Neural Computing and Applications
https://doi.org/10.1007/s00521-021-06109-0

ORIGINAL ARTICLE

A state-of-art optimization method for analyzing the tweets of earthquake-prone region

Nazmiye Eligüzel[1] · Cihan Çetinkaya[2] · Türkay Dereli[3]

Received: 22 July 2020 / Accepted: 5 May 2021

Abstract

With the increase in accumulated data and usage of the Internet, social media such as Twitter has become a fundamental tool to access all kinds of information. Therefore, it can be expressed that processing, preparing data, and eliminating unnecessary information on Twitter gains its importance rapidly. In particular, it is very important to analyze the information and make it available in emergencies such as disasters. In the proposed study, an earthquake with the magnitude of Mw = 6.8 on the Richter scale that occurred on January 24, 2020, in Elazig province, Turkey, is analyzed in detail. Tweets under twelve hashtags are clustered separately by utilizing the Social Spider Optimization (SSO) algorithm with some modifications. The sum-of intra-cluster distances (SICD) is utilized to measure the performance of the proposed clustering algorithm. In addition, SICD, which works in a way of assigning a new solution to its nearest node, is used as an integer programming model to be solved with the GUROBI package program on the test data-sets. Optimal results are gathered and compared with the proposed SSO results. In the study, center tweets with optimal results are found by utilizing modified SSO. Moreover, results of the proposed SSO algorithm are compared with the K-means clustering technique which is the most popular clustering technique. The proposed SSO algorithm gives better results. Hereby, the general situation of society after an earthquake is deduced to provide moral and material supports.

Keywords Clustering · Latent semantic analyses · Social spider optimization · Twitter

1 Introduction

Earthquakes have always been part of life, which occur every day all over the world. Some of the earthquakes cause loss of life and properties or some psychological effects on victims. According to the statistics, the number of earthquakes (Mw $\geq$ 5 on the Richter scale) worldwide from 2000 to 2019 can be seen in Fig. 1.[1]

In 2019, a total of 1637 earthquakes (Mw = 5 +) took place in the world. As can be seen from Fig. 1, the number of earthquakes is very high in the world. Turkey is one of the countries that experienced a lot of earthquakes every year. Elazig earthquake is one of them. In this study, the Elazig earthquake is taken into consideration. Brief information is given about Elazig Earthquake as follows:

Elazig and Malatya are the provinces located in Eastern Turkey. Turkey is in the earthquake region and these provinces are in the 2nd -degree earthquake zone. Turkey fault line map can be seen in Fig. 2. Elazig earthquake is one of the earthquakes that caused the loss of life and property. The Elazig earthquake with the magnitude of Mw = 6.8 on the Richter scale occurred on January 24, 2020, in Turkey. A total of 2795 aftershocks and 21 earthquakes measuring more than 4 on the Richter scale

✉ Nazmiye Eligüzel
nazmiye@gantep.edu.tr

[1] Industrial Engineering, Gaziantep University,
27310 Gaziantep, Turkey

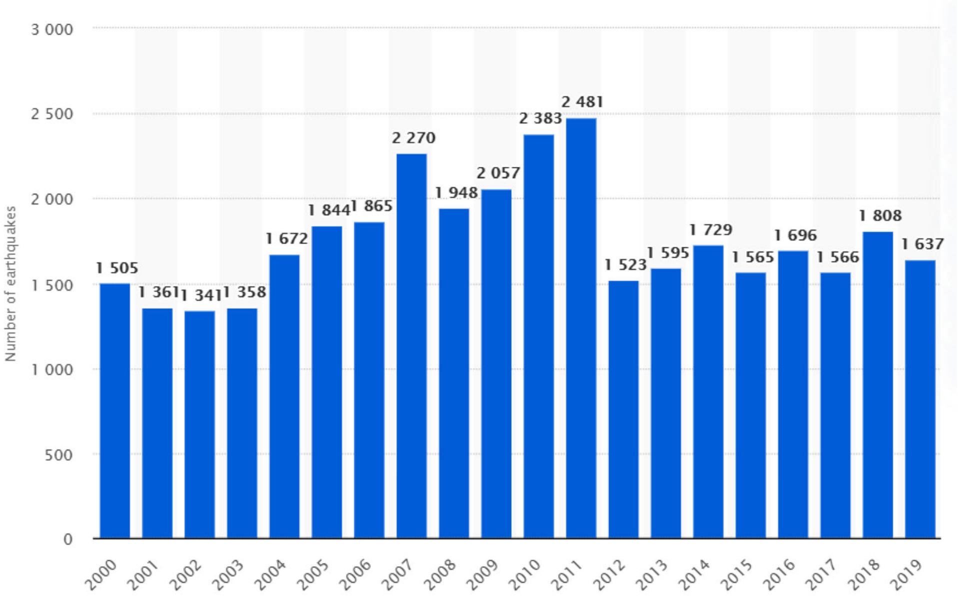
[2] Department of Management Information Systems, Adana
Alparslan Türkeş Science and Technology University,
01250 Adana, Turkey

