Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 10, No. 2, 2022, 159-169.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Clustering Methods to Analyze Social Media Posts during Coronavirus Pandemic in Iran

Fatemeh Amiri[1], Samira Abbasi[2*] and Mahboobe Babaie mohamadeh [3]

1. Department of Computer Engineering, Hamedan University of Technology, Hamedan, Iran.
2. Department of Biomedical Engineering, Hamedan University of Technology, Hamedan, Iran.
3. Society of Rural Social Development, University of Tehran, Tehran, Iran.

Article Info

Article History:
Received 17 October 2021
Revised 05 January 2021
Accepted 21 February 2022

DOI:10.22044/JADM.2022.11270.2285

Keywords:
Clustering, COVID-19, Iran, Social Media, Social Trust.

**Corresponding author: samira.abbasi@gmail.com (S. Abbasi).*

Abstract

During the COVID-19 crisis, we face a wide range of thoughts, feelings, and behaviors on the social media that play a significant role in spreading information regarding COVID-19. Trustful information, together with hopeful messages, could be used to control the people's emotions and reactions during pandemics. This work examines the Iranian society's resilience in the face of the Corona crisis, and provides a strategy to promote resilience in similar situations. It investigates the posts and news related to the COVID-19 pandemic in Iran in order to determine which messages and references have caused concern in the community, and how they could be modified? and also which references are the most trusted publishers? The social network analysis methods such as clustering are used in order to analyze the data. In the present work, we apply a two-stage clustering method constructed on the self-organizing map and K-means. Due to the importance of social trust in accepting messages, this work examines the public trust in social posts. The results obtained show that trust in the health-related posts is less than the social-related and cultural-related posts. The trusted posts are shared on Instagram and the news sites. The health and cultural posts with negative polarity affect the people's trust, and lead to negative emotions such as fear, disgust, sadness, and anger. Thus we suggest that non-political discourses are used in order to share the topics in the field of health.

1. Introduction

The COVID-19 pandemic has an essential influence on the utilization of social media by the people worldwide. Before the prevalence of COVID-19, people already relied on the social media in order to collect information and news but with the worldwide spread of the coronavirus, the individual activity on the social networks like Twitter and Facebook has risen [1]. Thus during the COVID-19 prevalence, the positive role of the social media in the public dissemination and discussion of important information about the pandemic has been revealed [2].

During the COVID-19 outbreak, we face a wide range of thoughts, feelings, attitudes, and behaviors on the social media, and this information is of particular importance. Public data shared on the social networks by the users around the world can be used to quickly recognize the main feelings, thoughts, attitudes, and issues that are occupying the people's minds concerning the COVID-19 outbreak. Several studies have reported that the social media play a vital role in understanding the society's attitudes and behaviors during crises as a way to protect the critical communication and health promotion messages [1, 3, 4]. Such data may help the policy-makers and healthcare professionals to review the primary issues of concern, and analyze them more appropriately [1]. It could also help understand the social status, effectiveness of the policies and

methods implemented in the societies in the face of this crisis, and help design the subsequent steps to overcome it. The trustful information and messages of hope and solidarity shared on social media could be used to control the people's emotions and reactions during pandemics, build safety nets, and promote resilience. This work examines the Iranian society's resilience in the face of the Corona crisis by investigating the data extracted from the social media and providing a strategy to promote resilience in similar situations. The extracted dataset includes Persian posts and news published by the Iranians on the social networks focusing on Covid-19. This content has been extracted from the Telegram channels, public Instagram, and Twitter pages, as well as domestic news. These social networks have been selected since they are popular networks for Iranians to share their opinions and thoughts.

The users' trust in the posts is an essential aspect of decision-making and accepting or not accepting the information posted on the social media [5]. Thus we aim to analyze trust in the posts shared by the Twitter, Instagram, and Telegram users and news related to the COVID-19 epidemic in Iran. The motivation of this work is to analyze these posts and investigate the impact on social media users' trust. The results can help the government to understand the current public attitude towards policy and the overall direction of public opinion. It also can provide insights for the in-charge organizations for an efficient control of the situation. In this work, we apply the clustering algorithms to analyze the data extracted from the social media. A two-step clustering method was used, which was formed from two main steps. The first step was a self-organizing map (SOM), and the second step was K-means clustering. These are common clustering approaches but the results are important. We analyze our clustering algorithm's performance on the dataset with 6339 posts and news published by the Iranians in the social media with a focus on COVID-19. This content has been published online from 21 January 2020 to 29 April 2020.

In order to achieve the aim of this work, it is necessary to address the subsequent questions.

-What messages and references have caused concern within the community, and how they may be modified?

-What were the foremost trusted publishers and messages shared during the Corona crisis?

This paper is arranged as what follows. The second section summarizes the related works on the analysis of social media, focusing specifically on COVID-19. The third section describes the materials and methods. The results of the analysis are given in the fourth section. Finally, the last section is a discussion, and highlights the implications of these results, and draws a conclusion.

2. Related Works

Social media are the lens through which people collect and share information in different situations [4, 6-9]. With the rapid growth of the social media content, studies such as sentiment analysis or online opinion mining of text have attracted attention from different areas [10]. Approaches to conducting sentiment analysis include the supervised machine learning approach, symbolic techniques, and unsupervised learning techniques. Among these methods, the supervised learning approaches give the highest accuracy but the cost is also high in terms of the human participation and time. On the other hand, the symbolic techniques run very fast but the accuracy rates are usually poor. The clustering sentiment analysis technique performs a balance on the aspects of both cost and accuracy, and its performance is higher compared to the average performance of supervised learning and symbolic approaches. In addition, when there is not a large dataset, the unsupervised techniques have a better performance [11].

