clustering is a process to group the data points into groups (clusters) based on some similarity criteria (Cruickshank & Carley, 2020). This approach does not require any pre-knowledge of labels, and this makes it more practical than the classification approach (AL-Sharuee et al., 2018). Clustering approach has been applied on some studies for sentiment analysis (Cruickshank & Carley, 2020; Yu et al., 2020). Its main idea is to cluster data into a certain number of clusters or groups then applying sentiment analysis on these clusters to determine the sentiment hold by each cluster. From the literature, there are limited studies on sentiment analysis using clustering approach. Therefore, there is a need for further investigation and exploring of this approach, its tools, and techniques.

Most of the published studies on COVID-19 sentiment clustering, are conducted under topic modelling (Boon-Itt & Skunkan, 2020; Gupta et al., 2020; Kabir & Madria, 2020; Medford et al., 2020; Xue et al., 2020; Yin et al., 2020). They have extracted the topics from the processed data then grouped these data based on the topics. Data representation is important task in analyzing textual data. Several data representation techniques have been used by previous studies that includes TF-IDF (Boon-Itt & Skunkan, 2020; Cruickshank & Carley, 2020; Kabir & Madria, 2020; Wang et al., 2020) , Word-Frequency (Samuel et al., 2020; Xue et al., 2020; Yin et al., 2020) , word2vec (Yu et al., 2020), N-grams representation (Ahmed et al., 2020; Medford et al., 2020), and TextBlob (Gupta et al., 2020; Manguri et al., 2020; Nemes & Kiss, 2020; Pokharel, 2020).

As a summary, the previous studies showed that the clustering approach is an effective and practical approach for sentiment analysis and zero human interruption. Moreover, most of the studies in the clustering approach are focused on topic modelling while there is a need for profound discovering of the posts contents and reveal the sentiment of hidden clusters using different clustering techniques. This study aims to cluster Twitter data by using K-Means algorithm with TF-IDF data representation. The main aim of this study is to discover and analyse the other different aspects of COVID-19 which are concerned and discussed by people in social network such as COVID-19 Vaccine and the COVID-19 life procedures like wearing masks, social distance and staying home.

3. Exploratory Data Analytics Framework

The exploratory data analysis is an important process in order to discover the insights and opinions hidden inside the data. Fig. 1 shows the main framework used in this study to explore and investigate the insights and sentiment of Twitter data. The detail description of each process is presented the following sections.

3.1. Data Collection

The data related to COVID-19 is used in this study and has been crawled from Twitter using Twitter API and Rapidminer software. The crawling process starts by connecting Rapidminer with Twitter API and searching for relevant tweets using the different hashtags keyword, such as: #Covid-19, #Coronavirus, #QuarantineandChill, #LockdownNow, #SocialDistancing, #Quarantine, #StayHome, #Lockdown. The data is collected from 25 November 2020 till 28 November 2020, and about 23,513 English tweets are saved in CSV file.

Table 1: Related studies of sentiment analysis for COVID-19

Approach	**Data**	**Study**	**Techniques**	**Data Representation**	**Sentiment Scoring**
Classification	Twitter	Samuel et al., 2020)	Naïve Bayes + logistic regression	Word Frequency	NRC sentiment lexicon using Syuzhet and SentimentR
		Manguri et al., 2020	TextBlob-Naïve Bayes	TextBlob	TextBlob + Emotional Guidance Scale
		Pokharel, 2020)			
		Nemes & Kiss, 2020)	Deep learning RNN	TextBlob	TextBlob
	Weibo	Wang et al., 2020	Bidirectional Encoder Representations from Transformers (BERT)	TF-IDF	Manually Labelled
Topic Modelling Clustering	Twitter	Kabir & Madria, 2020	LDA	TF-IDF	Sentiment intensity analyser from NLTK
		Boon-Itt & Skunkan, 2020			National Research Council (NRC) sentiment lexicon
		Yin et al., 2020	LDA	Word2id + word frequency by Gensim	VADER
		Gupta et al., 2020	LDA	LDA	CrystalFeel
		Medford et al., 2020	LDA	N-gram representation	Syuzhet R package
		Xue et al., 2020		document-term matrix	Emotion classifier for English Tweets

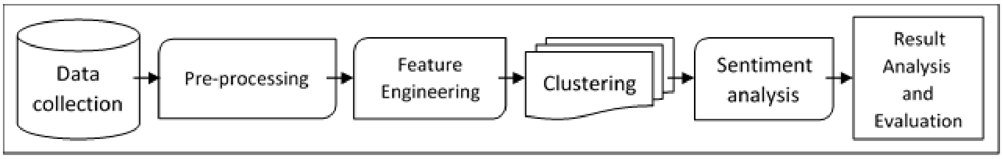
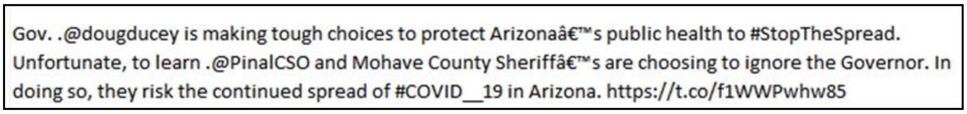


Fig. 1: Exploratory Data Analytics Framework

3.2. Data Pre-processing

The collected Twitter data typically contains much noise and irrelevant data that may affect the accuracy and performance of data analysis. Thus, the pre-processing is an important stage prior to subsequent analysis. Fig. 2 shows the raw sample of the collected raw tweet that has noise such as @, #, http://.



Gov. .@dougducey is making tough choices to protect Arizonaâ€™s public health to #StopTheSpread. Unfortunate, to learn .@PinalCSO and Mohave County Sheriffâ€™s are choosing to ignore the Governor. In doing so, they risk the continued spread of #COVID__19 in Arizona. https://t.co/f1WWPwhw85

Fig. 2: Sample of the collected raw tweet

3.2.1. Data Cleaning

Fig. 2 shows the collected tweets sample that usually contain other symbols and/or words that have no meaning in the context of data subject. These symbols can affect the performance and cause low accuracy results. These noisy data includes URLs (http://t.co/f1WWpwhw85), digits (19, 3,), punctuations and special characters (!"#$%&'()*+, -./:;<=>?@[\]^_`{|}~), mention symbols (@), hashtags (#), empty tweets, tweets with less than two words and duplicated tweets. This study cleans these noisy data by writing a python code to track and remove each of these unwanted symbols and words from each tweet. The result of this task is a cleaned tweet without any non-word characters.

3.2.2. Tokenization

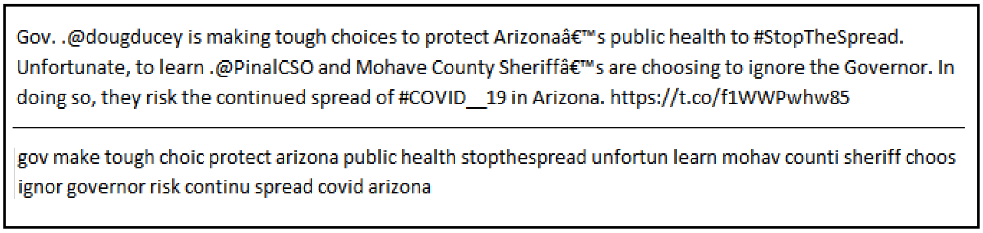
Tokenization is a process of breaking down the stream of text into a list of basic elements called token and each token is called feature. Machine learning analysis depends on examining these features in order to reveal the hidden patterns. The usual separator used in tokenization is the white space between words. There are many tools to perform tokenization such as spaCy library, Keras, Gensim and NLTK. The Natural Language Tool Kit (NLTK) word tokenizer were used to build the features lists in this study. The result of this task is a list of the tokenized tweets.