[3] Office of the President, Hasan Kalyoncu University,
Gaziantep, Turkey

[1] Statista, https://www.statista.com/statistics/263105/development-of-the-number-of-earthquakes-worldwide-since-2000/.

Published online: 24 May 2021

Fig. 1 Number of the earthquakes in the World (Statista)



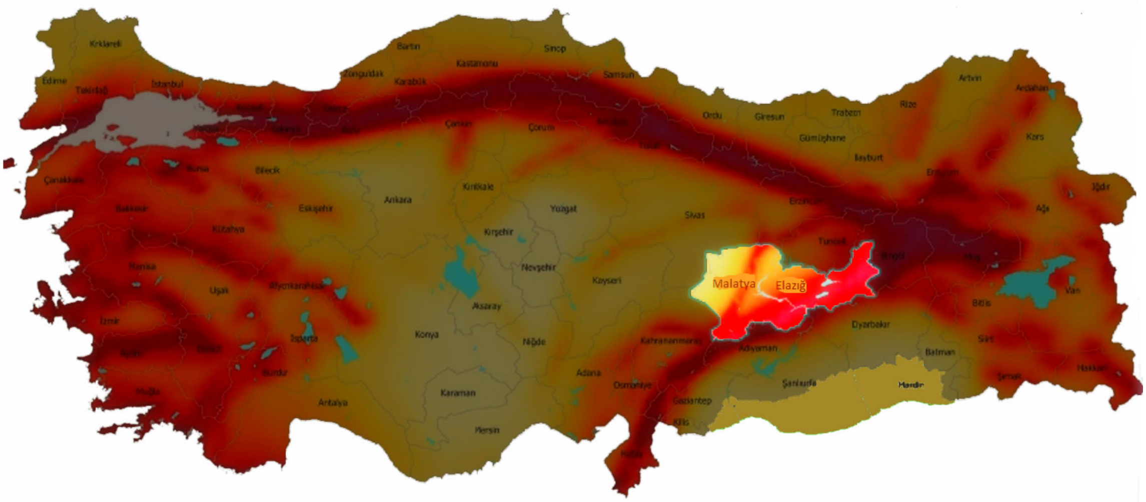


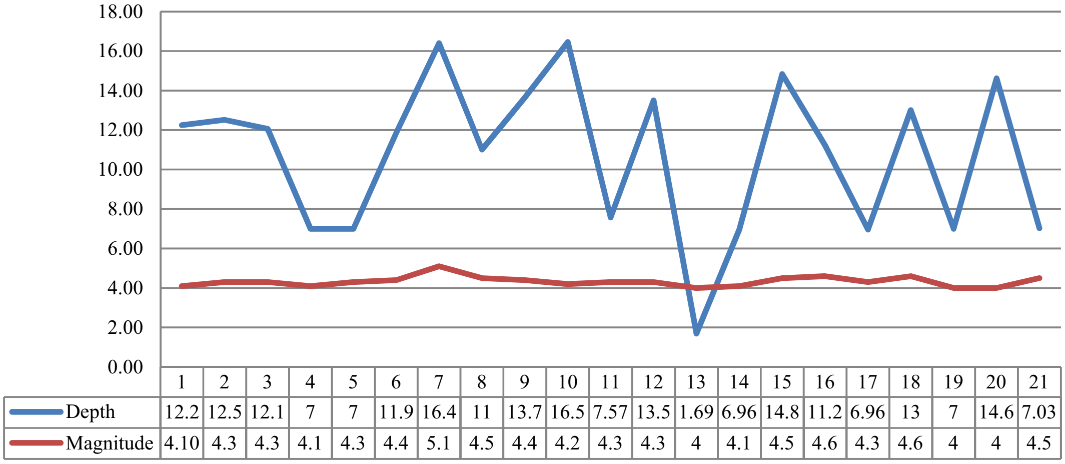
Fig. 2 Turkey fault line map

occurred after the earthquake, within 3 days in the region (Elazig and Malatya). In Fig. 3, the aftershocks in the earthquake region are given. In this disaster, 41 people died and 45 people were rescued from the debris. A total of 1631 people demand health care after the earthquake. Right after the earthquake, a total of 7245 personnel, 636 vehicles, and 24 search and rescue dogs participated in activities in the region under the coordination of the Disaster and Emergency Management Authority (AFAD). Different non-governmental organizations (NGOs) such as the Turkish Red Crescent Society, AKUT (Search and Rescue Association) and IHH (Humanitarian Relief Foundation) also participated in search and rescue operations. A total of 3846 shelters were distributed in the shelter area and 19,409 shelters were distributed individually in Elazig. In Malatya, a total of 5948 shelters were distributed individually. Five thousand eight hundred and fifty-one containers were distributed to Elazig, 4790 of them were placed in temporary shelter areas in the earthquake district at 6 different locations. In the region, psycho-social support activities were provided.[2]

In the fault line map, dark red areas represent regions with high earthquake potential, and highlighted part of the map indicates the recent earthquake region.

In Fig. 3, shocks after the earthquake are demonstrated with respect to depth. From Fig. 3, it can be expressed that depth is fluctuating and magnitude remains constant. When

[2] AFAD, https://www.afad.gov.tr/elazig-depremi-sonrasi-yapilan-yardimlar-merkezicerik.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Depth	12.2	12.5	12.1	7	7	11.9	16.4	11	13.7	16.5	7.57	13.5	1.69	6.96	14.8	11.2	6.96	13	7	14.6	7.03
Magnitude	4.10	4.3	4.3	4.1	4.3	4.4	5.1	4.5	4.4	4.2	4.3	4.3	4	4.1	4.5	4.6	4.3	4.6	4	4	4.5

Fig. 3 Aftershocks in the region for magnitude interval ≥ 4

the depth values from Fig. 3 taken into consideration, it is observed that aftershocks occurred too close to the surface.

