CRISE: Toward Better Understanding of COVID-19 Psychological Impact during Lockdown

Hussein Baalbaki[1], Hassan Harb[1,2], Chamseddine Zaki[2], Youssef Alakoury[3], Layla Tannoury[1], Michel Nabaa[1]

[1]Faculty of Sciences, Lebanese University, Beirut, Lebanon
[2]Faculty of Business Administration and Economics, Lebanese University, Rashaya, Lebanon
[3]Faculty of Letters and Human Sciences, Lebanese University, Zahle, Lebanon
Email: hsinbaalbaki@gmail.com, hassan.harb.1@ul.edu.lb, c.zaki@ul.edu.lb,
youssef.alakoury@ul.edu.lb, layla.tannoury@ul.edu.lb, michel.nabaa@ul.edu.lb

Abstract

Recently, the COVID-19 emerged in China and propagated around all the world has threatened millions of people and affected most countries and governments at several sides such as economical, educational, tourism, healthcare, etc. Indeed, one of the most important challenges that directly affect the people is the psychological side due to the harsh policies imposed by public authorities in most countries. In this paper, we propose a framework called CRISE that allows studying and understanding the psychological effect of COVID-19 during the lockdown period. Mainly, CRISE consists of four data stages: Collection, tRansformation, reductIon, and cluStEring. The first stage collects data from more than 2000 participants through a questionnaire containing attributes related to psychological effect before and during the lockdown. The second stage aims to preprocess the data before performing the study stage. The third stage proposes a model that finds the similarities among the attributes, based on the correlation matrix, to reduce its number. Finally, the fourth stage introduces a new version of Kmeans algorithm, called as Jaccard-based Kmeans (JKmeans), that allows to group participants having similar psychological situation in the same cluster for a later analysis. We show the effectiveness of CRISE in terms of clustering accuracy and understanding the psychological effect of COVID-19.

Keywords

COVID-19, Mental Health, Data Clustering, Similarity Function, Data Reduction, Correlation Matrix

1. Introduction

The importance of the public health sector comes from its role in protecting all other sectors from hazards such as infectious disease, outbreaks, and natural disasters. According to World Health Organization (WHO) [1], well functioning health care system requires several factors including a financing mechanism, a well trained and adequately paid workforce, reliable information to base decisions and policies on, and well maintained health facilities to deliver quality medicines and technologies. As a result, governments dedicate huge budgets for the sake of improving their health care systems functionality and increase the life expectancy. Mainly, such budgets help to support the healthcare industries (e.g. manufacturing medical equipment and drugs), to provide medical insurance for the largest possible population, to improve both patient-care capabilities and its potential for advanced clinical research. In addition to the local health care strategy, international organizations, such as WHO, work on wider scale and having a primary role in managing the international health within the united nations system. This includes the monitoring of the public health risks, the coordination of responses when health emergencies are detected, and the promotion of human health and well being.

Unfortunately, the emergence of the COVID-19 has clearly shown the weakness of the public health systems in both local and international levels, despite the massive budgets dedicated to build such systems. Due to its rapid spread, it affects more than 35 million people and causes the deaths of 1 million until the end of September, 2020. Consequently, the WHO declared the outbreak of COVID-19 as a global pandemic on 30 January 2020. Furthermore, COVID-19 has caused major impacts at both country and people levels. On one hand, the pandemic has tremendously affect several sectors, especially economic and business, that leads, according to UNDP, to lose more than $42 billion of gross domestic product of countries and to increase the unemployment rate by 1.2%. On the other hand, the impact of the pandemic at the people level and their daily lives is extensive and has far reaching consequences. Particularly, one of the most critical effects of COVID-19 is the mental health and psychological impacts on population as a result of the strict policies and lockdown adopted in most countries. Thus, several psychological problems related to COVID-19 have emerged progressively during the lockdown and quarantine period such as stress, emotional disturbance, mood alterations, anxiety, depression, frustration, insomnia, and uncertainty. Therefore, studying and evaluating the mental health of population is becoming essential to fight against the psychological impact of COVID-19 during the lockdown.

Indeed, the psychological reactions of COVID-19 pandemic are not the same for all countries and may differ from one person to another depending on several conditions such as age, income, culture, gender, job, etc. In this paper, we propose a framework called CRISE aiming to study the population mental health before and after the emergence of COVID-19. CRISE consists of four layers for

data studying and analyzing. The first and second stages aim to collect the data related to population mental health and preprocess them before the analysis process. The third stage aims to reduce the size of the collected data by introducing a model based on the correlation matrix. The last stage allows to grouping persons having similar mental health situations in the same clusters based on a new version of Kmeans algorithm, referred to as JKmeans. We evaluated the effectiveness of CRISE based on real collected data from more than 2000 persons while the results show an efficient clustering accuracy and clear understating of the pandemic psychological effect.

The remainder of this paper is organized as follows. Section 0 presents an overview about techniques proposed for studying the impacts of COVID-19. In section 0, we present the CRISE framework while detailing each data stage. Section 0.0.2 exposes the simulation and the discussion of the obtained results. Finally, section 3 concludes the paper and gives directions for future work.

2. Related Work

COVID-19 is bringing the world's scientist and health professional together to accelerate the research and development process to fight against this new threat. Consequently, the number of published articles and preprints in the first six months of 2020 increased up to 50% compared to 2019 where most of them are dedicated to analysis the effect and understand the behavior of this pandemic [2]. In addition, new open access journals have been launched to review COVID-19 preprints on a fast-tracked timeline while the peer-review process of articles related to this topic are accelerated 8 times than articles on other topics [3]. In such works, researchers are focused on studying the impacts of COVID-19 on several sides such as psychological [4], healthcare [5], economical [6], business [7], tourism [8], educational [9], etc. They aim to understand, analyze and predict the spread and behavior of COVID-19.

Unfortunately, from the psychological impact point of view, most of the proposed techniques prove that COVID-19 posed a negative effect on the mental health of the population. Thus, the spread of the Corona virus has led to high levels of anxiety and tension, insomnia and interrupted sleep hours, increased or loss of appetite, high rates of family violence, an increase in the number of divorces and suicide [10] [11] [12] [13] [14]. These effects can turn into symptoms psychiatric illness, such as generalized anxiety disorder, obsessive-compulsive disorder, major depressive disorder, and post-traumatic stress disorder.