Different machine learning approaches have been applied to assess the content analysis of the social media [12-16]. Some studies have found that the naïve Bayes classifier is a proven, effective, and simple method for text classification [16]. Boiy *et al*. [14] have shown that the machine learning approaches (e.g. support vector machine, Naïve Bayes classifier, and maximum Entropy) have a high accuracy (above 80%) in classifying sentiment while the symbolic techniques (i.e. based on the direction of words and force) have an accuracy less than 80%. Also logistics regression is one of the common and earlier methods for classification. Multi-nomial logistic regression could predict the sentiment of the Twitter users with an accuracy of 74% [15]. K-Nearest Neighbor (KNN) is a common non-parametric text classifier that classifies the documents or texts based on the similarity measurement [16]. Nemes and Kiss [17] have analyzed the sentiments of the Twitter users with Natural Language Processing, and classified them using the Recurrent Neural Network.

Also different studies have used unsupervised methods. For example, Li and Liu [18] have used the clustering method for sentiment analysis, and investigated this application. Wu *et al*. [19] have

used the sentiment analysis method based on K-means and online transfer learning to analyze the posts related to the online products. Kaveh-Yazdy *et al*. [20] have used an unsupervised method to analyze the news in the Persian language extracted from Telegram. The social media content can be used for different purposes. For example, social media in the United States is popular, and the reports show that 68% of the adults in America receive the news on the social media [21]. This is especially true for the health and science information. Social media also have been used to examine the public awareness, attitudes, and reactions to specific diseases [3, 22]. For example, the Twitter data has been widely used for crises analysis and tracking including pandemic analysis [23-26]. Many studies have been done on the social media analysis. Studies like [27] and [28] have shown that the community participation has been recognized as a method to involve people in the decision-making process and empower them towards an active participation to promote their health. Social networks have been very important in the Corona crisis also, and many studies have been conducted on the social networks data in different countries, with different goals and using different methods. For example, Li *et al*. [29] have shown that situational information is valuable to the people and authorities in responding to the epidemic. They applied natural language processing techniques to classify the COVID-19-related information into seven types of situational information. Lopez *et al*. [30] have analyzed the Twitter's multi-lingual dataset to understand how people in different countries react to the COVID-19 policies. Tao *et al*. [31] have analyzed the tweets in the Chinese language on Weibo related to oral health during the Covid-19 pandemic. Massaad and Cherfan [32] have investigated the Twitter public data related to telehealth during the Covid-19 pandemic using the clustering methods. Kaveh-Yazdy *et al*. [20] have investigated the content extracted from Telegram in the Persian language.

On the other hand, public trust is essential in the social attitude and behavior, especially in crises such as the COVID-19 pandemic [33-35]. Forsyth *et al*. [36] have defined trust as "a belief or attitude about a partner's grace and reliability in high-risk situations". Also Balog-Way and McComas [37] have denoted that the transparency of communication is critical in building the public trust. David Pastor-Escuredo and Carlota Tarazona [38] have suggested a framework to identify the leaders in Twitter based on the analysis of the social graph obtained from the activity in Twitter. The leaders could help to propagate trustful information with a positive influence. In times of crisis, in addition to trust, the quality of the information in social media is also essential. Dissemination of misinformation can weaken the public health strategies [39], and has potentially dangerous consequences [40, 41]. Some studies have worked on disseminating of misinformation on the social media and its effect on people [42-45]. In order to fight misinformation, content-based filtering is the most common method [46, 47]. The availability of Deep Learning tools makes this easier and more scalable [48, 49].

3. Materials and Methods

This work relies on the COVID-19 related posts on the social media in Iran. The clustering methods are unsupervised methods based on the similarities between the data that can have a good performance in the present work. The clustering process displays an overview of the groups in a set of documents. According to the mentioned research goal, in this research work, a two-stage clustering method (a combination of two clustering algorithms, K-means, and SOM) is used. The proposed research method diagram is shown in Figure 1, and is described in the following sub-sections.

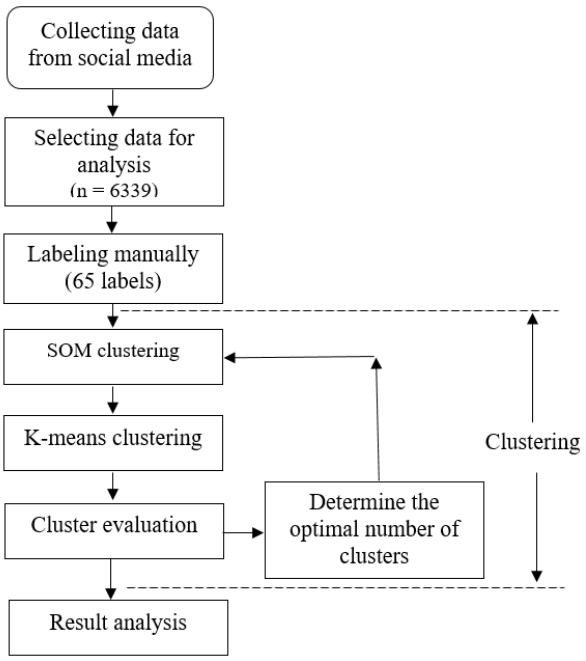

Figure 1. A flow chart of the work on social media posts about COVID-19 outbreak.

3.1. Collecting and labeling data

The posts and news focusing on COVID-19 from 21 January 2020 to 29 April 2020 were identified. All data including the full-text content, number of comments and likes, and owner information of

each post was extracted. These contents were in the Persian language and extracted from Telegram, Instagram, Twitter, and the domestic news published online. We call each post or news item a document.