Earthquakes cause physical and psychological effects on states, nations, and individuals. Therefore, the effects of earthquakes on societies should be examined and analyzed to improve physical and mental health. It is very important to analyze the information on social media and make it available in emergencies such as earthquakes. Social media is one of the tools utilized to share information. In 2019, it was concluded that almost 2.95 billion people were utilizing social media worldwide, a number projected to incline to almost 3.43 billion in 2023.[3] Twitter is one of the most commonly used social media tools among others that enable learning and conveying valuable information. As of the first quarter of 2019, there was an average of 330 million Twitter monthly active users.[4] During and after an earthquake, large volumes of data are generated by social media, especially by Twitter users. Analyses of these data are significant to find out problems and developing a solution. Therefore, data clustering is one of the most utilized unsupervised classification mechanisms in data mining for summarization huge amount of data-sets [1, 2]. Recently, swarm intelligence algorithms such as the Artificial Bee Colony approach [3], Particle Swarm Optimization (PSO) approach [4], Ant Colony Optimization (ACO) approach [5] and evolutionary algorithms such as the Genetic Algorithm approach [6, 7] have been applied by researchers to solve high-dimensional optimization problems in text clustering area. Nature-inspired algorithms have exploration and exploitation capabilities by moving the current solution to the near-global solution [8]. SSO (Social Spider Optimization) is among the swarm intelligence algorithms that demonstrate the communal behavior of organisms. SSO achieves a good balance between local and global searches [1, 8]. Therefore, in the proposed study, the SSO approach is used for tweet clustering. The most popular clustering technique is K-means clustering in the literature. However, it leads to some problems such as getting stuck into local optima [9]. Therefore, the SSO algorithm is utilized to overcome this problem. The main reason behind choosing SSO drives from its nature of mitigating the information loss and exchanges dynamically. During the optimization, current positions and exploration process lead to different searching behaviors which provide handling several local optima and converging global optima [10]. Another reason drives by its uniform structure of population which provides optimum search in complex problems [11]. All in all, SSO is preferred due to its ability to balance exploration and exploitation. In addition, its nature of providing a uniform structure of population for optimum searches.