3.2.3. Removing Stop Words

Stop words are the words that commonly used in each language. These words are meaningless such as Determiners (the, a, an, another), Coordinating conjunctions (for, nor, but, and, or, yet, so) and Prepositions (in, under, towards, before). Removing these words can give more chance for the processing task to focus on the other important words that carry sentiment and meaning. There are several published lists of stop words and there is no single universal list of stop words. This study used one of the famous stop words lists published by NLTK.

3.2.4. Stemming

Stemming is a process to reduce the inflected words to their stem by removing the suffixes and prefixes. It is one of the text normalization techniques. Stemming is important, as there may be several different words with the same single base. For instance, the words “protection”, “protections”, “protecting”, “protected” may be reduced to one single word “protect”. This study has used “snowball stemmer” for the English language to perform stemming of each tweet’s token. Fig. 3 shows the sample of tweet before and after pre-processing task, in which the tweet post has been cleaned, tokenized, and stemmed.



Gov. .@dougducey is making tough choices to protect Arizonaâ€™s public health to #StopTheSpread. Unfortunate, to learn .@PinalCSO and Mohave County Sheriffâ€™s are choosing to ignore the Governor. In doing so, they risk the continued spread of #COVID__19 in Arizona. https://t.co/f1WWPwhw85

gov make tough choic protect arizona public health stopthespread unfortun learn mohav counti sheriff choos ignor governor risk continu spread covid arizona

Fig. 3: Sample of tweet before and after pre-processing

3.3. Feature Engineering

The pre-processed data is still inadequate for analysis process. Data must be converted into a numerical form by using feature engineering. Feature engineering mainly has two essential parts: feature extraction and feature selection (Afify, Mohammed and Hassanien, 2020). In feature extraction, features are extracted from the pre-processed data using methods such as word frequency, word importance, n-grams and so on. In this study, Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) techniques are used to extract features from the pre-processed text. BOW is a simple model that represents each word (feature) in the text as a column in matrix and the documents (Tweets) as rows where each cell in that matrix hold the frequency of word appearance in the tweet. Hence, each tweet is represented as a numerical form of the frequency of its terms instead of textual terms. Table 2 shows the BOW model of two different tweets (d1, d2) selected from the collected corpus. The frequency of tweet terms (words) can be given by the Term Frequency (TF) model, which is presented by Equation 1:

$$tf(t,d) = f_{t,d} \tag{1}$$

where *t* is the tweet terms, *d* is the document (tweet) and $f_{t,d}$ is the number of times that term *t* occurs in document *d*.

Table 2: BOW model of two sample tweets

	georgia	**confirm**	**new**	**covid**	**case**	**updat**	**today**	**death**
d_1	1	1	1	1	1	0	0	0
d_2	0	0	0	2	2	1	1	1

d1 : georgia confirm new covid case

d2: covid case updat today covid case death

However, the frequency of terms (TF) only takes term frequency into consideration which is insufficient, as it fails to describe the tweet with high frequency and gives those terms high scores while they are not important related to the whole corpus. To overcome this limitation, the Inverse Document Frequency (IDF) is used to measure inverses fraction of terms that explain the importance of terms in the whole corpus. IDF formula is given by the Equation 2:

$$\mathrm{idf}(t, \mathrm{D}) = log_{10} \frac{|D|}{|\{d \in D : t \in d\}|} \tag{2}$$

where *t* is a term, *D* denotes all documents (tweets), |*D*| is the number of documents, and $|\{d \in D : t \in d\}|$ is the number of terms *t* that appear in all documents. The $\mathrm{idf}(t, D)$ is applied on *tf* to get the final importance weights of tweet's terms as described in the Equation 3:

$$\mathrm{tf\_idf}(t, \mathrm{d}, \mathrm{D}) = tf(t, d).idf(t, D) \tag{3}$$

Subsequently, the large number of extracted features are obtained which cause high dimensionality problem. To solve this problem, the second part of feature engineering is needed to reduce the high dimensionality of features by selecting relevant features. This can be achieved by several techniques such as Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Backward Feature Elimination (BFE) and Decision Tree Ensembles (DTE). SVD and PCA has been considered as the proper choice for Euclidean space of features representation (Lee & Song, 2020). However, PCA calculate and deal with covariance matrix, which make PCA more computationally complex than SVD (Yi, Park, Chen, & Caramanis, 2016). Therefore, this study will use the SVD technique for dimensionality reduction and selecting the most important and effective features.

Singular value decomposition (SVD) is a linear-algebra method used to decompose a matrix into its fundamental factors. SVD can be applied on any matrix, and this makes it applicable to be used as an effective method to reduce irrelevant features. SVD has numerous applications in several areas such as machine learning, image processing, denoising data and it is considered as a foundation of recommendation systems (Aggarwal, 2020). SVD decomposes matrix into three basic constituent low-rank matrices. Let A be an m×n matrix, the SVD can be given by Equation 4:

$$A = U.Segma.V^T \tag{4}$$

where U is $m \times m$ orthogonal matrix contains columns set called left-singular vectors of *A*. Sigma is $m \times n$ diagonal matrix contains singular values of A as a diagonal element. V^T is a transpose of matrix *V*. *V* is $n \times n$ orthogonal matrix contains columns set called right-singular vectors of *A*.

After the main matrix is decomposed into the basic matrices, the Sigma matrix will contain the singular values of the main matrix ordered descending. These singular values are non-negative values in which it shows what the magnitudes of their corresponding basis are. It is observed that only the first (*n*) singular values (components) are large. The key idea of reducing the dimension of the matrix is to select large component values and remove the other non-valuable values. This process will reduce the number of columns of matrix Sigma. On the other side, the columns, which are removed from Sigma, will be removed also from matrices *U* and *V*. The value of n is determined through explained variance, which can be calculated by using Equation 5:

$$Explained\ Variance = \frac{(Singulare\ Value)^2}{Sum\ ((Singulare\ Values)^2)} \tag{5}$$

By reconstructing the original matrix, A back from its fundamentals reduced matrices, the dimension of A is reduced, and only high important features will be counted while the other unimportant features are removed.

3.4. Data Clustering

Clustering is a process to partition the data points into a number of groups (clusters) depending on the similarity between those data points. K-means is a powerful unsupervised algorithm that divides a number of objects into separate clusters. The main idea of K-means is to define (k) using random fixed initial points called (centroids). The distance between each centroid and all data points in the Euclidean space is calculated then the data points are assigned to the nearest centroid based on the Euclidean distance. The Euclidean distance can be calculated by using Equation 6.

$$d(x, c) = \sqrt{\sum_{i=1}^{D} (x_i - c_i)^2} \tag{6}$$

where *x* is a data point, *c* a centroid, and *D* is the total number of data points in the Euclidean space. The mean distance of each centroid and its data points are calculated and the centroid is repositioned to mean position. The whole process is started again and iterated until there are no changes happen in centroid and the calculated mean. Equation 7 shows how to assign new centroid.

$$\mathrm{c} = \frac{1}{n} \sum_{\forall x_i \in A} x_i \tag{7}$$

where *x* is the data point and *n* is the total number of data points.

The limitation of K- Means is (k) number is required to be pre-defined before K-means can be started. There are some methods that can be used to determine an optimal number of clusters (k) such as Elbow method, Silhouette Coefficient and Calinski-Harabasz index. Elbow method is a widely used method for that purpose (Renuka et al., 2021). Elbow method depends on Within Sum of Squares (WSS) value, which is the square of distance between each data point and centroids. WSS can be calculated by using Equation 8.

$$\mathrm{WSS} = \sum_{i=1}^{k} (x_i - c_i)^2 \tag{8}$$

where x is the data point, c is the centroid of that data point.

After determining the number of clusters k, the K-means is employed to cluster the tweets in the corpus into k clusters. For more illustration, t-Distributed Stochastic Neighbour Embedding (t-SNE) technique is used to visualize the result of clustering where each cluster will be plotted with certain colour. t-SNE is a powerful technique used to visualize high dimensional data in 2-dimensional or 3-dimensional space(Van der Maaten & Hinton, 2008).