In [15], the authors examine the number of respondents who experienced a clinically significant change in psychopathological symptom levels from pre- to post-outbreak assessment or significant levels of COVID-19-related traumatic distress. The psychological services and techniques and the computer-aided psychotherapy had to react fast to the psychological problems that arose [16]. The contribution of the psychological interventions therapy like EMDR (Eye Movement Desensitization and Reprocessing) [17] is considered as an "evidence-based

intervention in adults as early intervention, in the acute phase". Then, it produces effects on the clinical reduction of symptoms and on the improvement of the functioning/quality of life. Thus, EMDR is an integrative method where patients take confidence in self-healing ability and reduce anxiety [18]. The authors of [19] [20] show the efficiency of computer-aided psychotherapy that is a computerized information technology that used patient input to make some psychotherapy decisions [19] [20]. Thus, computer-aided psychotherapy marks the possibility of applying data processing to solve psychological cases [21].

In [22], the authors surveyed the population in China to study the levels of psychological impact, depression, anxiety and stress during the initial phase of the COVID-19 outbreak. The survey included 1210 participants from various Chinese cities and it concluded that half of the participants reported the mental health as moderate-to-severe while one-third rated moderate-to-severe anxiety. The authors of [23] aimed to understand the impact of social distancing and isolation on the people mental health in Veitnam. They introduced a snowball sampling technique for selecting the participants and used the event scale-revised method to assessing the psychological impact of the pandemic. The results show that most participants reported moderate to extreme psychological conditions while the most affected people are female above 44 years old or having a higher number of children, persons having neighbors affected by the pandemic, the elderly persons, health care professionals. In [24], the authors studied the anxiety level caused by the pandemic and its impact on the university students in Malaysia. A survey dedicated to 983 students show that female gender, students having age below 18 years, pre-university level of education, and management studies were significantly associated with higher levels of anxiety.

Unfortunately, although the proposed mechanisms and techniques carry many advantages and study the impact of COVID-19 on several sides especially psychological side, but they mainly suffer from several drawbacks: 1) the number of respondents/participants is of hundreds, but not thousands, level; 2) the proposed techniques are mostly limited to one analysis model while our mechanism proposes a framework with several phases and many analytical models; 3) they do not mostly offer an analysis study for the psychological effect of participants; 4) the accuracy of the proposed techniques are generally of low level due to the limited number of participants.

3. CRISE Mechanism

Unfortunately, the confirmed cases by COVID-19 are still rapidly increasing in most countries due to the absence of vaccines until today. Thus, the psychological impact will continue to affecting additional millions of people around the world and drastically change their mental health. Subsequently, the psychological impact can have several risk levels depending on the person situation if it is confined at home, quarantined, has neighbors or friends affected by COVID-19, etc. Hence, it is important to understand the mental health of different popula-

tions and countries to provide interventions and government policies that reduce the psychological effect of the global epidemic. In this section, we introduce a new framework called CRISE that allows governments to be able to understand the psychological effect of the COVID-19 outbreak on their populations, then to design strategies and policies to reduce their impact. Heavily, CRISE uses data science techniques to classify persons that similarly affected by the pandemic and it is consisted on four stages described in the next sections. Figure 1 shows the architecture of the CRISE framework with various stages and algorithms used in each one.

3.1. Data Collection Stage

The first stage of our framework aims to collect data related to psychological impact of the COVID-19 pandemic. For this reason, we created a web-based survey composed of 22 questions launched between Mars and April, 2020. For the sake of ensuring the user's privacy, the survey was respected the anonymous of the participants, without requesting any information related to the real identity, and it was announced through the social media and university systems. The survey was addressed to Lebanese population in various governorates where more than 2000 (e.g. N) persons were participated in this study. The questions of the survey were carefully formulated and divided into three main sections:

- *General questions*: consists of 7 questions referred to the participant gender, age, educational level, major, hobbies, political and religion situations.
- *Psychological-based questions before the COVID-19 pandemic*: this type of questions focused on the psychological and physical health of the participants before the outbreak of the coronavirus. They ask information about mental health and physical symptoms appeared before the emergence of virus with their severity, e.g. between 1 and 5, along with the hobbies performed to overcome such symptoms.
- *Psychological-based questions after the COVID-19 pandemic*: the questions in this section are similar to those of the previous one however they aim to understand the participant situation after the lockdown period and the harsh policies settled by most governments. Consequently, we aimed to verify the new symptoms appeared with the virus and understand their severity.
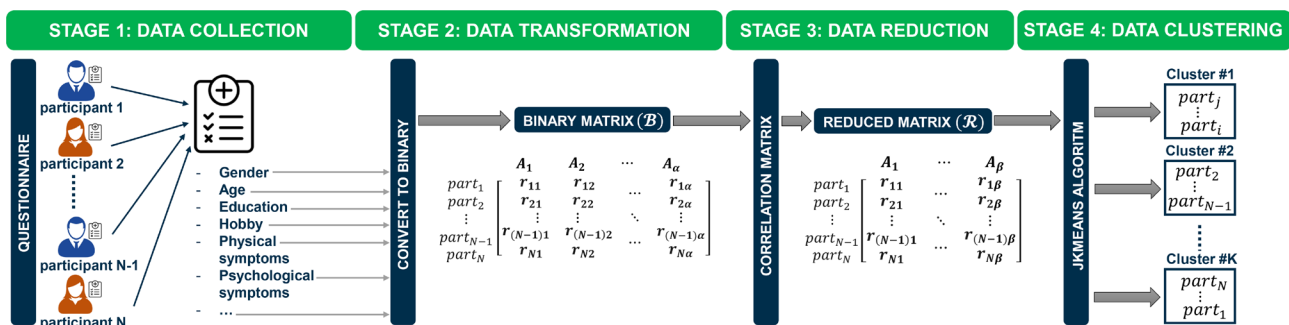

Figure 1. Architecture of CRISE framework.

Indeed, the traditional way of survey's construction mostly contains multiple-choice questions that force the participants to select their answers among a set of predefined words. Unfortunately, such way of data collection is not accurate, especially in the case of new disease (e.g. COVID-19), because the survey administrator may not be able to estimate all the participant reactions. Hence, our survey focused on free-text questions where the participants may express their actual situations using their own words. For instance, in the hobby field, the participant can insert a single or multiple words expressing their preferences. Although such data insertion can increase the complexity of the survey analysis, it allows a better understanding and deep comprehension of the psychological impact of the pandemic.

3.2. Data Transformation Stage

After the data being collected, the second stage of CRISE framework aims to preprocess the data before performing analysis task. The objective of the preprocessing is to cleaning and filtering the data by removing redundancy and useless information that may complicate the mission of the decision makers. Mainly, the data preprocessing task in the data transformation stage follows two main steps:

3.2.1. Feature Extraction

Mathematically, we formulate the analysis problem of the COVID-19 psychological impact as follows: given the survey $\mathcal{S}$ consisting of a set of attributes $\mathcal{A} = \{A_1, A_2, \cdots, A_{22}\}$, where each attribute indicates a column. Then, we define the set of participants $\mathcal{P} = \{P_1, P_2, \cdots, P_N\}$, where N is the total number of participants. Each participant $P_i \in \mathcal{P}$ has its own set of records $R_i = \{r_1, r_2, \cdots, r_{22}\}$ correspond to the attribute list $\mathcal{A}$, where r_i is the data related to the attribute A_i. Therefore, our objective is to study and analyze the data sets in $\mathcal{S}$ using clustering approach in order to understand the behavior of psychological impact of COVID-19 in the participants.