In this paper, the SSO approach is proposed with the LSA method to cluster Elazig Earthquake Tweets and analyze these tweets in detail. The aim of the study is to deduce the general situation of society after the earthquake in order to provide moral and material supports.

Word embeddings have been utilized in a wide variety of fields such as artificial intelligence [12], information retrieval [13, 14], sentiment analyses [15], and psychiatric [16]. There are several word embedding techniques such as Word2Vec, Glove, Latent Semantic Analysis (LSA), Doc2Vec, skip-gram, fasttex. To represent a text, the most common and simple technique is the vector space model that demonstrates each unique term in the vocabulary as one dimension in the feature space. However, it needs a huge number of features to demonstrate high dimensions

[3] Statista, https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[4] Statista, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users//.

[6]. LSA utilizes the singular value decomposition (SVD) that has the capability of mapping high-dimensional count vectors and lower-dimensional representation [17]. LSA is a counter-based technique. Although prediction-based models have severely raised in popularity, it is not certain whether they demonstrate better performance than classical counter-based models [18]. Pennington et al. [19] demonstrated that traditional methods especially LSA can be more useful than Word2Vec which is a prediction-based technique. LSA is a more stable technique compared with Word2Vec and independent from corpus size [20, 21]. According to Altszyler et al. [18] when the corpus size is decreased (under 10 million words) LSA proves the good performance. Therefore, it becomes the more appropriate tool for a small corpus size. In the proposed study, our corpus size is not very large because the data are analyzed based on the hashtags. LSA word representation quality can be increased by decreasing dimensionality while speeding up the processing time. With respect to the aforementioned facts, LSA is utilized in the proposed study.

The study is organized as follows. In Sect. 2, related works are given on data clustering. In Sect. 3, the background of the SSO and LSA are introduced. Section 4 presents the methodology including pre-processing stage and experimental results of the SSO algorithm with LSA. In Sect. 5, results and discussions are presented. Finally, in Sect. 6, conclusions are summarized.

2 Literature review

There are several studies in the literature for text clustering and feature selection by utilizing meta-heuristic algorithms. These meta-heuristic algorithms can be considered mostly under two frames which are swarm intelligence and evolutionary-based algorithms.

There are several studies in the literature that represent the implementation of PSO, which is one of the swarm intelligence algorithms, for text and document clustering. A new feature selection technique by utilizing the PSO algorithm was proposed to solve feature selection problems by generating a novel sub-set of informative text [4]. The experimental analysis demonstrated the effectiveness of this new sub-set of features by improving the performance of the document clustering and reducing the computational time. This algorithm utilized the term frequency-inverse document frequency (TF-IDF) as a feature selector. Hua and Wei [22] proposed a hybrid technique that combines PSO and K-Means algorithms for clustering Mongolian elements which is a new technology that integrates artificial intelligence with image processing. They demonstrated the effectiveness of their method by utilizing the internal and external evaluation indexes as an evaluation tool. Janani and Vijayarani [23] proposed a new Spectral Clustering algorithm, which is widely applied in the machine learning area with PSO to improve document clustering. Their study aims to overcome the huge dimension of text documents. Three algorithms were used to make a comparison with the proposed algorithm; Spherical K-means technique, Expectation–Maximization technique (EM), and standard PSO approach. The Spectral Clustering algorithm with PSO provided better clustering accuracy.

ACO is the swarm intelligence-based technique that has also been utilized for text clustering in the literature. Nema et al. [24] focused on multi-label text categorization where multiple features participate in a common class that causes some problems related to appropriate feature selection. The feature optimization process is conducted by utilizing the ACO algorithm. The algorithm was compared with the fuzzy relevance method and other classification methods. The proposed ACO algorithm yielded better results than the other algorithms. Hailong et al. [25] tried to improve ant-based text clustering methods by considering text similarity computation, termination condition, and parametric adoption. The experimental results showed that improved ant-based text clustering provides better performance and has irreplaceable advantages over standard ant-based text clustering methods.