For further analysis, 6339 posts were randomly selected, called total data. The total data includes 788 news items, 2073 posts on Instagram, 1630 Tweets, and 1848 posts on Telegram. In choosing the posts, care was taken that duplicate posts were selected only once, and also the promotional document or blank documents only containing hashtags (#) were not selected. Then the 6339 sampled data was labeled manually to 65 different labels[1]. These labels were grouped into five categories: topic, emotion, polarity, sub-topic (with or without polarity), and document value. The labels in the emotion category were based on plutchik's model [50] and included joy, sadness, trust, anticipation, fear, anger, surprise, disgust, stress, and other emotions. There were three types of poles to show the document's polarity: positive, negative, and neutral. The topic categories were scientific, religious, health, political, cultural and social, humor, fantasy, economics, and others. The sub-topic category was partitioned into 34 different sub-categories such as the poor/strong performance of Europe, the poor/good performance of America, the poor/good performance of institutions and people in Iran, and the positive/negative news and information. The document value category involved the empty post, the non-relevant post, the promotional post, and the incomplete post. It should be noted, in this work that we focused on three main label categories: topic, emotion, and polarity.

Each document may have several different labels simultaneously; for example, a document might be labeled as the health category and humor category. Each label is considered as a feature in the dataset with a binary value. That is, if a document has a specific label, the value of that label will be 1 for the document; otherwise, it is zero. Table 1 shows the set of features of the dataset, their type, and feature numbers in the experiments. The features include the document's ID, content, topic, emotion, and polarity.

The content of documents shows that about 40% of the documents are about the performance of institutions and people in the Corona crisis, which has led to different emotions in the society. We also observed that about 72% of the documents were in the field of health, culture, and society, 6% in religion, 3% in science, and the rest in politics or humor. The documents with non-positive polarity (negative or neutral) are about 50% of all documents.

Table 1. Dataset categories and specifications.

Category	Features	type	feature #
ID	Post Id,	Number	1
	Document_Id		2
Content of document	Content,	String	3
	User		4
Topic	Scientific,	Number	5
	religious,	{0 ,1}	6
	health,		7
	Political,		8
	cultural and social,		9
	humour and fantasy,		10
	economics,		11
	other topics		12
Emotion	Joy,	Number	13
	Sadness,	{0,1}	14
	fear,		15
	disgust,		16
	anger,		17
	surprise,		18
	trust,		19
	anticipation,		20
	other emotions,		21
	stress		22
Polarity	Positive,	Number	23
	negative,	{0,1}	24
	neutral		25

3.2. Clustering

The self-organizing map (SOM) is one of the most widely used clustering algorithms for data analysis [43]. It is trained using an unsupervised learning algorithm. SOM's specific property can map a high-dimensional input space to a lower-dimensional space and preserve the dataset's original topology while doing so.

SOM is an artificial neural network (ANN) consisting of an input layer and a competitive layer. The input layer receives a high-dimensional data and aggregates information to the competitive layer by weight vectors. The competitive layer is applied to produce the output results. The competitive layer neurons' arrangement takes three forms including 1D linear, 2D array, and 3D grid array. The 2D array is the most typical structure, and is similar to the image of the cerebral cortex [51].

The running process of the SOM network consists of the training stage and the mapping stage. During the training phase, the vectors from the dataset are presented to the map in a random

[1] The data set was collected in the Social Networks Laboratory of the Faculty of Electrical and Computer Engineering, the University of Tehran, with the support of the Cognitive Sciences and Technologies Council. Labeling of the dataset have been done in the Persian language processing laboratory of Shahid Beheshti University. The dataset is available at https://covidchallenge.cogc.ir/.

order. In the output layer for a particular input pattern, one winning node generates the most significant response. At the beginning of the training phase, which node will generate the maximum response in the output layer is unknown. When the category of the input pattern is modified, the 2D plane's winning node will also vary. The algorithm adjusts the weights of nodes in the output layer with a high number of training samples. Each output layer node is sensitive to a particular pattern class [52]. The trained SOM produces a mapping of the input space onto a 2D plane so that similar data points are placed close to each other [45].

3.3. K-means clustering

K-means is a well-known clustering algorithm. Its goal is to find groups of data with similar features, and assign them into several clusters.

K-means defines the clusters by (K) centers. A point belongs to a particular cluster if it is closer to the centroid of that cluster (in the Euclidean sense) other than any other centroid. In other words, K-means determines K centroids, and then assigns the data points to the nearest cluster. The term 'means' in K-means refers to computing the average of the data points, finding the centroid.

The cluster's centroids are calculated and updated at each iteration during training; for each cluster, the algorithm computes the weighted average of all data points, becoming the new centroid. We must set K (number of clusters) manually.

Training the K-means algorithm is as follows:

1) The algorithm takes a set of data points and the number of clusters (K) as the input.
2) It selects K centroids randomly, and calculates the distances between the data points and centroids. The distance may be measured by Manhattan, Euclidean or other distance measures. The choice of distance depends on the dataset and objective.
3) K-means is assigned data points to the cluster with the nearest distance from its centroid.
4) The algorithm updates the cluster centroids by computing the new mean values.

The algorithm continues looping until convergence. Typically, the training phase is done when the centroid of clusters stops moving or one could determine how many iterations should be done.

The implementation of K-means is fast and straightforward but it has some significant defects [53]. For example, the result of clustering mainly depends on the initial centroids; also it is required to determine the value of K. Algorithm operates poorly on the high-dimensional data.

3.4. Two-stage clustering

As mentioned earlier, K-means is simple and calculates quickly. The clustering results are still affected by the initial centroids, leading to the results potentially falling into the local minimum [47]. A clustering algorithm based on SOM and K-means uses the advantage of the automatic clustering of SOM. In order to specify the cluster centroids, the input data point will first be clustered by SOM; the SOM results will provide the initial center vector for K-means. Combining these algorithms can compensate for their shortcomings and improve the clustering output [44]. Thus training of this two-stage algorithm is as follows:

1) First, we cluster all the 6339 documents by SOM. This step's output will be a set of weight vectors. This step can help to decrease the training time by stopping SOM clustering before a complete convergence; for instance, 300 cycles are enough for SOM.
2) Secondly, we use the weight vectors of the SOM clustering outputs as the initial cluster centers. K-means is initialized, and then the data points are clustered with K-means.