Indeed, due to the advanced nature of our survey, a participant may insert several values for each record $r_i \in R_i$, which makes the analysis of participants as non-trivial task. Hence, we propose to extract the set of words $W_i = \{w_1, w_2, \cdots, w_\gamma\}$ that define each attribute A_i based on the **Algorithm 1**. Basically, the algorithm takes, as input, the set of attributes with their record sets and returns a list of short word set for each attribute, e.g. $\mathcal{W}$. The process starts by creating a word set for each attribute (lines 2-3), then we split each record inserted by a participant into several words. This is done by applying a function called split() that returns a list of words separated using a separator, for instance comma or space, defined by the user. Thus, only the unique and not repeated words are added to the final list of the attributes (lines 6-11).

3.2.2. Data Binary Conversion

In this step, we aim to uniform the various types of data in $\mathcal{S}$. Since the

---

Algorithm 1 Word Extraction Algorithm.

Require: Survey: $\mathcal{S}$; Set of attributes: $\mathcal{A}$; Participant number: N.
Ensure: List of word sets: $\mathcal{W} = \{W_1, W_2, \ldots, W_{22}\}$.
  1: $\mathcal{W} \leftarrow \emptyset$
  2: **for** each attribute $A_i \in \mathcal{A}$ **do**
  3:     $W_i \leftarrow \emptyset$
  4:     **for** $j = 1$ to N **do**
  5:       $T \leftarrow \text{split}(r_i)$
  6:       **for** each $w_k \in T$ such that $w_k \notin W_i$ **do**
  7:         $W_i \leftarrow W_i \cup \{w_k\}$
  8:       **end for**
  9:     **end for**
10: **end for**
11: $\mathcal{W} \leftarrow \mathcal{W} \cup \{W_i\}$
12: **return** $\mathcal{W}$

---

attributes are expressed in different data types (numerical, text, binary, etc.), it is important to convert them into the same type to be able to study the similarity between them. Hence, we propose a binary conversion algorithm that allows to transform the data in $\mathcal{S}$ into the binary form. **Algorithm 2** shows the binary conversion process that takes the attributes of $\mathcal{S}$ and convert it into a binary matrix $\mathcal{B}$. The idea behind our algorithm is to convert each attribute $A_i \in \mathcal{A}$ into a set of attributes according to its word number in W_i (lines 2-5). Thus, each word will be transformed into an attribute where its records are filled by 1 or 0 according to the participant if it inserted or not such word (lines 6-10).

Algorithm 2 Binary Conversion Algorithm.

Require: Survey: $\mathcal{S}$; Set of attributes: $\mathcal{A}$; Set of participants: $\mathcal{P}$, List of word sets: $\mathcal{W}$.
Ensure: Binary matrix: $\mathcal{B}$.
  1: $\mathcal{B} \leftarrow \emptyset$
  2: **for** each attribute $A_i \in \mathcal{A}$ **do**
  3:     **for** each word $w_j \in W_i$ **do**
  4:       $A_{i_j} \leftarrow \emptyset$
  5:       **for** each participant $P_k \in \mathcal{P}$ **do**
  6:         **if** w_j is inserted by P_k **then**
  7:           $A_{i_j} \leftarrow A_{i_j} \cup \{1\}$
  8:         **else**
  9:           $A_{i_j} \leftarrow A_{i_j} \cup \{0\}$
10:         **end if**
11:       **end for**
12:       $\mathcal{B} \leftarrow \mathcal{B} \cup \{A_{i_j}\}$
13:     **end for**
14: **end for**
15: **return** $\mathcal{B}$

After applying the **Algorithm 2**, the data transformation stage will generate a binary matrix $\mathcal{B}$ of dimensions $N \times \alpha$, where the rows represent the partici-

pants' records and the columns indicate the new attributes with size α, as follows:

$$
\begin{array}{c}
\\
P_1 = \\
P_2 = \\
\vdots \\
P_N =
\end{array}
\begin{pmatrix}
A_1 & A_2 & \cdots & A_\alpha \\
r_{11} & r_{12} & \cdots & r_{1\alpha} \\
r_{21} & r_{22} & \cdots & r_{2\alpha} \\
\vdots & \vdots & \cdots & \vdots \\
r_{N1} & r_{N2} & \cdots & r_{N\alpha}
\end{pmatrix}
\tag{1}
$$

Finally, $\mathcal{B}$ will be sent to the next stage in the CRISE framework for further analysis and process.

3.3. Data Reduction Stage

At the end of data transformation stage, the binary matrix may contain a massive number of attributes due to the binary conversion algorithm. This will provide two main challenges: first, it complicates the data processing operation during the data analysis, especially when the initial number of attributes in the survey becomes bigger. Second, some attributes may be correlated due to the similarities between data inserted by the participants. Thus, eliminating attribute redundancy is becoming essential in order to not affecting the clustering accuracy in the next stage. Therefore, to overcome these challenges, the data reduction stage aims to search the correlations among the attributes, and then try to remove the correlated ones. Thus, this stage is consisted on two steps as follows:

- *Attribute Correlation's Computation*: in literature, one can find many techniques for calculating the correlations between data sets such as association rules, principal component analysis, backward feature elimination, low variance filter, etc. [25]. In this paper, we are interested in the correlation matrix (CM) method that is considered as one of the most well-known and efficient approaches in calculating feature similarities. Basically, CM is a square matrix that searching the similarity between pairs of features according to the Hamming distance. Then, features with high similarity are identified by thresholding the metric. Therefore, two attributes A_i and A_j are considered similar, based on the Hamming distance, if the difference between them is less than a defined threshold as follows:

$$
H_{i,j} = \frac{\sum_{k=1}^{N} 1_{r_i = r_j}}{N} \leq \mathcal{H}
\tag{2}
$$

where $r_i \in R_i$, $r_j \in R_j$ and $\mathcal{H}$ is the Hamming distance threshold, in range $[0,1]$, defined by the user depending on the application requirements. Indeed, a value of $\mathcal{H}$ close to zero will increase the number of attributes and lead to increase the accuracy of the obtained results, and vice versa. In our simulation, we changed the value of $\mathcal{H}$ to several numbers in order to study its effect on the analysis.