The clustering result is evaluated by using Silhouette coefficient analysis. Silhouette analysis is a technique used to calculate the distance between the resulted clusters by measuring how close the data point is in one cluster to the other data points in the neighbour clusters. The plotting of silhouette analysis helps to assess clustering performance visually. The silhouette coefficient score has a range of [-1, 1] and can be calculated by Equation 9:

$$Sil\ (x) = \frac{b\ (x) - a\ (x)}{Max\ (a, b)} \tag{9}$$

where x is data point in one cluster, a is mean distance of data point x and other data points in the same cluster, and b is mean distance between the data point x and other data points in the nearest neighbour cluster. If the silhouette score is closed to 1, then the data point is well clustered and it is far from the nearest neighbour cluster. If the silhouette score is close to -1, it means that the data point is wrongly clustered in which it is far from other data points in the same cluster and close to other data points in the nearest neighbour cluster. If the silhouette score equal to 0, that means the data point is on the overlapping clusters (centre between clusters).

3.5. Sentiment Scoring

After dividing the data into clusters using K-means algorithm, the sentiment of each cluster is obtained by calculating the polarity of tweets text. Text polarity is the amount of neutrality, positivity or negativity carried by that text. For example, the word ‘good’ has a positive polarity and the word ‘bad’ has a negative polarity while the word ‘behaviour’ has a neutral polarity. Sentiment carried by text can be obtained by determining the accumulative polarity of its terms. The accumulative polarity can be the summation of polarities of each term or can be the average. The scoring process is the task of assigning polarity to certain word or document. There are some dictionary-based tools available for scoring documents (tweets) such as Affin, Vader and TextBlob. The basic idea of these tools is to assign each term in the document with a polarity score then apply an aggregating function to come out with one polarity score for the document by using sum or average. The aggregated polarity can be computed by using Equation 10.

$$(T)_{polarity} = \sum_{i=0}^{n} t_p \tag{10}$$

where t is token in tweet T, t_p is the polarity amount of token t and n is the total number of tweet tokens. In this study, TextBlob tool is used to calculate the polarity and assign a polarity score to each tweet in the corpus. TextBlob is a python library used for text processing and sentiment analysis. Later, sentiment included in the tweets is determined by using threshold value to differentiate between positive, negative and neutral sentiment. Usually, zero polarity indicates to the neutral sentiment while the polarity greater than zero indicates to the positive sentiment and the polarity less than zero for negative sentiment.

4. Experimental Results and Discussion

In this study, the experiment results are divided into five parts. The first part is related to the feature engineering results in which the value of (n) components used in the SVD technique is determined. The second part is related to the finding of k (number of clusters) which is achieved by applying Elbow method. The third part is related to the tweets clustering and labels assigning for each tweet in the corpus. In the fourth part, the sentiment for each cluster is analysed. Finally, the clustering analysis and evaluation results are presented. All experiments are written in Python programming language with Anaconda Jupiter Notebook editor and run under Windows operating system on Intel Core i7-4770 CPU @ 3.40GHz with 8 GB of RAM.

4.1. Feature Engineering Results

A total number of 12661 tweets is obtained after the data has been pre-processed. These tweets are transformed into a numerical form by using TF-IDF technique and 20200 features are acquired. To reduce the number of features dimensions, the feature selection process has been applied using SVD. The optimal dimensions can be selected based on the amount of explained variance. Typically, the percentage of selected explained variance can be chosen between (95% - 99%) (Alhowaide et al., 2020; Stella et al., 2020). The explained variance is plotted, and the cut-off point of the curve is observed. Fig. 5 shows the explained variance curve of twitter data, the x-axis represents the number of samples (Features) while the y-axis represents the explained variances ratios, which obtained by using SVD. As noted from the curve, the explained variances values are decreased sharply when the analysed samples are increased. In this study, 95% of the SVD explained variances are selected and this percentage is resulted by the summation of variance ratio of the first (6403) features as the SVD explained variances are automatically have descending order.

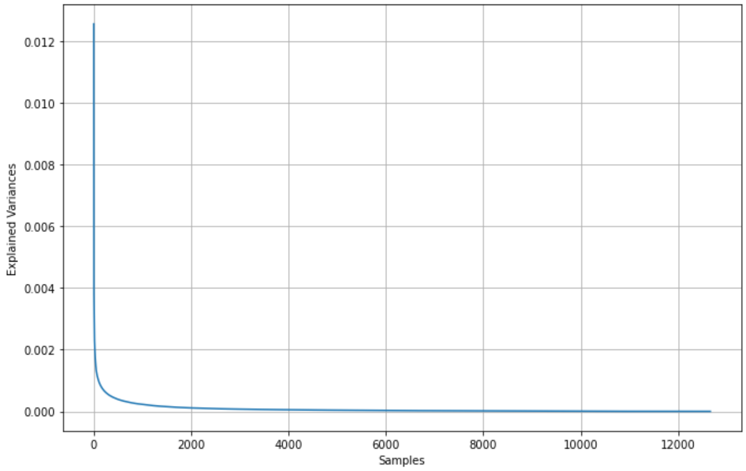


Fig. 5: The explained variances curve

4.2. Finding the Number of Clusters (k)

Prior to clustering data, the Elbow method is applied to determine the optimal number of (k). For that purpose, Yellowbrick visualization tool are used to run K-means algorithm for 20 times and each time the WSS (also called distortion score) is calculated using Equation 8. The Yellowbrick has plotted the result of the suggested values of k and the average WSS values. As shown in the Figure 6, the dashed line, which is produced by Yellowbrick, indicates that the elbow is at the cluster number 9 with distortion score (11625.711) is consider as the good choice for k. Fig. 6 shows the result of applying Elbow analysis on our data in which Yellowbrick determined the “elbow” point with a dashed line.

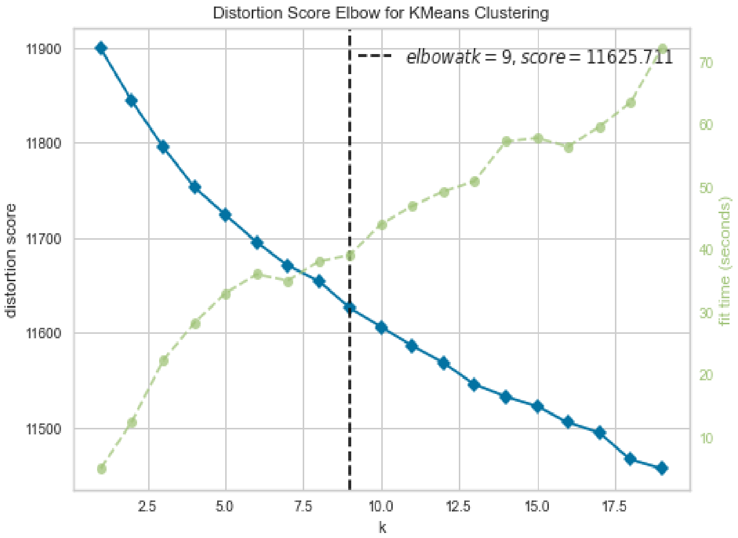
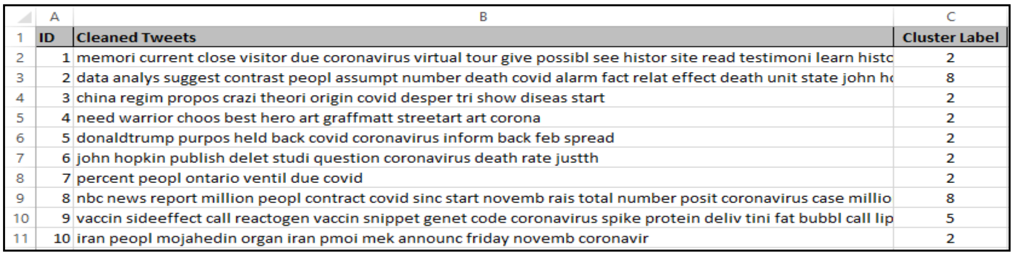


Fig. 6: The result of Elbow analysis

4.3. Data Clustering

The K-means algorithm is initialized with its default parameters values and ran with (k=9). Once the K-means is ran on our dataset, the data has been divided into 9 clusters which are labelled from (0 to 8). Fig. 7 shows the first 10 tweets in the corpus with the assigned labels. t-SNE technique is used to illustrate these clusters with coloured visualization as showed in Fig. 8.