Our two-stage algorithm has the self-organization characteristics of SOM and the high-efficiency features of K-means. Therefore, the SOM network's long convergence time and the possibility of mistakenly choosing cluster centroids of K-means are made up [44]. In this work, the labels of the post contents were applied as the inputs to SOM. The cluster centers in the output of SOM were used as the initial center points of K-means.

3.5. Determining optimal number of clusters

Defining the optimal number of clusters is a core problem in implementing various clustering methods; this parameter must be pre-specified before performing the clustering. This parameter is either identified by the users based on a prior knowledge or determined in a particular way [54].

At first, this work analyzed the results of the SOM clustering with 65 labels. The results obtained showed that one of the labels did not affect the training process. Thus we removed that label and continued our analysis with 64 labels by the two-

stage clustering method.
There are different clustering evaluation measures to assess the optimal number of clusters [49]. In this work, we used the Calinski-Harabasz (CH) measure. The optimal cluster number was calculated by maximizing the CH measure.

4. Experimental Results and Discussion

The implementations were built in Python version 3.7 and MATLAB R2015a. In order to run the algorithms such as clustering, the MATLAB software was used. The experiments and simulations were performed on a PC with a 2.50 GHz Intel Core i7 CPU and 8.0G RAM. The analysis of the experimental results is described in detail as below. Trust in the message content is an essential aspect of accepting or not accepting information in the social media. Low levels of trust often lead to a lack of citizen participation in the public activities. Thus we analyzed trust in the messages shared on the social media during the Corona pandemic in Iran. Figure 2 shows the correlation coefficient between each feature pair in Table 1. It should be noted that we considered the features numbered 5 to 25 in Table 1. According to Figure 2, there is a high correlation between the emotion of anger and disgust. The correlation between sadness and fear and stress is higher than the others, and fear is highly correlated with stress. There is a high correlation between disgust and anger, and the linear correlation between disgust and stress is small and around zero. We also observed a statistically significant linear relationship between the trust and economic labels and also between the trust and cultural and social labels.

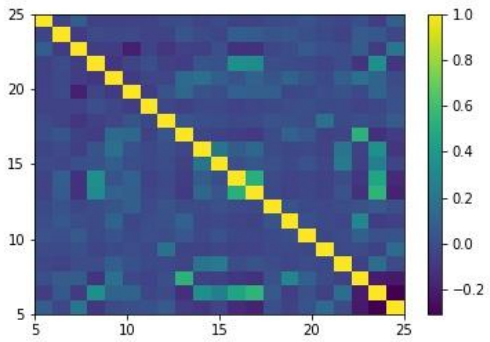

Figure 2. Correlation coefficient between each feature pair in Table 1 (features numbered 5 to 25).

The posts including the emotions of fear, sadness, disgust, stress or anger have a negative polarity. Hopeful post-corona documents lead to the feeling of joy in the society. Fear about Covid-19 takes the emotion of sadness in the society. The posts related to the lack of healthcare facility evoke the emotion fear. 27% of the documents including news, information, and statistics are negatively polarized.

4.1. Experimental results of clustering

This section presents the results of two-stage clustering, and provides a discussion of the results. The Calinski-Harabasz clustering evaluation was applied in order to find the optimal number of clusters. This measure was calculated for the number of clusters in the range of 1-150 in the two-stage clustering method. The results obtained showed that the optimal number of clusters was two, and the second number was four (Figure 3). Thus the data was grouped into two clusters using two-stage clustering, and the clusters were analyzed. We also investigated the proposed method's results by grouping in 64 clusters since there were 64 labels in the original dataset.
Trust is an important aspect of decision-making and accepting or not accepting the information posted on the social media [5]; thus we examined the trust label in the clusters. Investigation of the documents having a trust label showed that the trustful information was published on Instagram (about 50% of trustful information has been published on Instagram, and the rest has been published on other social networks).
When we partitioned the 6339 documents into two clusters, one cluster (cluster 1) included 4446 posts (70%), and another cluster (cluster 2) contained 1320 posts (30%). The number of observed labels is provided in Table 2. In this table, there are five columns: label name, label category, total data, cluster 1 (large cluster), and cluster 2 (small cluster). It should be noted that important labels are presented in this table.

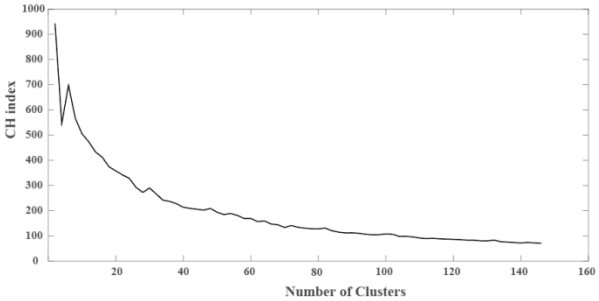

Figure 3. Calinski-Harabasz (CH) index versus different K-values.

In each row of the table, the number of labels seen in the total data, the large cluster, and the small cluster are shown. For example, according to the first row, the number of health labels in the total data is 2686. Given that there are 6339 documents in the total data, this label is about 42% of the

total data. In the large cluster, 1201 health labels have been observed, and in the small cluster, there are 1485 health labels.

As shown in Table 2, in the total data column, many posts are related to the health, cultural, and social topics. The non-positive polar (negative and neutral) posts are about 50% of the total data. We found out that 27% of the posts was related to the news, information, and statistic.

Table 2. Summary of grouping data into two clusters.