Thus, after calculating the similarity between all pairs of attributes, the CM will be represented as follows:

$$A_1 = \begin{matrix} & A_1 & A_2 & \cdots & A_\alpha \\ A_1 \\ A_2 \\ \vdots \\ A_\alpha \end{matrix} \begin{pmatrix} 0 & H_{1,2} & \cdots & H_{1,\alpha} \\ H_{2,1} & 0 & \cdots & H_{2,\alpha} \\ \vdots & \vdots & \cdots & \vdots \\ H_{\alpha,1} & H_{\alpha,2} & \cdots & 0 \end{pmatrix} \tag{3}$$

- *Attribute Correlation's Elimination*: in this step, let first define the list of similar pairs' attributes among those calculated in the CM:

$\mathcal{O} = \left\{ \left(A_i, A_j \right) \text{ such that } H_{i,j} \leq \mathcal{H} \right\}$. Our objective is to select a set of attributes among $\mathcal{A}$ to perform the data clustering stage in a way to remove the attributes similarity without affecting the accuracy of the obtained clusters. **Algorithm 3** shows the process of eliminating the attribute similarity by taking the list of similar attributes, e.g. $\mathcal{O}$, and returns the selected list of final attributes, e.g. $\mathcal{F}$. Briefly, for each pairs of similar attributes, the algorithm selects one having more similar attributes and it added to the final list of selected attributes (lines 2-4). Then, it deletes all the pairs from $\mathcal{O}$ containing one of similar attributes (line 5).

Algorithm 3   Attribute Selection Algorithm.

Require: Set of attributes: $\mathcal{A}$; List of similar attributes: $\mathcal{O}$.
Ensure: List of selected attributes: $\mathcal{F}$.
1: $\mathcal{F} \leftarrow \emptyset$
2: **for** each pair of attributes $(A_i, A_j) \in \mathcal{A}$ **do**
3:    consider A_i has more number of similar attributes than A_j
4:    $\mathcal{F} \leftarrow \mathcal{F} \cup \{A_i\}$
5:    delete all pairs of attributes containing one of A_i or A_j
6: **end for**
7: **return** $\mathcal{F}$

3.4. Data Clustering Stage

In machine learning, clustering is one of the most unsupervised algorithms that is used to classify unlabeled data into a set of clusters. Consequently, data within the same cluster will be more similar compared to those in other clusters. In the literature, clustering approach is broadly adapted to many sectors including, but not limited to, security, business, transportation, and healthcare applications [26] [27] [28]. Researchers have proposed several data clustering algorithms such as mean-shift, density-based spatial clustering of applications with noise, expectation-maximization using Gaussian mixture models, agglomerative hierarchical, etc. However, K-means still the most well-known clustering algorithm that is used in a lot of introductory data science and machine learning classes. In this paper, we are interested in the K-means algorithm in order to group the participants affected similarly by the psychological COVID-19 impact into the same clusters. Thus, the data clustering stage allows psychiatrists to study the severity of COVID-19 inside each cluster to determine the appropriate treatment.

However, we face two main challenges when using the traditional K-means algorithm: the selection of the cluster number (K) and the convergence criterion function. From one hand, selecting the number of clusters is a crucial decision because it determines the accuracy of the obtained clusters. On the other hand, the number of iterations generated by K-means is highly dependent on the selection of the convergence criterion function; thus, an inappropriate criterion function can lead to increase the computation process of K-means then increase the processing complexity. Therefore, in order to overcome these challenges, we propose a new version of K-means, called Jaccard-based K-means or JKmeans, that adapts the Jaccard coefficient to the traditional K-means algorithm. In the next sections, we first recall the K-means algorithm along with the Jaccard coefficient then we present our new clustering algorithm called JKmeans.

3.5. Recall of Kmeans Clustering Algorithm

As mentioned before, K-means [29] is one of the most flexible and simple algorithm in data clustering. Typically, the K-means is an iterative algorithm in which the process starts by randomly selecting an initial centroid for each cluster. Then, each data point is assigned to the nearest centroid using the Euclidean distance and the first round of cluster formation is performed. After that, the cluster centroids are updated and the process is repeated until the convergence of the criterion function (**Algorithm 4**). Subsequently, one of the most criterion functions that have been used in Kmeans is the mean square errors.

Algorithm 4   Kmeans Algorithm.

Require: Set of participants: $\mathcal{P}$; Cluster number: K.
Ensure: Set of obtained clusters: $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$.
 1: **for** $j \leftarrow 1$ to K **do**
 2:    randomly choose centroid c_j among $\mathcal{P}$ belongs to C_j
 3: **end for**
 4: **repeat**
 5:    **for** each data set $P_i \in \mathcal{P}$ **do**
 6:       Assign P_i to the cluster C_j with nearest c_j
          (i.e., $|P_i, c_{j*}| \leq |P_i, c_j|$; $j \in \{1, \ldots K\}$)
 7:    **end for**
 8:    **for** each cluster C_j, where $j \in \{1, \ldots K\}$ **do**
 9:       Update the centroid c_i to be the centroid of all
          records currently in C_j
10:    **end for**
11: **until** all K clusters meet the criterion function convergence
12: return $\mathcal{C}$

3.6. Jaccard Coefficient

Typically, the Jaccard coefficient is used for gauging the similarity and diversity between sample sets. Thus, it has been used in a wide range of applications including community detection in social networks [30], document and web pages plagiarism [31], attack detection [32], market analysis [33], etc. In this paper, we

use the Jaccard coefficient to measure the similarity between data record inserted by two participants. Thus, two data records R_i and R_j inserted by participants P_i and P_j respectively are considered similar according to the Jaccard coefficient if and only if:

$$\mathcal{J}\left(R_i, R_j\right) = \frac{\sum_{k=1}^{\alpha} 1_{r_i = r_j}}{2 \times \alpha - \sum_{k=1}^{\alpha} 1_{r_i = r_j}} \geq t_{\mathcal{J}} \tag{4}$$

where $t_{\mathcal{J}} \in [0,1]$ is the Jaccard threshold defined by the experts; 1 indicates that the participants inserted the same data for all attributes while 0 indicates none of the attributes are similar.

3.7. Jaccard-Based Kmeans Algorithm: JKmean

In this section, we integrate the Jaccard coefficient to the traditional K-means in order to produce a more accurate clustering algorithm, e.g. JKmeans. Indeed, JKmeans is two-fold: first, it dynamically finds the optimal number of clusters without using trial-and-error method adapted in most existing algorithms. Second, it uses a new convergence criterion to increase the accuracy of the obtained clusters. Basically, JKmeans assumes that all participants inserted different records of all attributes, thus each participant is considered a cluster. Then, it recursively merges the clusters every time a similarity between them is detected. The process of cluster joining is stopped when no similarity between clusters is existing, which used as a criterion function in our algorithm. **Algorithm 5** describes the process of JKmeans that takes, as input, the record sets of all participants, and finds the clusters of participants have similar psychological symptoms, e.g. $\mathcal{C}$. First, we randomly select a record set and assign as the centroid of the first cluster. For the other record sets, we calculate the similarity of each one

---

Algorithm 5   JKmeans Algorithm.