ID	Cleaned Tweets	Cluster Label
1	memori current close visitor due coronavirus virtual tour give possibl see histor site read testimoni learn histc	2
2	data analys suggest contrast peopl assumpt number death covid alarm fact relat effect death unit state john hc	8
3	china regim propos crazi theori origin covid desper tri show diseas start	2
4	need warrior choos best hero art graffmatt streetart art corona	2
5	donaldtrump purpos held back covid coronavirus inform back feb spread	2
6	john hopkin publish delet studi question coronavirus death rate justth	2
7	percent peopl ontario ventil due covid	2
8	nbc news report million peopl contract covid sinc start novemb rais total number posit coronavirus case millio	8
9	vaccin sideeffect call reactogen vaccin snippet genet code coronavirus spike protein deliv tini fat bubbl call lip	5
10	iran peopl mojahedin organ iran pmoi mek announc friday novemb coronavir	2

Fig. 7: The first 10 tweets in the corpus with the assigned labels

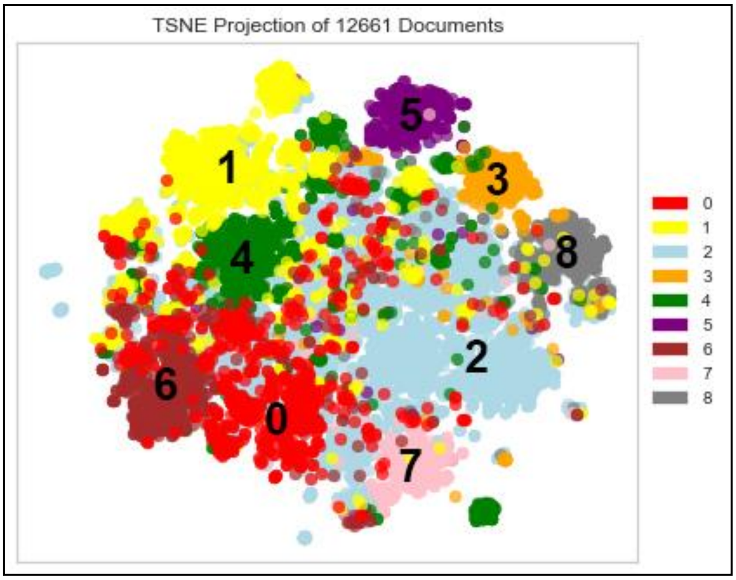
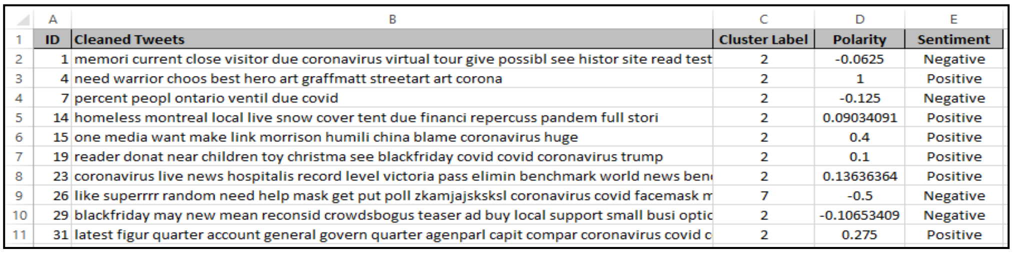


Fig. 8: The t-SNE visualization of the resulted clusters

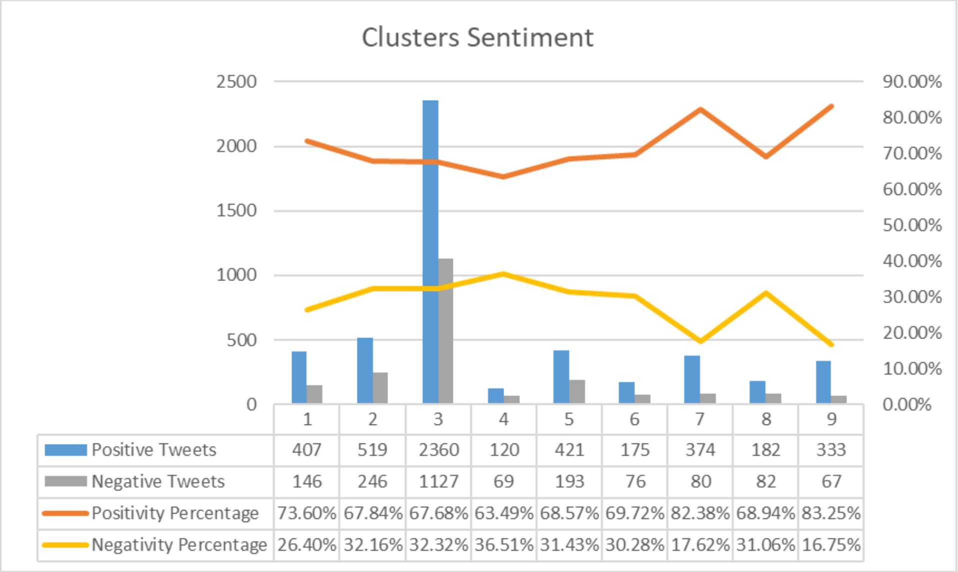
4.4. Sentiment Analysis

The tweets in the corpus are now divided into 9 clusters. To explore the sentiment of these clusters, this study ran an experiment to calculate the polarity of each tweet. TextBlob is employed to calculate the positive and negative polarity of the tweets. Each tweet has been fed into TextBlob in order to obtain its polarity score. Positive score (>0) indicates to a positive tweet, while the negative score (<0) indicates to a negative tweet, and the score (=0) indicates to a neutral tweet. According to this scoring process, the tweets has been labelled with the sentiment (Positive, Negative or Neutral). As a result, 4891/12661 (42.05%) tweets have been labelled with positive sentiment, while 2086/12661 (14.50%) tweets were labelled with negative sentiment, and 5684/12661 (43.45%) tweets have a neutral sentiment. While the neutral sentiment has no effect on the overall sentiment, this study ignored the neutral tweets and focus on the other polarity tweets. Fig. 9 illustrates the first 10 tweets in the corpus with the final results of clustering and sentiment analysis. Fig. 10 illustrates the distribution of the sentiment (Positive and Negative) over each cluster.