No.	Label name	Label category	Total data	Cluster 1 (large cluster)	Cluster 2 (small cluster)
1	Health	Topic	2686	1201	1485
2	Cultural and social	Topic	1906	1334	572
3	Economic	Topic	482	399	83
4	Negative	Polarity	1511	1426	85
5	Neutral	Polarity	1609	615	994
6	Trust	Emotion	249	233	16
7	Fear	Emotion	331	293	38
8	Sadness	Emotion	436	393	43
9	Anger	Emotion	636	621	15
10	Disgust	Emotion	498	488	10

There are many non-positive polar (negative and neutral) posts in the large cluster compared to the small cluster. A majority of the documents with the trust label are in the large cluster. 50% of the documents in the large cluster show the poor performance of institutions, leading to negative feelings such as fear, anger, sadness, and disgust. In the small cluster, about 70% of the documents are in the field of health. We found out that 38% of the documents included the content showing the good performance of institutions. In this cluster, and 50% of the documents was neutrally polarized and did not cause negative emotions (fear, anger, and disgust).

According to the result in cluster 1, the posts with non-positive polarity show the poor performance of institutions in the health or social topics. These posts decreased the society's trust. On the other hand, in cluster 2, the posts related to the institutions' good performance in health did not affect the people's trust, and their polarity has often been neutral. We also investigated the performance of our clustering method for 64 clusters. When the data was partitioned into 64 clusters, we examined two clusters with the highest and lowest trust labels. The cluster with the lowest number of trust labels had 69 documents, while the cluster with the highest number of trust labels had 92 documents. The number of different labels in these two clusters is shown in Table 3. In this table, there are four columns: label name, label category, cluster 1 (with the lowest trust label), and cluster 2 (with the highest trust label).

The rows of the table show the number of labels seen in cluster 1 and cluster 2. For example, according to the first row, the health label is in the topic category and has been seen in 31 documents in cluster 1, while 78 documents have the health label in cluster 2. There were 45 neutral posts in cluster 1. The majority of posts (about 70%) shared the news, information, and statistics.

Table 3. Summary of clustering dataset into 64 clusters.

No.	Label name	Label category	Cluster 1 (with lowest trust label)	Cluster 2 (with highest trust label)
1	Health	Topic	31	78
2	Cultural and social	Topic	5	13
3	Economic	Topic	0	50
3	Negative	Polarity	1	88
4	Neutral	Polarity	45	1
5	Trust	Emotion	0	74
6	Fear	Emotion	0	3
7	Sadness	Emotion	0	3
8	Anger	Emotion	0	15
9	Disgust	Emotion	0	10

The analysis showed that the trusted posts were published through Instagram and news channels in cluster 1. In cluster 2, there were 74 posts with the trust labels and 88 posts with the negative polarity labels. In the cluster, the documents in health show the poor performance of Iran's institutions, and have decreased the society trust. Also the number of documents with negative polarity is high. These documents have led to negative emotions such as anger and sadness.

We found that hopeful posts were more trusted than the other messages. On the other hand, religion has always been an essential factor in building trust in the societies [55]. The results obtained showed that religion could build trust by emphasizing on the national identity components. It was expected that the posts shared in health increase people’s trust but unfortunately, even the trust in health posts was less than it in the socio-cultural topics. The findings from this research work indicated that Twitter influenced on destroying public trust. Also the social activists were more effective than the politicians, religious, and scientific characters in destroying or building trust. The social activists posed most of the messages shared in Covid-19; they have effectively built public trust. The results obtained showed that if there were messages about improving health, the audience would have distrusted them.

According to the experimental results, we suggest that the non-political messages should be used to deliver health messages to the people. We also

found that if there were any evidence of health improvement, the audience would have distrusted them. People did not accept the documents describing the state of health or readiness to face this pandemic in the country. Also the documents against politicians and their impact on the spread of coronavirus were trusted.

The results show that the posts related to the poor performance of institutions and governments lead to disgust, and a lack of government transparency leads to anger. The news and statistic related to the poor performance of the government lead to the reduction of public trust.

The social activists have been very influential in gaining and destroying the post audiences' trust (about 62% of trustful information is in the culture and society, which is about the social activists and the Corona crisis). The documents sharing a message of hope have been more trusted than the other documents. Religion has always been one of the essential factors in gaining the social trust. The analysis of results showed that religion could build trust by using the components of national identity. Also the analysis showed that people had trust in the documents opposing the politicians, although the corona pandemic is a disease and, as expected, the documents published by the health organizations should be trusted. Unfortunately, trust in the health topics has been less than in the socio-cultural field. It can be concluded that to improve the trust that has been destroyed by different media, at first, the sources of distrust should be identified, and then an appropriate confrontation should be selected.

4.2. Comparison with other clustering methods

In this section, the results of the proposed method (a combination of K-means and SOM) are compared with the other clustering methods like K-means, fuzzy C-means [56], and SOM, separately. Fuzzy C-means is a clustering method where each data point is assigned different probability scores to belong to several clusters. A probability score shows a degree of membership to a cluster. At the end, it is assumed that each data belongs to the cluster with the highest degree of membership [56].

The clustering results for the input parameter 2 (2 clusters) are presented in Tables 4-6. Each column displays information about a cluster including the number of labels seen in that cluster. Table 4 shows the results of clustering using K-means on 6399 documents. According to the K-means algorithm, there are 3857 documents in the first cluster and 2482 documents in the second cluster. The results obtained show that the labels are scattered in both clusters. According to Figure 2, the clusters are expected to contain the label of fear and sadness together. It is also expected that a significant number of economic and trust labels to be observed in the same cluster, which is not seen in Table 4.

Table 5 shows the results of the SOM clustering algorithm. The results show that the labels are scattered in the clusters, and significant linear relationships between the trust and other labels (such as economic or cultural and social labels) are not observed in Table 5.

Table 4. Summary of grouping data into two clusters using K-means algorithm.