Recently, the literature focuses on the SSO algorithm that is swarm-based. SSO proposed by Cuevas [26] that is utilized in the large area such as image processing, basic physical and mathematical sciences, neural networks, optical flow, web service selection, etc. [27]. However, there is no enough study on disaster management especially on earthquakes, related to tweet clustering by applying the SSO algorithm to understand victims and affected people in a disaster. Several studies have focused on feature selection, feature reduction, data clustering, and text and document clustering through SSO. A system based on SSO for feature selection was proposed that aims to find out the optimal feature set by increasing classification performance [28]. A comparison was done with PSO and genetic algorithm by utilizing different data-sets, and results demonstrate that SSO outperformed PSO and genetic algorithms based on evaluation metrics such as mean, standard criterion, etc. Bas and Ulker [29] proposed a binary SSO to solve the feature selection problem by utilizing eight transfer functions that provide mapping continuous exploration space to binary exploration space. In the feature selection part, K-nearest neighbor and support vector machines were utilized as classifiers. Transfer functions were compared with different evaluation metrics among themselves and the best results were demonstrated. Feature reduction or attribute reduction is also important when considering a huge amount of data. El Aziz and

Hassanien [30] proposed an improved SSO algorithm by utilizing rough sets to solve the attribute reduction problem. Experimental results showed that the proposed attribute reduction algorithm is superior to existing swarm-based considering classification accuracy while confining the number of features. SSO-based algorithms were proposed for data clustering and comparisons have been made that would contribute to the literature [1, 8]. Chandran et al. [31] proposed the SSO algorithm for text clustering to reduce premature convergence and deal with local minima. K-means method was utilized for comparison. SSO algorithm gave better results than the K-means clustering method. In other work proposed by Chandran et al. [32], a hybrid SSO algorithm was implemented for text clustering to improve quality. They utilized two-hybrid clustering methods namely SSO with K-means and K-means with SSO. These methods were compared with different approaches such as K-means, PSO, ACO, and Bee Colony Optimization by considering several evaluation techniques namely average cosine similarity, intra-cluster distance, inter-cluster distance, and accuracy.

There are several studies in the literature by considering disaster events through the SSO algorithm [33–35]. However, to the best of our knowledge, there is no study applied the SSO algorithm to the tweet categorization by considering disaster events to determine center topics in disaster tweets and find physical and psychological effects on victims.

The main contribution of this study is to determine the reaction of victims and other people to the Elazig earthquake during and after the disaster, based on the SSO algorithm by utilizing LSA to represent the relationship between terms. In addition, a different approach for SSO is applied under three modifications. The first modification is to select the best three spiders instead of the worst one for updating operations. The second modification is to take fitness values into consideration instead of the weight of spiders for updating the procedure of the best three spiders. Lastly, newly generated spiders are assigned to the nearest existing spider by calculating Euclidean distance.

3 Background of SSO algorithm

The SSO algorithm is a nature-inspired algorithm that is divided into two categories called social members (Fig. 4) and communal web [26]. The social members are also categorized into two groups as males and females. SSO algorithm focuses on the social members.

The number of female spiders is more than the number of male spiders in the population. The female spiders offer an attraction or dislike to other spiders. Male spiders are categorized into two groups, dominant and non-dominant. Fitness characteristics of dominant males are better than non-dominant males. The offspring is produced within a particular range through mating operation between a dominant male and one or all females [1]. Each spider is demonstrated by a position in each dimension based on weight and vibrations received from the other spiders. Spider position represents candidate solutions within the search space. To transmit information that is encoded as tiny vibrations among the social members, the communal web is utilized. If the total population is formed of npop spiders, the number of females N_f is randomly chosen within the range of 65%–90% of npop. The rest of the other spiders are considered male spiders N_m. N_f and N_m are calculated utilizing the following equations, respectively.

$$Nf = floor[(0.9 - random(0,1) * 0.25 = *npop] \tag{1}$$

$$Nm = npop - Nf \tag{2}$$

Each spider position is randomly chosen according to the upper (p_{high}) and lower (p_{low}) bounds of each dimension of objective function as seen in Eq. 3.

$$s_{i,j} = plow_j + random(0,1) * (phigh_j - plow_j) \tag{3}$$

The greatest distance between two spiders can be calculated utilizing the following:

$$d_{max} = \sqrt{\sum_{j=1}^{n} (phigh_j - plow_j)^2} \tag{4}$$

Distance between two spiders can be calculated according to Eq. 5.