	A	B	C	D	E
1	ID	Cleaned Tweets	Cluster Label	Polarity	Sentiment
2	1	memori current close visitor due coronavirus virtual tour give possibl see histor site read test	2	-0.0625	Negative
3	4	need warrior choos best hero art graffmatt streetart art corona	2	1	Positive
4	7	percent peopl ontario ventil due covid	2	-0.125	Negative
5	14	homeless montreal local live snow cover tent due financi repercuss pandem full stori	2	0.09034091	Positive
6	15	one media want make link morrison humili china blame coronavirus huge	2	0.4	Positive
7	19	reader donat near children toy christma see blackfriday covid covid coronavirus trump	2	0.1	Positive
8	23	coronavirus live news hospitalis record level victoria pass elimin benchmark world news ben	2	0.13636364	Positive
9	26	like superrrr random need help mask get put poll zkamjajsksksl coronavirus covid facemask m	7	-0.5	Negative
10	29	blackfriday may new mean reconsid crowdsbogus teaser ad buy local support small busi optic	2	-0.10653409	Negative
11	31	latest figur quarter account general govern quarter agenparl capit compar coronavirus covid c	2	0.275	Positive

Fig. 9: The results of clustering and sentiment analysis

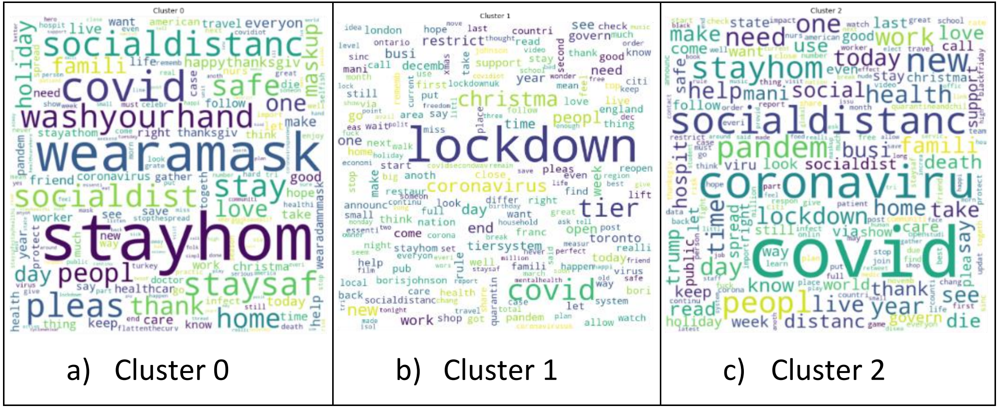


	1	2	3	4	5	6	7	8	9
Positive Tweets	407	519	2360	120	421	175	374	182	333
Negative Tweets	146	246	1127	69	193	76	80	82	67
Positivity Percentage	73.60%	67.84%	67.68%	63.49%	68.57%	69.72%	82.38%	68.94%	83.25%
Negativity Percentage	26.40%	32.16%	32.32%	36.51%	31.43%	30.28%	17.62%	31.06%	16.75%

Fig. 10: Distribution of the sentiment (Positive and Negative) over each cluster

4.5. Cluster Analysis and Evaluation

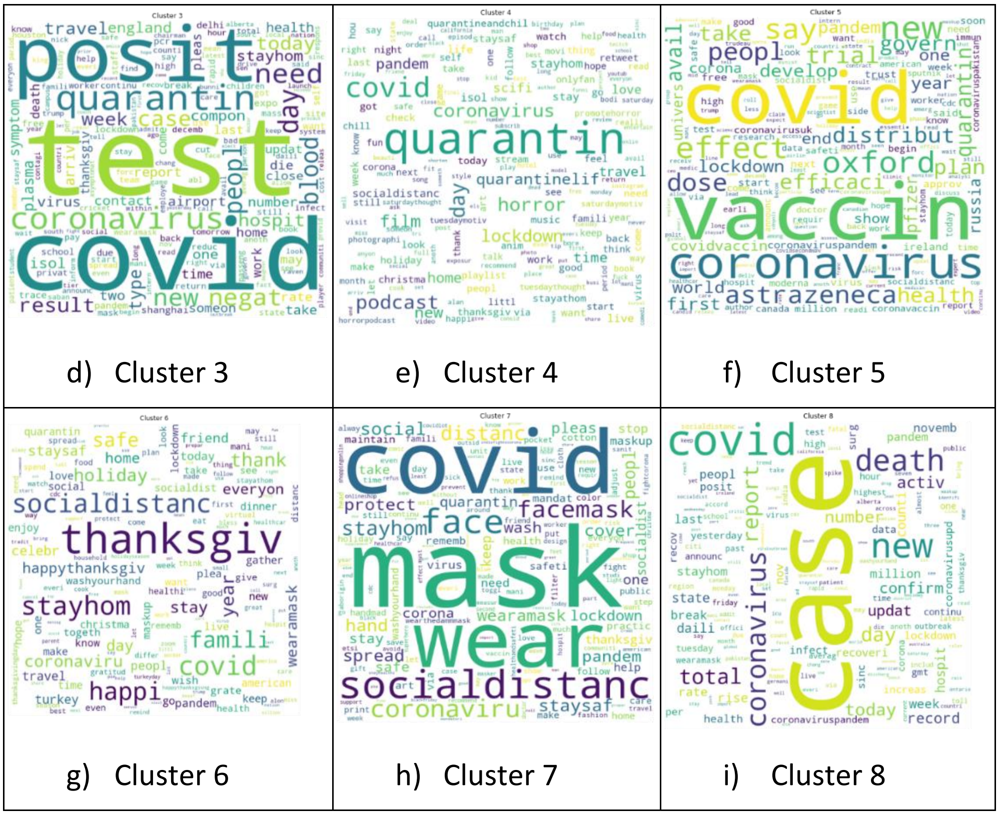
The result of K-means clustering is 9 clusters of tweets, which are labelled from 0 to 8. These tweets are grouped based on the similarity between them. As a result, there are some common topics, in which these tweets are grouped based on. The analysis of the cluster topic is done by exploring the dominant words in each cluster. In Fig. 11, each cluster and its word cloud are presented.



a) Cluster 0

b) Cluster 1

c) Cluster 2



d) Cluster 3

e) Cluster 4

f) Cluster 5

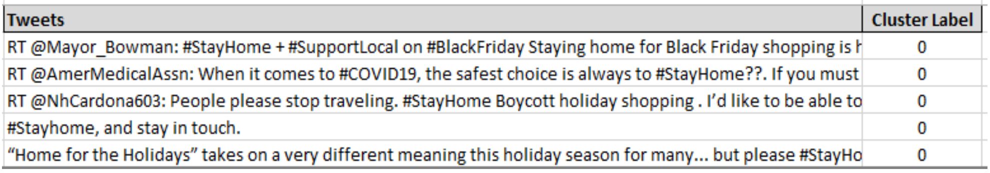
g) Cluster 6

h) Cluster 7

i) Cluster 8

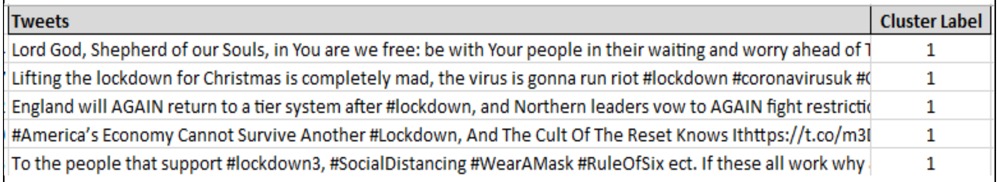
Fig. 11: Word cloud of each cluster

Cluster 0 (553 Tweets, Positive 73.60%, Negative 26.40%) discussed the rule (stay home) which are applied to limit the physical communication between people in order to control the spreading of COVID-19. Fig. 12 a) shows some tweets examples of cluster 0.