No.	Label name	Label category	Cluster 1	Cluster 2
1	Health	Topic	2187	499
2	Cultural and social	Topic	1312	594
3	Economic	Topic	202	280
4	Negative	Polarity	58	1453
5	Neutral	Polarity	1094	515
6	Trust	Emotion	133	116
7	Fear	Emotion	97	234
8	Sadness	Emotion	337	99
9	Anger	Emotion	420	216
10	Disgust	Emotion	474	24

Table 5. Summary of grouping data into two clusters using SOM.

No.	Label name	Label category	Cluster 1	Cluster 2
1	Health	Topic	1593	1093
2	Cultural and social	Topic	676	1230
3	Economic	Topic	235	247
4	Negative	Polarity	484	1027
5	Neutral	Polarity	1045	564
6	Trust	Emotion	120	129
7	Fear	Emotion	88	243
8	Sadness	Emotion	103	333
9	Anger	Emotion	552	84
10	Disgust	Emotion	289	209

Table 6. Summary of grouping data into two clusters using fuzzy C-means.

No.	Label name	Label category	Cluster 1	Cluster 2
1	Health	Topic	1230	1456
2	Cultural and social	Topic	595	1311
3	Economic	Topic	221	261
4	Negative	Polarity	18	1493
5	Neutral	Polarity	1037	572
6	Trust	Emotion	180	69
7	Fear	Emotion	264	67
8	Sadness	Emotion	57	379
9	Anger	Emotion	623	13
10	Disgust	Emotion	14	484

Table 6 shows the results of the fuzzy C-means algorithm. The labels are scattered in the clusters. However, in our proposed method, the correlation between the labels (seen in Figure 2) can be observed in clustering, and the characteristics of the documents in each cluster can be analysed, which shows the efficiency of the proposed

method compared to each one of the methods including K-means, SOM, and fuzzy C-means separately.

5. Conclusion

Social media is a fundamental source of information on the day-to-day events, and reflects the social interactions and responses [57]. It also has played a significant role in affecting the public emotions and attitudes during the COVID-19 crisis. Therefore, analysis of the posts shared on the social media could be used as a tool to assess our community's and governments' requirements to prepare during the pandemics. Analyzing the social media can help the policy-makers and healthcare organizations assess the societies' requirements, and address them appropriately and relevantly. Therefore, we investigated the Persian-language posts shared on the social media (Twitter, Instagram, Telegram, and online news) about the COVID-19 crisis in Iran. The data was analyzed using a two-stage clustering method based on the SOM and K-means algorithms. Since the public trust is an essential aspect of deciding whether or not to accept the information shared on the social media, the trust label in clusters was examined.

We concluded that first, the sources of creating distrust should be identified in order to improve the destroyed trust, and then a suitable confrontation should be selected. According to the results obtained, we suggested that the non-political messages should be used to share the messages related to the people's health status in the community. Our results can help the health care providers to be prepared for the future epidemics or health situations. It is also essential to engage the Ministry of Health and Medical Education. They may use our results to decide on the best performance in the present situation and similar situations. In addition, this paper could be of interest to the researchers in emotion analysis, neuroscience, social media, and data analysis. Similar works could also be done in the industry and business. In the future work, we will apply other machine learning methods (such as deep learning) for the social media analysis.

Acknowledgment

The authors would like to extend their gratitude to the Cognitive Sciences and Technologies Council, Tehran, Iran for providing the datasets.

References

[1] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi, and Z. Shah, "Top concerns of tweeters during the COVID-19 pandemic: infoveillance study" *Journal of medical Internet research*, vol. 22, no. 4, pp. e19016, 2020.

[2] A. R. Ahmad and H. R. Murad, "The impact of social media on panic during the COVID-19 pandemic in Iraqi Kurdistan: online questionnaire study" *Journal of Medical Internet Research*, vol. 22, no. 5, p. e19556, 2020.

[3] X. Ji, S. A. Chun, and J. Geller, "Monitoring public health concerns using twitter sentiment classifications" in *2013 IEEE International Conference on Healthcare Informatics*, 2013, pp. 335-344.

[4] M. Smith, D. A. Broniatowski, M. J. Paul, and M. Dredze, "Towards real-time measurement of public epidemic awareness: Monitoring influenza awareness through twitter" in *AAAI spring symposium on observational studies through social media and other human-generated content*, 2016.

[5] W. Sherchan, S. Nepal, and C. Paris, "A survey of trust in social networks" *Computing Surveys*, vol. 45, no. 4, pp. 1-33, 2013.

[6] L. Mollema *et al.*, "Disease detection or public opinion reflection? Content analysis of tweets, other social media, and online newspapers during the measles outbreak in The Netherlands in 2013" *Journal of medical Internet research*, vol. 17, no. 5, pp. e128, 2015.

[7] A. J. Lazard, E. Scheinfeld, J. M. Bernhardt, G. B. Wilcox, and M. Suran, "Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat" *American journal of infection control*, vol. 43, no. 10, pp. 1109-1111, 2015.

[8] T. Tran and K. Lee, "Understanding citizen reactions and Ebola-related information propagation on social media" in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2016, pp. 106-111.

[9] M. Miller, T. Banerjee, R. Muppalla, W. Romine, and A. Sheth, "What are people tweeting about Zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention" *JMIR public health and surveillance*, vol. 3, no. 2, p. e38, 2017.

[10] L.-C. Chen, C.-M. Lee, and M.-Y. Chen, "Exploration of social media for sentiment analysis using deep learning" *Soft Computing*, vol. 24, no. 11, pp. 8187-8197, 2020.

[11] G. Li and F. Liu, "Sentiment analysis based on clustering: a framework in improving accuracy and recognizing neutral opinions" *Applied intelligence*, vol. 40, no. 3, pp. 441-452, 2014.

[12] V. K. Vijayan, K. Bindu, and L. Parameswaran, "A comprehensive study of text classification algorithms" in *2017 International Conference on Advances in Computing, Communications and Informatics*, 2017, pp. 1109-1113.