$$d_{i,j} = \sqrt{\sum_{j=1}^{n} (s_{i,j} - s_{j,i})^2 / d_{\max}} \tag{5}$$

The solution quality is represented with the weight w_i of each spider s_i. The following formula is used to calculate the weight of every spider.

$$w_i = (f(s_i) - worst_s)/(best_s - worst_s) \tag{6}$$

where $f(s_i)$ is the fitness value, $worst_s$ and $best_s$ are defined as minimum weight and maximum weight, respectively (considering a minimization problem). The vibrations via information transmitted perceived by the spider s_i from spider s_j can be calculated utilizing Eq. 7.

$$vib_{i,j} = w_j * e^{-d_{i,j}^{\,2}} \tag{7}$$

where w_j is the weight of spider s_j, $d_{i,j}$ is the Euclidean distance between the spiders.

The female spiders offer an attraction or dislike to other spiders. There are several random phenomena for the final movement of attraction or dislike. In this process, a uniform random number between [0,1] is produced. If this

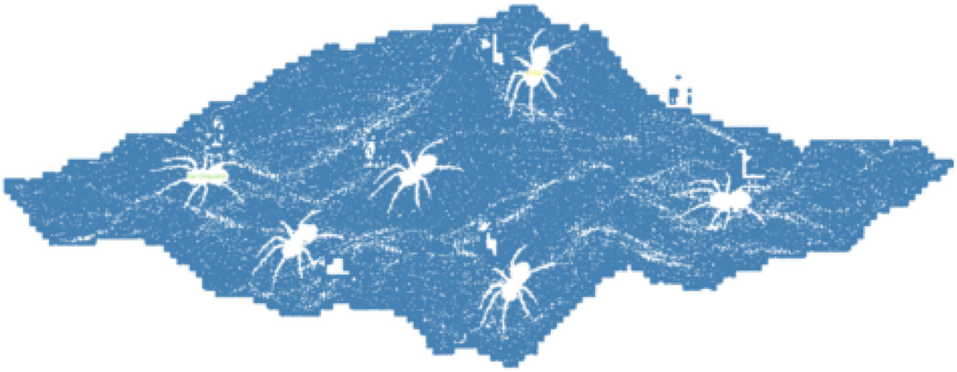


Fig. 4 Social members of SSO algorithm

number is smaller than threshold value PT, an attraction movement is produced; otherwise, a dislike movement is produced. These processes can be modeled using Eq. 8 and 9, respectively.

$$f^{next}_{i,j} = f^{current}_{i,j} + \propto *vibc_i * \left(s_{c,j} - f^{current}_{i,j}\right) + \beta * vibb_i * \left(s_{b,j} - f^{current}_{i,j}\right) + \gamma * (random(0,1) - 0.5 \quad (8)$$

$$f^{next}_{i,j} = f^{current}_{i,j} - \propto *vibc_i * \left(s_{c,j} - f^{current}_{i,j}\right) - \beta * vibb_i * \left(s_{b,j} - f^{current}_{i,j}\right) + \gamma * (random(0,1) - 0.5) \quad (9)$$

where $vibc_i$, $vibb_i$ are vibrations perceived from the nearest better spider ($s_{c,j}$) with a higher weight and from the globally best spider ($s_{b,j}$), $\propto$, β,γ are random numbers generated between 0 and 1. $f^{next}_{i,j}$ is the next position of female spider $\mathrm{s_i}$ in the jth dimension and $f^{current}_{i,j}$ is current position of female spider.

The weight of dominant male spider is above the median weight of the male population. The weights of the non-dominant male spiders are under the median of the male population. The next position of dominant and non-dominant males in the jth dimension can be calculated utilizing Eqs. 10 and 11, respectively.

$$m^{next}_{i,j} = m^{current}_{i,j} + \propto *vibf_i * \left(f_{c,j} - m^{current}_{i,j}\right) + \gamma * (random - 0.5) \quad (10)$$

$$m^{next}_{i,j} = m^{current}_{i,j} + \alpha * \left(W - m^{current}_{i,j}\right) \quad (11)$$

where $vibf_i$ is vibration perceived from the nearest female spider ($f_{c,j}$), $m^{current}_{i,j}$ is the current male spider, $\propto$ and γ are random numbers generated between 0 and 1, W is the weighted mean calculated utilizing Eq. 12.