Require: Set of participants: $\mathcal{P}$; Jaccard threshold: $t_{\mathcal{J}}$.
Ensure: Set of obtained clusters: $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$.
1: select a random record set R_i to be the centroid c_1 of C_1
2: $C_1 \leftarrow \{R_i\}$
3: $\mathcal{C} \leftarrow \mathcal{C} \cup \{C_1\}$
4: **for** each record set R_j of participant P_j **do**
5:      **for** each cluster $C_k \in \mathcal{C}$ **do**
6:          **if** $\mathcal{J}(R_j, c_k) \geq t_{\mathcal{J}}$ **then**
7:             $C_k \leftarrow C_k \cup \{R_j\}$
8:             break
9:          **end if**
10:      **end for**
11:      **if** R_j is not similar to any cluster centroid **then**
12:          $C_k \leftarrow \{R_j\}$
13:          $\mathcal{C} \leftarrow \mathcal{C} \cup \{C_k\}$
14:      **end if**
15: **end for**
16: **return** $\mathcal{C}$

---

with each of the cluster centroids, then we assign it to the cluster having high similarity, e.g. greater than the Jaccard threshold, with its centroid (lines 4-10). Otherwise, if the record set does not have any similarity with the cluster centroids, it assigned as a centroid for a new cluster (lines 11-14).

4. Performance Evaluation

In this section, we show the relevance of our mechanism, CRISE, in terms of understanding the psychological impact of the COVID-19 outbreak. We implemented CRISE based on the Python language under Anaconda framework and we used the answers of survey collected by more than 2000 as described in section 0. We have considered the system environment of Windows 10 operating system that possesses 64-bit Quad core processor packed with 2GB NVDIA CUDA graphic card packed with 16 GB RAM. Table 1 summarizes the parameter description used in our simulation with their tested values.

4.1. Study of the Attribute Number Variation

In this section, we study the performance of the transformation and reduction stages in CRISE framework in terms of optimizing the number of attributes. Figure 2 shows the original number of attributes in the survey, e.g. raw attributes, and after applying the data transformation and reduction stages. The obtained results are highly dependent on the Hamming distance threshold ($\mathcal{H}$) that is varied between 0.1 and 0.3. We show that the binary conversion algorithm (see **Algorithm 2**) proposed in the data transformation stage increases the number of raw attributes from 22 in the original survey to 123 after extracting all

Table 1. Simulation configuration.

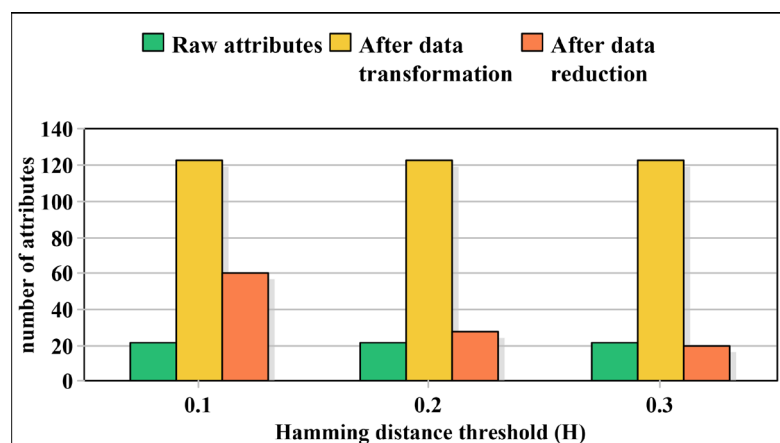| Parameter	Description	Values
$\mathcal{H}$	Hamming distance threshold	0.1, 0.2, 0.3
$t_{\mathcal{J}}$	Jaccard coefficient threshold	0.5, 0.55, 0.6

Figure 2. Variation of the attribute numbers after applying transformation and reduction stages.

word list inserted by the participants. This confirms the behavior of our framework by allowing the participants to express their psychological impact in their own words. We can also observe that the number of attributes is then decreased after applying the data reduction stage. Subsequently, the attribute numbers are reduced by 51%, 77% and 83% when varying the Hamming threshold to 0.1, 0.2 and 0.3 respectively. This is due to the similarity between the attributes that increases with the increase of $\mathcal{H}$, thus, the reduction stage will eliminate more redundant attributes.

4.2. Study of Participant Distribution over Clusters

Indeed, one of the most advantages of JKmeans algorithm is its ability to dynamically find the optimal number of clusters according to the Jaccard coefficient criterion. **Figure 3** shows the number of obtained clusters along with their sizes, e.g. the number of participants inside each one, when varying the Hamming and the Jaccard threshold values. The obtained results show that the cluster formation is highly affected by the variation of $\mathcal{H}$ and $t_{\mathcal{J}}$ allowing more understanding of the psychological impact of COVID-19 outbreak. Mainly, we observe that the number of clusters arrives to 17 and 40 when varying the Hamming and Jaccard thresholds respectively, while their corresponding cluster sizes arrive to 1226 and 424 respectively. Furthermore, the following observations are eminent:

*   The number of clusters K increases with the increase of the value of $\mathcal{H}$ (**Figure 3(a)**). For instance, K increases from 6 to 17 when $\mathcal{H}$ varied from 0.1 to 0.3. This is because the number of attributes will reduce when the value of Hamming threshold increases, which make the Jaccard similarity condition more difficult to meet thus increase the number of clusters.

*   The number of clusters K increases with the increase of the value of $t_{\mathcal{J}}$ (**Figure 3(b)**). For instance, K increases from 9 to 40 when $t_{\mathcal{J}}$ varied from 0.5 to 0.6. This is because the similarities between the participants decrease when the value of Jaccard threshold increases, which increase the number of obtained clusters.
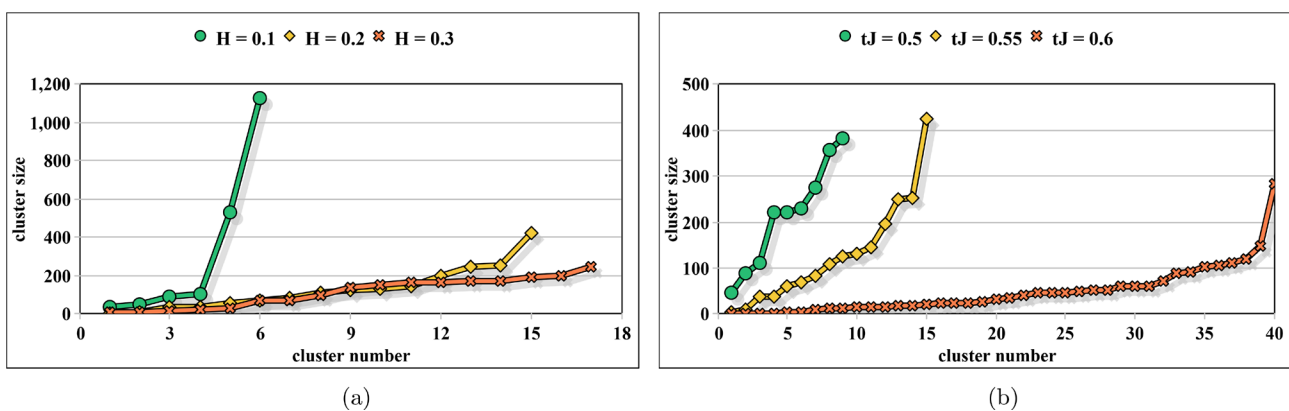