[13] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, "Scalable sentiment classification for big data analysis using naive bayes classifier" in *2013 IEEE international conference on big data*, 2013, pp. 99-104.

[14] E. Boiy, P. Hens, K. Deschacht, and M. F. Moens "Automatic Sentiment Analysis in On-line Text" in *ELPUB*, 2007, pp. 349-360.

[15] W. Ramadhan, S. A. Novianty, and S. C. Setianingsih, "Sentiment analysis using multinomial logistic regression" in *2017 International Conference on Control, Electronics, Renewable Energy and Communications*, 2017, pp. 46-49.

[16] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey" *Information*, vol. 10, no. 4, pp. 150, 2019.

[17] L. Nemes and A. Kiss, "Social media sentiment analysis based on COVID-19" *Journal of Information and Telecommunication*, pp. 1-15, 2020.

[18] G. Li and F. Liu, "Application of a clustering method on sentiment analysis" *Journal of Information Science*, vol. 38, no. 2, pp. 127-139, 2012.

[19] S. Wu, Y. Liu, J. Wang, and Q. Li, "Sentiment analysis method based on Kmeans and online transfer learning" *Cmccomputers Materials and Continua*, vol. 60, no. 3, pp. 1207-1222, 2019.

[20] F. Kaveh-Yazdy and S. Zarifzadeh, "Track Iran's national COVID-19 response committee's major concerns using two-stage unsupervised topic modeling" *International journal of medical informatics*, vol. 145, pp. 104309, 2021.

[21] K. E. Matsa and E. Shearer, "News use across social media platforms 2018" *Pew Research Center*, vol. 10, 2018.

[22] P. Hitlin and K. Olmstead, "The science people see on social media. Pew Research Center" ed, 2018.

[23] X. Ye, S. Li, X. Yang, and C. Qin, "Use of social media for the detection and analysis of infectious diseases in China" *International Journal of Geo-Information*, vol. 5, no. 9, pp. 156, 2016.

[24] I. C.-H. Fung *et al.*, "Pedagogical Demonstration of twitter data analysis: A case study of world AIDS day, 2014" *Data*, vol. 4, no. 2, pp. 84, 2019.

[25] E. H.-J. Kim, Y. K. Jeong, Y. Kim, K. Y. Kang, and M. Song, "Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news" *Journal of Information Science*, vol. 42, no. 6, pp. 763-781, 2016.

[26] J. Samuel, M. Rahman, G. Ali, Y. Samuel, and A. Pelaez, "Feeling Like it is Time to Reopen Now? COVID-19 New Normal Scenarios based on Reopening Sentiment Analytics" *Nawaz and Samuel, Yana and Pelaez, Alexander, Feeling Like it is Time to Reopen Now*, 2020.

[27] A. Mosam, S. Goldstein, A. Erzse, A. Tugendhaft, and K. Hofman, "Building trust during COVID 19: Value-driven and ethical priority-setting" *SAMJ: South African Medical Journal*, vol. 110, no. 6, pp. 1-4, 2020.

[28] A. Oksanen, M. Kaakinen, R. Latikka, I. Savolainen, N. Savela, and A. Koivula, "Regulation and trust: 3-month follow-up study on COVID-19 mortality in 25 European countries" *JMIR Public Health and Surveillance*, vol. 6, no. 2, pp. e19218, 2020.

[29] L. Li *et al.*, "Characterizing the propagation of situational information in social media during covid-19 epidemic: A case study on weibo" *IEEE Transactions on Computational Social Systems*, vol. 7, no. 2, pp. 556-562, 2020.

[30] C. E. Lopez, M. Vasu, and C. Gallemore, "Understanding the perception of COVID-19 policies by mining a multilanguage Twitter dataset" *arXiv preprint arXiv:2003.10359*, 2020.

[31] Z.-Y. Tao *et al.*, "Nature and diffusion of COVID-19–related oral health information on Chinese social media: analysis of tweets on weibo" *Journal of Medical Internet Research*, vol. 22, no. 6, pp. e19981, 2020.

[32] E. Massaad and P. Cherfan, "Social media data analytics on telehealth during the COVID-19 pandemic" *Cureus*, vol. 12, no. 4, 2020.

[33] V. P. Vinogradac, J. P. Vukičević, and I. C. Mraović, "Value system as a factor of young people's trust in education during the Covid-19 pandemic in three countries of southeast europe" *Društvene i humanističke studije,* vol. 5, no. 3 (12), pp. 331-354, 2020.

[34] D. H. Balog-Way and K. A. McComas, "COVID-19: Reflections on trust, tradeoffs, and preparedness" *Journal of Risk Research*, vol. 23, no. 7-8, pp. 838-848, 2020.

[35] A. Deslatte, "The erosion of trust during a global pandemic and how public administrators should counter it" *The American Review of Public Administration*, vol. 50, no. 6-7, pp. 489-496, 2020.

[36] P. B. Forsyth, C. M. Adams, and W. K. Hoy, "Collective trust" *Why schools can't improve,* 2011.

[37] D. H. Balog-Way and K. A. McComas, "COVID-19: Reflections on trust, tradeoffs, and preparedness" *Journal of Risk Research*, pp. 1-11, 2020.

[38] D. Pastor-Escuredo and C. Tarazona, "Characterizing information leaders in Twitter during COVID-19 crisis" *arXiv preprint arXiv:2005.07266*, 2020.

[39] J. Zarocostas, "How to fight an infodemic," *The Lancet*, vol. 395, no. 10225, pp. 676, 2020.

[40] L. Bode and E. K. Vraga, "See something, say something: Correction of global health misinformation

on social media" *Health communication*, vol. 33, no. 9, pp. 1131-1140, 2018.