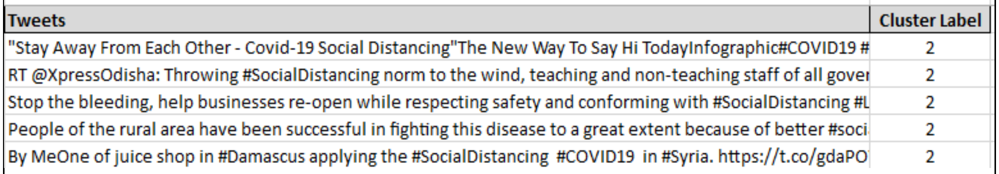
Tweets	Cluster Label
RT @Mayor_Bowman: #StayHome + #SupportLocal on #BlackFriday Staying home for Black Friday shopping is h	0
RT @AmerMedicalAssn: When it comes to #COVID19, the safest choice is always to #StayHome??. If you must	0
RT @NhCardona603: People please stop traveling. #StayHome Boycott holiday shopping . I'd like to be able to	0
#Stayhome, and stay in touch.	0
"Home for the Holidays" takes on a very different meaning this holiday season for many... but please #StayHo	0

a) Tweet samples of cluster 0



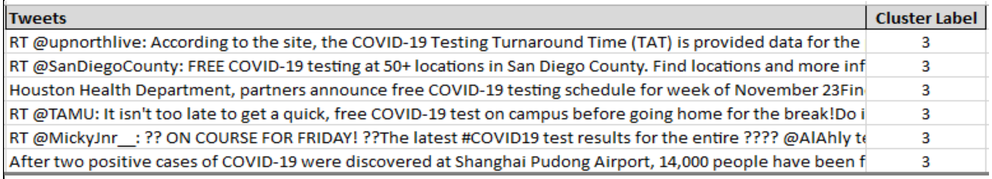
Tweets	Cluster Label
Lord God, Shepherd of our Souls, in You are we free: be with Your people in their waiting and worry ahead of T	1
Lifting the lockdown for Christmas is completely mad, the virus is gonna run riot #lockdown #coronavirusuk #(	1
England will AGAIN return to a tier system after #lockdown, and Northern leaders vow to AGAIN fight restrictio	1
#America's Economy Cannot Survive Another #Lockdown, And The Cult Of The Reset Knows Iththttps://t.co/m3I	1
To the people that support #lockdown3, #SocialDistancing #WearAMask #RuleOfSix ect. If these all work why	1

b) Tweet samples of cluster 1



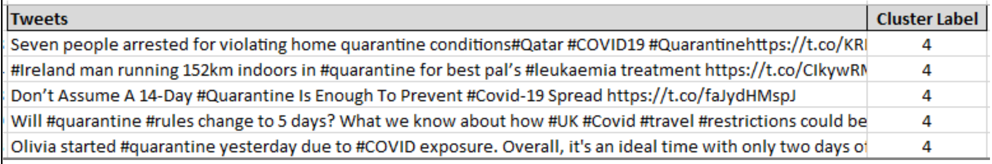
Tweets	Cluster Label
"Stay Away From Each Other - Covid-19 Social Distancing"The New Way To Say Hi TodayInfographic#COVID19 #	2
RT @XpressOdisha: Throwing #SocialDistancing norm to the wind, teaching and non-teaching staff of all gover	2
Stop the bleeding, help businesses re-open while respecting safety and conforming with #SocialDistancing #L	2
People of the rural area have been successful in fighting this disease to a great extent because of better #soci	2
By MeOne of juice shop in #Damascus applying the #SocialDistancing #COVID19 in #Syria. https://t.co/gdaPO	2

c) Tweet samples of cluster 2



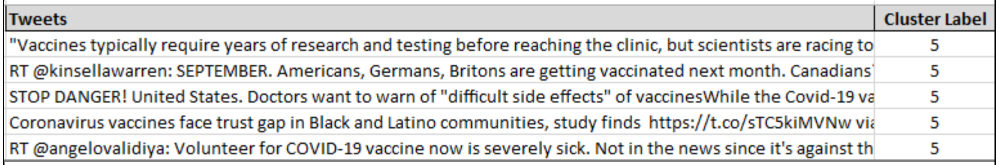
Tweets	Cluster Label
RT @upnorthlive: According to the site, the COVID-19 Testing Turnaround Time (TAT) is provided data for the	3
RT @SanDiegoCounty: FREE COVID-19 testing at 50+ locations in San Diego County. Find locations and more inf	3
Houston Health Department, partners announce free COVID-19 testing schedule for week of November 23Fin	3
RT @TAMU: It isn't too late to get a quick, free COVID-19 test on campus before going home for the break!Do i	3
RT @MickyJnr__: ?? ON COURSE FOR FRIDAY! ??The latest #COVID19 test results for the entire ???? @AlAhly te	3
After two positive cases of COVID-19 were discovered at Shanghai Pudong Airport, 14,000 people have been f	3

d) Tweet samples of cluster 3



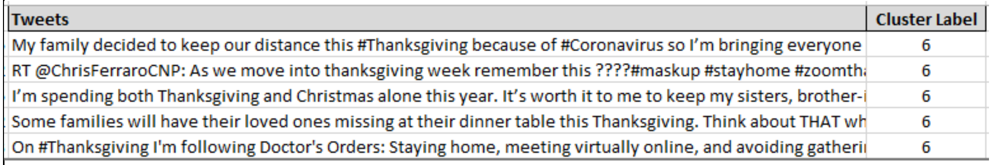
Tweets	Cluster Label
Seven people arrested for violating home quarantine conditions#Qatar #COVID19 #Quarantinehttps://t.co/KRI	4
#Ireland man running 152km indoors in #quarantine for best pal's #leukaemia treatment https://t.co/ClkywRN	4
Don't Assume A 14-Day #Quarantine Is Enough To Prevent #Covid-19 Spread https://t.co/faJydHMspJ	4
Will #quarantine #rules change to 5 days? What we know about how #UK #Covid #travel #restrictions could be	4
Olivia started #quarantine yesterday due to #COVID exposure. Overall, it's an ideal time with only two days of	4

e) Tweet samples of cluster 4



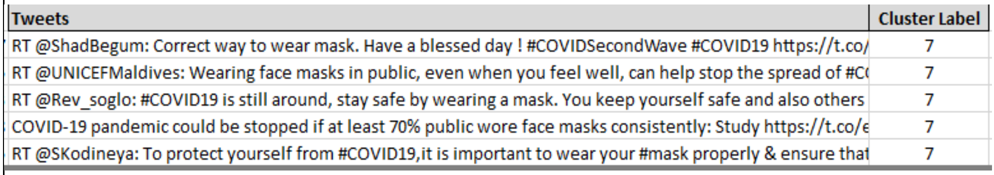
Tweets	Cluster Label
"Vaccines typically require years of research and testing before reaching the clinic, but scientists are racing to	5
RT @kinsellawarren: SEPTEMBER. Americans, Germans, Britons are getting vaccinated next month. Canadians	5
STOP DANGER! United States. Doctors want to warn of "difficult side effects" of vaccinesWhile the Covid-19 va	5
Coronavirus vaccines face trust gap in Black and Latino communities, study finds https://t.co/sTC5kiMVNw via	5
RT @angelovalidiya: Volunteer for COVID-19 vaccine now is severely sick. Not in the news since it's against th	5

f) Tweet samples of cluster 5



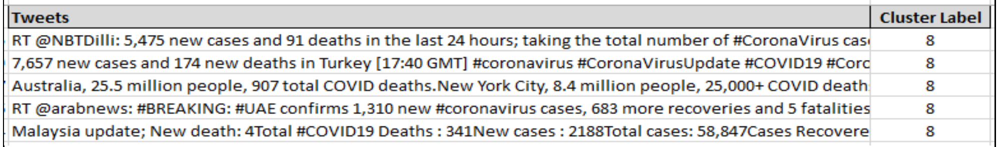
Tweets	Cluster Label
My family decided to keep our distance this #Thanksgiving because of #Coronavirus so I'm bringing everyone	6
RT @ChrisFerraroCNP: As we move into thanksgiving week remember this ????#maskup #stayhome #zoomth	6
I'm spending both Thanksgiving and Christmas alone this year. It's worth it to me to keep my sisters, brother-i	6
Some families will have their loved ones missing at their dinner table this Thanksgiving. Think about THAT wh	6
On #Thanksgiving I'm following Doctor's Orders: Staying home, meeting virtually online, and avoiding gatheri	6