Figure 3. Participants' distribution over clusters after applying data clustering stage. (a) in function of Hamming threshold; (b) in function of Jaccard threshold.

- The average number of participants per cluster decreases with the increase of $\mathcal{H}$ or $t_{\mathcal{J}}$. Subsequently, each cluster contains an average of 323, 129 and 114 participants when $\mathcal{H}$ changed to 0.1, 0.2 and 0.3 respectively. While, the average number of participants per cluster changed to 215, 129 and 48 when $t_{\mathcal{J}}$ changed to 0.5, 0.55 and 0.6 respectively. This will lead to increase the similarities between the records of participants in each cluster and enhance the understanding of the pandemic outbreak of such participants.

4.3. JKmeans Clustering Accuracy

An important metric to asses any clustering algorithm is by calculating the accuracy. Indeed, a good clustering accuracy is obtained when the difference within clusters is minimized and that between clusters is maximized. In this paper, we focus on the mean square error (MSE) as a well-known and most used metric to calculate the clustering accuracy. Figure 4 shows the average MSE of all clusters by applying traditional K-means, JKmeans and the Enhanced K-means (EK-means) proposed in [34]. The obtained results show that JKmeans outperforms K-means and EK-means in terms of reducing the clustering error in all cases. Subsequently, JKmeans optimizes the clustering accuracy up to 53% and 46% comparing to K-means and EK-means respectively. We can also observe that JKmeans ensures a high level of similarity within the clusters, which strongly confirms the behavior of our framework by grouping participants having similar psychological impact in the same cluster. Furthermore, we can notice the following:

- The clustering accuracy of JKmeans increases with the increase of $\mathcal{H}$ (Figure 4(a)). For instance, the JKmeans error reduced from 0.21 to 0.18 when $\mathcal{H}$ varied from 0.1 to 0.3. This is due to the number of clusters that increases when $\mathcal{H}$ increases, which reduces the average number of participants per cluster and reduce the error inside each cluster.
- The clustering accuracy of JKmeans increases with the increase of $t_{\mathcal{J}}$ (Figure 4(b)). For instance, the error of JKmeans decreased from 0.23 to 0.15 when $t_{\mathcal{J}}$ varied from 0.5 to 0.6 (for the same previous reason).
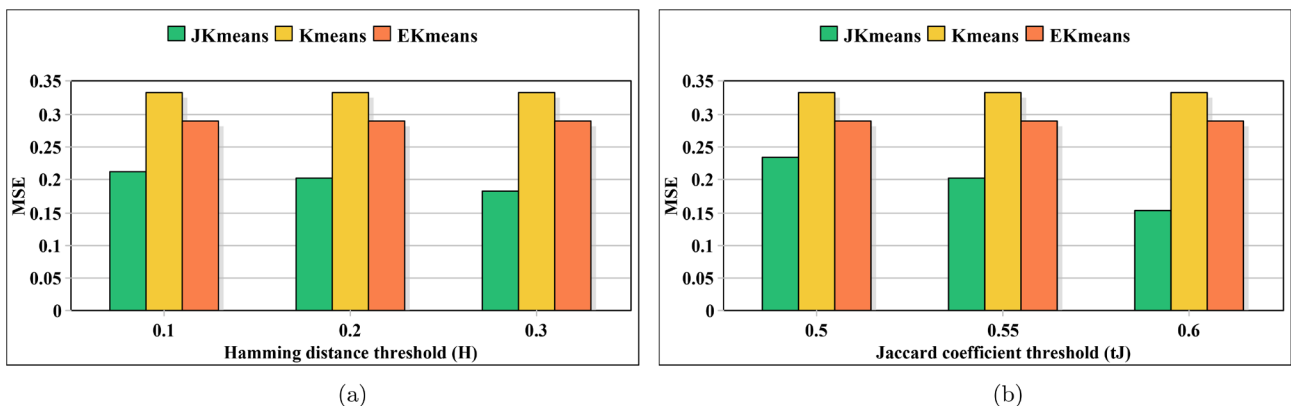

(a) (b)

Figure 4. Comparison of clustering accuracy. (a) in function of Hamming threshold; (b) in function of Jaccard threshold.

- Psychological-based Clustering Analysis In this section, we analyze the psychological impact of COVID-19 during the lockdown for each obtained cluster, after applying JKmeans. Indeed, in the light of the obtained results, we distinguish between two types of psychological impact of COVID-19: cluster-based and global-based impacts. In the first type, we show some specific characteristics of each cluster separately while, in the second type, we analyze the effect of the pandemic on the whole population, e.g. all clusters.

From the cluster-based point of view, we show common characteristics of psychological impact of COVID-19 that describe each cluster as follows:

- In the first set of clusters, we remark that the activities, prayers, and family life characterize the individuals who declare a comfort in their past "past comfort". Indeed, if these characteristics are repeated in the period just before COVID-19, we report less anxious reactions than those who do not practice them recently or in their past.

- In the second set of clusters, we show that the absence of the "past comfort" (non-comfort indicator that reaches the degree of four over four is reported in the present and the distant past) is the main characteristic of the individuals of these clusters. Individuals that lack decision making and confidence in social and political situations also mark these clusters. On the other hand, the presence of "past comfort" for individuals in these clusters decreases the degree of presence of anxiety symptoms in the distant and recent past. This last degree increases with the apparition of COVID-19, which can be seen as a normal and transient reaction during the pandemic.

- In the third set of clusters, we show that work, walking, playing, reading, friends, and relaxation are parallel to lower degrees of anxiety manifestations. These individuals are women who practice these comfort activities, with a specialization in Psychology. It shows that these people emphasize the psychological effect of these activities in the management of anxiety.