[41] P. M. Waszak, W. Kasprzycka-Waszak, and A. Kubanek, "The spread of medical fake news in social media–the pilot quantitative study" *Health policy and technology*, vol. 7, no. 2, pp. 115-118, 2018.

[42] L. Singh *et al.*, "A first look at COVID-19 information and misinformation sharing on Twitter," *arXiv preprint arXiv:2003.13907*, 2020.

[43] P. Wicke and M. M. Bolognesi, "Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter" *arXiv preprint arXiv:2004.06986*, 2020.

[44] S. Llewellyn, "Covid-19: how to be careful with trust and expertise on social media," *BMJ*, vol. 368, 2020.

[45] R. Kouzy *et al.*, "Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter" *Cureus*, vol. 12, no. 3, 2020.

[46] F. Pierri and S. Ceri, "False news on social media: a data-driven survey" *ACM Sigmod Record*, vol. 48, no. 2, pp. 18-27, 2019.

[47] B. Ghanem, P. Rosso, and F. Rangel, "An emotional analysis of false information in social media and news articles" *ACM Transactions on Internet Technology*, vol. 20, no. 2, pp. 1-18, 2020.

[48] K. Popat, S. Mukherjee, A. Yates, and G. Weikum, "Declare: Debunking fake news and false claims using evidence-aware deep learning" *arXiv preprint arXiv:1809.06416*, 2018.

[49] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection" in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 797-806.

[50] R. Plutchik, "A general psychoevolutionary theory of emotion" in *Theories of emotion*: Elsevier, 1980, pp. 3-33.

[51] V. Khachidze, T. Wang, S. Siddiqui, V. Liu, S. Cappuccio, and A. Lim, "Contemporary research on E-business technology and strategy" in *Conference proceedings iCETS*, 2012, pp. 43.

[52] Y.-C. Liu, M. Liu, and X.-L. Wang, "Application of self-organizing maps in text clustering: a review chapter", 2012.

[53] Y. P. Raykov, A. Boukouvalas, F. Baig, and M. A. Little, "What to do when k-means clustering fails: a simple yet principled alternative algorithm" *PloS one*, vol. 11, no. 9, pp. e0162259, 2016.

[54] C. Patil and I. Baidari, "Estimating the optimal number of clusters k in a dataset using data depth" *Data Science and Engineering*, vol. 4, no. 2, pp. 132-140, 2019.

[55] S. H. Chuah, S. Gächter, R. Hoffmann, and J. H. Tan, "Religion, discrimination and trust across three cultures" *European Economic Review*, vol. 90, pp. 280-301, 2016.

[56] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm" *Computers and geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[57] A. Lakizadeh and E. Moradizadeh, "Text sentiment classification based on separate embedding of aspect and context" *Journal of AI and Data Mining*, vol. 10, no.1 , pp. 139-149, 2022.

تحلیل پست‌ها در رسانه‌های اجتماعی در زمان شیوع ویروس کرونا در ایران با استفاده از خوشه‌بندی

فاطمه امیری[۱]، سمیرا عباسی[۲،*] و محبوبه بابایی[۳]

[۱] گروه مهندسی کامپیوتر، دانشگاه صنعتی همدان، همدان، ایران .

[۲] گروه مهندسی پزشکی، دانشگاه صنعتی همدان، همدان، ایران.

[۳] جامعه توسعه روستایی، دانشگاه تهران، تهران، ایران.

ارسال ۲۰۲۱/۱۰/۱۷؛ بازنگری ۲۰۲۱/۰۱/۰۵؛ پذیرش ۲۰۲۲/۰۲/۲۱

چکیده:

در طول بحران کرونا، با طیف گسترده‌ای از افکار، احساسات و رفتارها در رسانه‌های اجتماعی روبرو هستیم که نقش مهمی در انتشار اطلاعات در مورد کرونا بازی می‌کنند. می‌توان از اطلاعات موثق به همراه پیام‌های امید برای کنترل احساسات و عکس‌العمل‌های مردم در زمان همه‌گیری استفاده کرد. کار حاضر به بررسی تاب‌آوری جامعه ایران در مواجهه با بحران کرونا می‌پردازد و راهکاری برای ارتقاء تاب‌آوری در شرایط مشابه ارائه می‌دهد. این مقاله پست‌ها و اخبار مربوط به بیماری همه‌گیر کوید ۱۹ در ایران را بررسی می‌کند تا مشخص شود کدام پیام‌ها و مراجع باعث نگرانی در جامعه می‌شوند و چگونه می‌توان آنها را اصلاح کرد؟ و همچنین کدام مراجع مورد اعتمادترین ناشران هستند؟ برای تحلیل داده‌ها از روش‌های تحلیل شبکه‌های اجتماعی مانند خوشه‌بندی استفاده می‌شود. در این کار، از روش خوشه‌بندی دو مرحله‌ای مبتنی بر نقشه خود سازمانده و K-میانگین استفاده می‌شود. به دلیل اهمیت اعتماد اجتماعی در پذیرش پیام‌ها، اعتماد به پست‌های موجود در مجموعه داده بررسی می‌شود. نتایج بدست آمده نشان می‌دهد که اعتماد به پست‌های مربوط به حوزه سلامت کمتر از موضوعات اجتماعی و فرهنگی است. پست‌های مورد اعتماد در اینستاگرام و سایت‌های خبری به اشتراک گذاشته می‌شوند. پست‌های بهداشتی و فرهنگی با قطبیت منفی بر اعتماد مردم تأثیر گذاشته و منجر به احساسات منفی مانند ترس، انزجار، ناراحتی و عصبانیت می‌شوند. از اینرو، پیشنهاد می‌شود که از گفتمان‌های غیرسیاسی برای به اشتراک گذاشتن موضوعات در زمینه سلامت استفاده شود.

کلمات کلیدی: خوشه‌بندی، کوید ۱۹، ایران، رسانه های اجتماعی، اعتماد اجتماعی.