g) Tweet samples of cluster 6



Tweets	Cluster Label
RT @ShadBegum: Correct way to wear mask. Have a blessed day ! #COVIDSecondWave #COVID19 https://t.co/	7
RT @UNICEFMaldives: Wearing face masks in public, even when you feel well, can help stop the spread of #C(	7
RT @Rev_soglo: #COVID19 is still around, stay safe by wearing a mask. You keep yourself safe and also others	7
COVID-19 pandemic could be stopped if at least 70% public wore face masks consistently: Study https://t.co/e	7
RT @SKodineya: To protect yourself from #COVID19,it is important to wear your #mask properly & ensure that	7

h) Tweet samples of cluster 7



Tweets	Cluster Label
RT @NBTDilli: 5,475 new cases and 91 deaths in the last 24 hours; taking the total number of #CoronaVirus cas	8
7,657 new cases and 174 new deaths in Turkey [17:40 GMT] #coronavirus #CoronaVirusUpdate #COVID19 #Corc	8
Australia, 25.5 million people, 907 total COVID deaths.New York City, 8.4 million people, 25,000+ COVID death	8
RT @arabnews: #BREAKING: #UAE confirms 1,310 new #coronavirus cases, 683 more recoveries and 5 fatalities	8
Malaysia update; New death: 4Total #COVID19 Deaths : 341New cases : 2188Total cases: 58,847Cases Recovere	8

i) Tweet samples of cluster 8

Fig. 12: Tweet samples for each cluster

Cluster 1 (765 Tweets, Positive 67.84%, Negative 32.16%) has discussed the lockdown procedure. Many countries have applied this procedure in multiple tiers. Fig. 12 b) shows some of the tweets belong to this cluster. Cluster 2 (3487 Tweets, Positive 67.68%, Negative 32.32%) is related the rule of (social distance). This rule enforces people to take a safety distance of one meter at least and it claims that prevent spreading virus among people. Fig. 12 c) shows some tweets examples of this cluster. Cluster 3 (189 Tweets, Positive 63.49%, Negative 36.51%) debates on the COVID-19 testing procedure. Fig.12 d) shows some tweets examples of this cluster.

Cluster 4 (614 Tweets, Positive 68.57%, Negative 31.43%) has discussed the quarantine procedure. This procedure applied mostly on passengers and for those people who have physical contacting with other people which have positive COVID-19. Fig. 12 e) shows some tweets examples in this cluster. Cluster 5 (251 Tweets, Positive 69.72%, Negative 30.28%) has discussed the COVID-19 vaccine. Fig. 12 f) shows some tweets examples belong to this cluster. Cluster 6 (454 Tweets, Positive 82.38%, Negative 17.62%) has discussed the thanksgiving celebration with the changing in the life due to the rules of COVID-19. Some tweets examples of this cluster is shown in Fig. 12 g). Cluster 7 (264 Tweets, Positive 68.94%, Negative 31.06%) has discussed the procedure of wearing mask and its impact on controlling the COVID-19 cases. Fig. 12 h) shows some tweets examples of this cluster. Cluster 8 (400 Tweets, Positive 83.25%, Negative 16.75%) has discussed the reports of the confirmed recovery, new and death cases of COVID-19. examples of some tweets presented this cluster is depicted in Fig. 12 i). The clustering results that are obtained have been evaluated by using Silhouette coefficient analysis and has achieved coefficient average (silhouette score) of 0.0070. Fig. 13 shows the silhouette coefficient for each cluster.

In Fig.13, the dotted red vertical line indicates to the average of silhouette score obtained by clustering process. Each cluster has been showed with different colour. The height of each cluster indicates to the amount of data points belongs to that cluster. The width indicates to the amount of silhouette score obtained by that cluster, in which implies also whether the cluster is clustered well or not. It is notable that the cluster 2 has the highest number of data points and almost all those data points achieved negative silhouette score (< 0.0) which indicates that the data points are not clustered well. On the other hand, the cluster 0, cluster 1 and cluster 4 have average amount of data points which most of them achieved low positive silhouette score (>0.0) while few of them achieved negative silhouette score (<0.0). This indicates that those clusters almost overlapped with their neighbour clusters. Finally, the cluster 3, cluster 5, cluster 6 and cluster 8 have average amount of data points and achieved higher positive silhouette score (> 0.0) than the other clusters which indicates that they are clustered well. In general, the average silhouette score of all clusters was 0.0070.

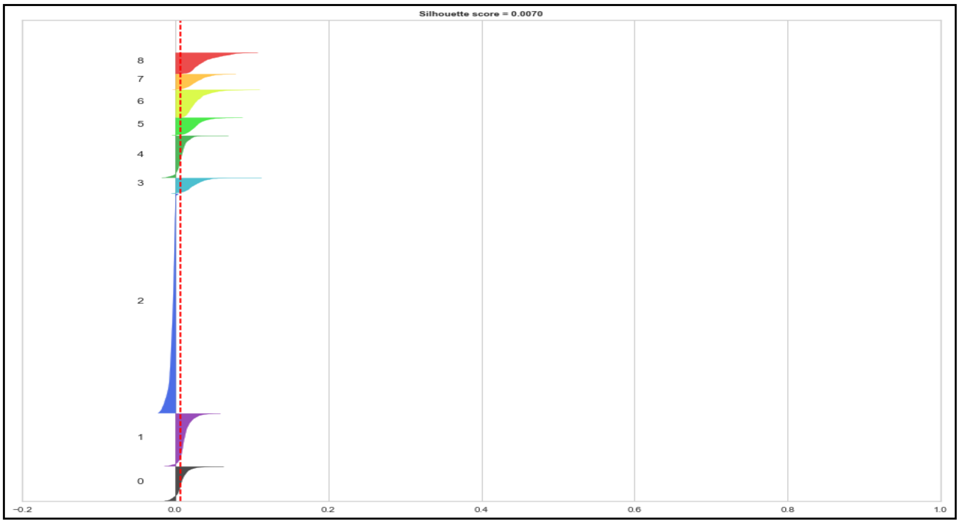


Fig. 13: Silhouette coefficient analysis

From the above evaluation of the results, it is clear that the clustering process has achieved low silhouette score (0.0070). This score indicates that the data points are clustered, but the clusters are almost overlapped. The main reason of this result related to the challenges of unstructured short textual data analysis, such as the data representation, weighting scheme, high dimensionality, sparse feature vector and the word synonymy. Moreover, there are some challenges faced by the clustering technique of textual data such as the similarity measures, determine the optimal centres and the huge number of outliers.

It is important to choose the suitable data representation and weighting schema because the weighting of features can make the data points become similar and it is difficult to put them in different clusters with reasonable distance. As seen in cluster 2, there is a high number of data points that are not clustered well this can be due to the issue of data representation and weighting the features as well as the issue of the Euclidean measure which is used by K-Means to cluster the data points. In addition, the cluster 0 has discussed the rule of “wearing mask” which is discussed mainly in cluster 7 with the feature “mask”. This occurred due to the issue of the word synonymy challenge of textual data that two terms have same meaning. Furthermore, the TF-IDF weighting process resulted a very sparse feature matrix, which most of the vector values are zeros (Yoon & Joung, 2020) . This is due to the high dimensionality of processed features and short length of tweets and to overcome this issue the suitable feature selection technique must be chosen carefully.