- In the fourth set of clusters, individuals are characterized by a high degree of "comfort today" and "comfort past" and the activity of "listening to music". The anxiety was zero or medium in the recent past and the distant past. Anxiety reactions during COVID-19 are found to be moderate and transient.

- In the fifth set of clusters, the faith as well as the protection of the family, and social contact, constitutes positive resources against the appearance of anxiety during COVID-19. It seems that the activities in the past and present reduce the anxiety reactions during the period of COVID-19. The absence or the minimal intensity of anxiety during the COVID-19 pandemic is the main characteristic of these clusters.

From the global-based point of view, we show some common characteristics describing the psychological impact of COVID-19 on the whole individuals in the sample taken from Lebanon. The following observations are eminent:

- Individuals show more anxious reactions due to a quasi-absence of comfort during and before COVID-19, and even in their long past.

- What characterizes the individuals in terms of anxiety symptoms is that most manifest anxious reactions with the appearance of COVID-19. These symptoms reach level 4 at the physical and mental level. It should be noted that the presence of this anxiety was triggered before the pandemic. Therefore, we have anxious individuals; that chronic anxiety appears on them from the distant past.

- The psychological reactions of anxiety individuals depend on their behaviors and personal history, and the accumulation of negative situations or situations that generate symptoms of anxiety. Those individuals that fail to remove memories that are a source of anxiety, will keep the same symptoms experienced before.

According to the above analysis, we give directions for other open issues that are very important to be addressed by the researchers in the future works in order to deeply understand the psychological impact of COVID-19. Some of the most relevant questions include: Is it a matter of a memory that memorizes cognitive patterns and negative events that cause anxiety? Is it a collective or individual memory? Is it a matter of learning individual or collective behavior? Is it in relation with a triggering factor? Who are the invulnerable ones? What are their characteristics? Who need a psychological intervention? How we can identify severs cases?

Finally, in this research, we grouped individuals with similar characteristics and behaviors together. This non supervised classification helped us to identify the characteristics that highly impact anxiety and identify individual and groups who probably need psychological treatment. A very specific and "revolutionary" psychotherapies such as EMDR, consists of reprocessing negative events, cognitions and emotions stored in memory, sources of anxiety, can be recommend for certain individuals. The EMDR therapy is a behavioral therapy that takes care of the physical reactions through specific techniques. This therapy helps individuals re-adjusts their anxious reactions to reach the normal level.

5. Conclusions

The COVID-19 pandemic is totally changed the lifestyle and the working modes of the people by imposing quarantine and lockdown. Unfortunately, such situations are negatively affected the mental health of most people and provided many psychological problems such as stress, depression, and so on. This motivated the public authorities of most countries to respond to psychological impact of COVID-19 pandemic and help their citizens to coping with current situation in a healthy way. In this paper, we proposed a psychological-based framework for understanding the mental health of the population during COVID-19 pandemic. The framework is called CRISE and consists of four stages: collection, transformation, reduction and clustering. Mainly, CRISE used real data collected from more than 2000 persons followed by a preprocess for cleaning and filtering purposes. Then, CRISE aimed to grouping psychological persons having similar

situations into the same clusters for a later studying and treatment.

We have two main directions to enhance the performance of CRISE. First, we plan to test another convergence criterion in clustering stage, such as distance functions, in order to optimize the clustering error. Second, we seek to add an evaluation technique that allows to assess the psychological level of each cluster for a later appropriate treatment.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] World Health Organization (2007) Everybody's Business—Strengthening Health Systems to Improve Health Outcomes: WHO's Framework for Action, World Health Organization, Geneva.

[2] Dimensions Database (n.d.). https://app.dimensions.ai/discover/publication

[3] Nature Index (2020) Coronavirus Research Publishing: The Rise and Rise of COVID-19 Clinical Trials. https://www.natureindex.com/news-blog/how-coronavirus-is-changing-research-practices-and-publishing

[4] Li, S., Wang, Y., Xue, J., Zhao, N. and Zhu, T. (2020) The Impact of COVID-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. *International Journal of Environmental Research and Public Health*, **17**, Article No. 2032. https://doi.org/10.3390/ijerph17062032

[5] Spoorthy, M.S., Pratapa, S.K. and Mahant, S. (2020) Mental Health Problems Faced by Healthcare Workers Due to the COVID-19 Pandemic—A Review. *Asian Journal of Psychiatry*, **51**, Article ID: 102119. https://doi.org/10.1016/j.ajp.2020.102119

[6] Nicola, M., Alsafi, Z., Sohrabi, C., Kerwan, A., Al-Jabir, A., Iosifidis, C., Agha, M. and Agha, R. (2020) The Socio-Economic Implications of the Coronavirus Pandemic (COVID-19): A Review. *International Journal of Surgery*, **78**, 185-193. https://doi.org/10.1016/j.ijsu.2020.04.018

[7] Bartik, A.W., Bertrand, M., Cullen, Z., Glaeser, E.L., Luca, M. and Stanton, C. (2020) The Impact of COVID-19 on Small Business Outcomes and Expectations. *Proceedings of the National Academy of Sciences of the United States of America*, **117**, 17565-17666. https://doi.org/10.1073/pnas.2006991117

[8] Gössling, S., Scott, D. and Michael, H.C. (2020) Pandemics, Tourism and Global Change: A Rapid Assessment of COVID-19. *Journal of Sustainable Tourism*, **29**, 1-20. https://doi.org/10.1080/09669582.2020.1758708

[9] Crawford, J., Butler-Henderson, K., Rudolph, J., Malkawi, B., Glowatz, M., Burton, R., Magni, P. and Lam, S. (2020) COVID-19: 20 Countries' Higher Education Intra-Period Digital Pedagogy Responses. *Journal of Applied Learning & Teaching*, **3**, 9-28. https://doi.org/10.37074/jalt.2020.3.1.7

[10] Rodrguez-Rey, R., Garrido-Hernansaiz, H. and Collado, S. (2020) Psychological Impact and Associated Factors During the Initial Stage of the Coronavirus (COVID-19) Pandemic among The general Population in Spain. *Frontiers in psychology*, **11**, Article No. 1504. https://doi.org/10.3389/fpsyg.2020.01540

[11] Xiang, Y.-T., Yang, Y., Li, W., Zhang, L., Zhang, Q., Cheung, T. and Ng, C.H.

(2020) Timely Mental Health Care for the 2019 Novel Coronavirus Outbreak Is Urgently Needed. *The Lancet Psychiatry*, **7**, 228-229.
https://doi.org/10.1016/S2215-0366(20)30046-8

[12] Bao, Y., Sun, Y., Meng, S., Shi, J. and Lu, L. (2020) 2019-nCoV Epidemic: Address Mental Health Care to Empower Society. *The Lancet*, **395**, E37-E38.
https://doi.org/10.1016/S0140-6736(20)30309-3

[13] Brooks, S.K., Webster, R.K., Smith, L.E., Woodland, L., Wessely, S., Greenberg, N. and Rubin, G.J. (2020) The Psychological Impact of Quarantine and How to Reduce It: Rapid Review of the Evidence. *The Lancet*, **395**, 912-920.
https://doi.org/10.1016/s0140-6736(20)30460-8

[14] Duan, L. and Zhu, G. (2020) Psychological Interventions for People Affected by the COVID-19 Epidemic. *The Lancet Psychiatry*, **7**, 300-302.
https://doi.org/10.1016/S2215-0366(20)30073-0

[15] Schäfer, S., Sopp, R., Schanz, C., Staginnus, M., Göritz, A.S. and Michael, T. (2020) Impact of COVID-19 on Public Mental Health and the Buffering Effect of Sense of Coherence. https://doi.org/10.31219/osf.io/4ptfz

[16] Fernandez, I. and Formenti, L. (2020) EMDR Europe: Psychological Support for the Different Populations Exposed to Covid-19 Pandemic.
https://victimsupport.eu/news/emdr-europe-psychological-support-for-the-different-populations-exposed-to-covid-19-pandemic/

[17] Solomon, R.M. and Hensley, B.J. (2020) EMDR Therapy Treatment of Grief and Mourning in Times of COVID-19. *Journal of EMDR Practice and Research*, **14**, 162-174. https://doi.org/10.1891/EMDR-D-20-00031

[18] Lantheaume, S. (2018) Utilisation de la thérapie EMDR dans le traitement d'un ESPT après cancer du sein. *Journal de Thérapie Comportementale et Cognitive*, **28**, 3-16. https://doi.org/10.1016/j.jtcc.2017.07.001

[19] Shubina, I. (2015) Cognitive-Behavioral Therapy of Patients with PTSD: Literature Review. *Procedia-Social and Behavioral Sciences*, **165**, 208-216.
https://doi.org/10.1016/j.sbspro.2014.12.624

[20] Marks, I., Lovell, K., Noshirvani, H., Livanou, M. and Thrasher, S. (1998) Treatment of Posttraumatic Stress Disorder by Exposure and/or Cognitive Restructuring: A Controlled Study. *Archives of General Psychiatry*, **55**, 317-325.
https://doi.org/10.1001/archpsyc.55.4.317

[21] King, M.W. and Resick, P.A. (2014) Data Mining in Psychological Treatment Research: A Primer on Classification and Regression Trees. *Journal of Consulting and Clinical Psychology*, **82**, 895-905. https://doi.apa.org/doi/10.1037/a0035886

[22] Wang, C., Pan, R., Wan, X., Tan, Y., Xu, L., Ho, C.S. and Ho, R.C. (2020) Immediate Psychological Responses and Associated Factors during the Initial Stage of the 2019 Coronavirus Disease (COVID-19) Epidemic among the General Population in China. *International Journal of Environmental Research and Public Health*, **17**, Article No. 1729. https://doi.org/10.3390/ijerph17051729

[23] Le, X.T.T., Dang, K.A., Toweh, J., Nguyen, Q.N., Le, H.T., Toan, D.T.T., Phan, H.B.T., Nguyen, T.T., Pham, Q.T., Ta, N.K.T. and others (2020) Evaluating the Psychological Impacts Related to COVID-19 of Vietnamese People under the First Nationwide Partial Lockdown in Vietnam. *Frontiers in Psychiatry*, **11**, Article No. 824. https://doi.org/10.3389/fpsyt.2020.00824

[24] Sundarasen, S., Chinna, K., Kamaludin, K., Nurunnabi, M., Baloch, G.M., Khoshaim, H.B., Hossain, S.F.A. and Sukayt, A. (2020) Psychological Impact of COVID-19 and Lockdown among University Students in Malaysia: Implications and Policy

Recommendations. *International Journal of Environmental Research and Public Health*, **17**, Article No. 6206. https://doi.org/10.3390/ijerph17176206

[25] Huang, X., Wu, L. and Ye, Y. (2019) A Review on Dimensionality Reduction Techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, **33**, Article ID: 1950017. https://doi.org/10.1142/S0218001419500174

[26] Landauer, M., Skopik, F., Wurzenberger, M. and Rauber, A. (2020) System Log Clustering Approaches for Cyber Security Applications: A Survey. *Computers & Security*, **92**, Article ID: 101739. https://doi.org/10.1016/j.cose.2020.101739

[27] Senouci, O., Harous, S. and Aliouat, Z. (2020) Survey on Vehicular ad HOC Networks Clustering Algorithms: Overview, Taxonomy, Challenges, and Open Research Issues. *International Journal of Communication Systems*, **42**, Article No. e4402. https://doi.org/10.1002/dac.4402

[28] Ghosal, A., Nandy, A., Das, A.K., Goswami, S. and Panday, M. (2020) A Short Review on Different Clustering Techniques and their applications, Emerging Technology in Modelling and Graphics. In: Mandal, J. and Bhattacharya, D., *Emerging Technology in Modelling and Graphics*, Vol. 937, Springer, Singapore, 69-83. https://doi.org/10.1007/978-981-13-7403-6_9

[29] MacQueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the* 5*th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 281-297.

[30] Guo, K., He, L., Chen, Y., Guo, W. and Zheng, J. (2020) A Local Community Detection Algorithm Based on Internal Force between Nodes. *Applied Intelligence*, **50**, 328-340. https://doi.org/10.1007/s10489-019-01541-1

[31] Diana, N.E. and Ulfa, I.H. (2019) Measuring Performance of N-Gram and Jaccard-Similarity Metrics in Document Plagiarism Application. *Journal of Physics: Conference Series*, **1196**, Article ID: 012069. https://doi.org/10.1088/1742-6596/1196/1/012069

[32] Li, B., Gao, M., Ma, L., Liang, Y. and Chen, G. (2019) Web Application-Layer DDoS Attack Detection Based on Generalized Jaccard Similarity and Information Entropy. *International Conference on Artificial Intelligence and Security*, New York, 26-28 July 2019, 576-585. https://doi.org/10.1007/978-3-030-24271-8_51

[33] Moodley, R., Chiclana, F., Caraffini, F. and Carter, J. (2019) Application of Uninorms to Market Basket Analysis. *International Journal of Intelligent Systems*, **34**, 39-49. https://doi.org/10.1002/int.22039

[34] Siddiqui, M.K., Morales-Menendez, R., Gupta, P.K., Iqbal, H.M., Hussain, F., Khatoon, K. and Ahmad, S. (2020) Correlation between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases Based on Machine Learning Analysis. *Journal of Pure and Applied Microbiology*, **14**, 1017-1024. https://doi.org/10.22207/JPAM.14.SPL1.40