5. Conclusion and Future Works

Clustering analysis on social media data can offer a better understanding of issues discussed around the globe and illustrate the correlation between data elements. Fast spread of COVID-19 makes it the most discussed topic in social networks. Cluster analysis on COVID-19 outbreak can be incredibly helpful for decision makers and government in monitoring and planning proper plans. Sentiment analysis is a valuable way to figure out people's opinions and feelings towards events or other aspects. In this study, cluster analysis on COVID-19 outbreak using K-means algorithm is presented. The Twitter data related to COVID-19 is crawled and represented using TF-IDF technique as a weighting schema. The K-means algorithm with Euclidean distance measure is used to cluster the data. In addition, this study discovered sentiment towards COVID-19 by applying sentiment analysis for each cluster using TextBlob lexicon-based tool. With the use of data visualization tools such as t-SNE and Word Cloud, the results of the analysis are illustrated. The results of the experiments showed that there are relatively 9 clusters that are obtained with different topics ranging with highest score of 83.25% positivity and 16.75% of negativity being reported. Dominant topics are explored using word cloud and the clustering results have been evaluated with 0.0070 Silhouette coefficient. This study also has discovered several challenges in the clustering of unstructured textual data in terms of the high dimensionality, data sparsity and data representation. The study suggests in using different data representation such as Doc2Vec (Document to Vector) and Word2Vec (Word to Vector) techniques in order to obtain higher results. In addition, the study suggests using different similarity measures with K-Means like cosine distance. As the amount of data in social networks increase tremendously and cause issues such as high dimensionality, thus there is a need to develop new and adaptive methods to cope with this huge amount of data.

Acknowledgement

This research is supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2020/ICT02/UUM/02/3).

References

Aggarwal, C. C. (2020). Singular Value Decomposition *Linear Algebra and Optimization for Machine Learning* (pp. 299-337): Springer.

Ahmed, M. E., Rabin, M. R. I., & Chowdhury, F. N. (2020). COVID-19: Social Media Sentiment Analysis on Reopening. *arXiv preprint arXiv:2006.00804*.

AL-Sharuee, M. T., Liu, F., & Pratama, M. (2018). Sentiment analysis: an automatic contextual analysis and ensemble clustering approach and comparison. *Data & Knowledge Engineering, 115,* 194-213.

Alhowaide, A., Alsmadi, I., & Tang, J. (2020). *PCA, Random-Forest and Pearson Correlation for Dimensionality Reduction in IoT IDS.* Paper presented at the 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS).

Boon-Itt, S., & Skunkan, Y. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance, 6*(4), e21978.

Contratres, F. G., Alves-Souza, S. N., Filgueiras, L. V. L., & DeSouza, L. S. (2018). *Sentiment analysis of social network data for cold-start relief in recommender systems.* Paper presented at the World Conference on Information Systems and Technologies.

Cruickshank, I. J., & Carley, K. M. (2020). Characterizing communities of hashtag usage on twitter during the 2020 COVID-19 pandemic by multi-view clustering. *Applied Network Science, 5*(1), 1-40.

El-Din, D.M., Hassanein, A.E., Hassanein, E.E. and Hussein W.M.E. E-Quarantine (2020): A Smart Health System for Monitoring Coronavirus Patients for Remote Quarantine. Journal of System and Management Sciences 10 (4), 102-124.

Gupta, R. K., Vishwanath, A., & Yang, Y. (2020). COVID-19 Twitter Dataset with Latent Topics, Sentiments and Emotions Attributes. *arXiv preprint arXiv:2007.06954*.

Gupta, S., & Arora, S. (2016). A hybrid firefly algorithm and social spider algorithm for multimodal function *Intelligent Systems Technologies and Applications* (pp. 17-30): Springer.

Imran, A. S., Doudpota, S. M., Kastrati, Z., & Bhatra, R. (2020). Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning--a Case Study on COVID-19. *arXiv preprint arXiv:2008.10031*.

Kabir, M. Y., & Madria, S. (2020). CoronaVis: A real-time COVID-19 tweets data analyzer and data repository.

Kemp, S. (2021). Digital 2021: Global Overview Report. from https://datareportal.com/reports/digital-2021-global-overview-report

Lee, S., & Song, B. C. (2020). Transformation of Non-Euclidean Space to Euclidean Space for Efficient Learning of Singular Vectors. *IEEE Access, 8*, 127074-127083.

Liu, D., Wang, Y., Wang, J., Liu, J., Yue, Y., Liu, W., & Wang, Z. (2020). Characteristics and Outcomes of 599 Patients Infected with COVID-19 in Wuhan, China: Based on an Online Reported Sample. *Journal of Medical Internet Research*.

Manguri, K. H., Ramadhan, R. N., & Amin, P. R. M. (2020). Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks. *Kurdistan Journal of Applied Research,* 54-65.

Medford, R. J., Saleh, S. N., Sumarsono, A., Perl, T. M., & Lehmann, C. U. (2020). *An "infodemic": leveraging high-volume Twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak.* Paper presented at the Open forum infectious diseases.

Moutidis, I., & Williams, H. T. (2020). Good and bad events: combining network-based event detection with sentiment analysis. *Social Network Analysis and Mining, 10*(1), 1-12.

Narasamma, V. L., Sreedevi, M., & Vijay Kumar, G. (2020). TweetShort Text Data Analysis on COVID-19 Out Break. *International Journal of Advanced Science and Technology, 29(7), 10013-10021.*

Nemes, L., & Kiss, A. (2020). Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication,* 1-15.

Orkphol, K., & Yang, W. (2019). Sentiment analysis on microblogging with K-means clustering and artificial bee colony. *International Journal of Computational Intelligence and Applications, 18*(03), 1950017.

Ou, X., Cao, Y., and Mu, X. (2015). Classification of Sentiment Sentences Based on Naive Bayesian Classifier. Journal of Logistics, Informatics and Service Science, 2(1), 48-58.

Pokharel, B. P. (2020). Twitter sentiment analysis during covid-19 outbreak in nepal. *Available at SSRN 3624719.*

Renuka, S., Kiran, G. R., & Rohit, P. (2021). An Unsupervised Content-Based Article Recommendation System Using Natural Language Processing *Data Intelligence and Cognitive Informatics* (pp. 165-180): Springer.

Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2020). Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers. *Future Generation Computer Systems, 106,* 92-104.

Samuel, J., Ali, G., Rahman, M., Esawi, E., & Samuel, Y. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. *Information, 11*(6), 314.

Shchoholiev, M., Andriichuk, O., Tsyganok, V., & Tretynyk, V. (2021). *Decision-making and computational linguistic tools application for overall estimation of the level of social tension.* Paper presented at the Journal of Physics: Conference Series.

Stella, A., Bonnier, F., Tfayli, A., Yvergnaux, F., Byrne, H. J., Chourpa, I., . . . Tauber, C. (2020). Raman mapping coupled to self-modelling MCR-ALS analysis to estimate active cosmetic ingredient penetration profile in skin. *Journal of Biophotonics, 13*(11), e202000136.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(11).

Venigalla, A. S. M., Chimalakonda, S., & Vagavolu, D. (2020). *Mood of India During Covid-19-An Interactive Web Portal Based on Emotion Analysis of Twitter Data.* Paper presented at the Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing.

Wang, T., Lu, K., Chow, K. P., & Zhu, Q. (2020). COVID-19 Sensing: Negative sentiment analysis on social media in China via Bert Model. *IEEE Access, 8,* 138162-138169.

Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using Latent Dirichlet Allocation for topic modeling on Twitter. *PloS one, 15*(9), e0239441.

Yi, X., Park, D., Chen, Y., & Caramanis, C. (2016). *Fast algorithms for robust PCA via gradient descent.* Paper presented at the Advances in neural information processing systems.

Yin, H., Yang, S., & Li, J. (2020). *Detecting topic and sentiment dynamics due to Covid-19 pandemic using social media.* Paper presented at the International Conference on Advanced Data Mining and Applications.

Yoon, J and Joung, S. (2020). A Big Data Based Cosmetic Recommendation Algorithm. Journal of System and Management Sciences, 10 (2), 40-52.Yu, X., Zhong, C., Li, D., & Xu, W. (2020). *Sentiment analysis for news and social media in*

COVID-19. Paper presented at the Proceedings of the 6th ACM SIGSPATIAL International Workshop on Emergency Management using GIS.