$$W = \frac{\sum_{h=1}^{N_m} m_h * w_{Nf+h}}{\sum_{h=1}^{N_m} w_{Nf+h}} \quad (12)$$

Before the mating process, a specific range is defined. This range is calculated utilizing Eq. 13.

$$range = \frac{\sum_{j=1}^{n} (phigh_j - plow_j)}{2 * n} \quad (13)$$

where n is the dimension number in the objective function. After the mating operation, a new population is generated. Assume that $\mathrm{s_{dominant}}$ is a dominant male spider, F is the all-female spiders within the specific mating range and E is the set of all spiders that are getting involved in the mating process. E can be calculated utilizing Eq. 14.

$$E = s_{dominant} \cup F \quad (14)$$

To generate new spider, roulette wheel method is utilized as seen in Eq. 15.

$$P_i = \frac{w_i}{\sum w_i} \quad (15)$$

where P_i is the influence probability of each member. If the spiders have a heavier weight, they are more likely to influence generated new spiders.

In the literature, several swarm algorithms such as PSO, ACO, and Artificial Bee Colony (ABC) are the most popular to solve complex optimization problems. However, they present important shortcomings such as not getting the right balance between exploration and exploitation, and trouble to overcome local minima [36]. SSO algorithm provides efficient exploration and broad exploitation through female and male spiders. Therefore, it can find the global optimal solution by achieving a balance between exploration and exploitation [1].

4 Methodology

This section presents the applied steps through the application of the LSA feature extraction technique by utilizing the SSO algorithm for the determination of center tweets among the Elazig earthquake Twitter data-set. The

Table 1 Data utilized in the proposed study

Hashtag	Labels	The total number of tweet	The total number of sentence	The number of cleaned sentence	Sample size (sentence)
Hashtag-1	Victim name	4965	13,663	11,149	1976
Hashtag-2	Earthquake province	6999	17,040	14,136	2053
Hashtag-3	State is together with nation	6871	14,641	12,215	2007
Hashtag-4	Earthquake province	4737	10,196	8747	1884
Hashtag-5	Earthquake province	7179	15,426	13,377	2036
Hashtag-6	Earthquake province	7161	13,891	11,424	1985
Hashtag-7	Mayor of a metropolitan municipality	3639	6539	6539	1757
Hashtag-8	Earthquake observatory	1004	1956	1679	989
Hashtag-9	Earthquake province	6946	15,143	12,879	2024
Hashtag-10	Earthquake province	7566	11,263	9413	1914
Hashtag-11	Syrian rescuer	202	520	419	419
Hashtag-12	Turkish red crescent society	6525	14,154	10,684	1961
Total		63,794	134,432	112,661	21,005

proposed tweet clustering structure includes sequential steps and some techniques as follows:

4.1 Data-set preparation

For the proposed study experiments, Iris and Wine datasets from UCI Machine Learning Repository[5] are utilized in order to test the proposed SSO algorithm. Approximately, a total of 63,794 tweets are collected on January 25 and 26, 2020. However, in the proposed study, the total number of sentences, which equals 134,432, is analyzed, in order to handle different subjects in the same tweet. The number of 112.661 sentences among collected tweets is pre-processed and vectorized by utilizing the Latent Semantic Analyses (LSA) information retrieval method. The number of 21,005 sentences gathered by sampling is utilized under twelve different hashtags in the proposed study. The pre-processing stage that is the operation of cleaning data and making data ready for the other operations is applied to 112,661 sentences by utilizing Python 3.6 software as follows:

- Date and time fields in the tweets are removed.
- URLs (Uniform Resource Locator) referred to as web addresses are removed because of causing misclassification of texts.
- Usernames that do not have any importance in making the classification are removed.
- Punctuation does not make sense alone. Therefore, punctuations are removed.
- Infrequent words in the text are removed.
- Tokenization that is the operation of splitting the strings into pieces is applied.

At the end of the pre-processing stage, the LSA feature extraction technique is implemented on sentences. Sampling is applied in order to obtain the portray of the population [37] by taking a confidence interval into consideration due to the huge amount of tweets and gathered representative samples. The confidence level indicates the interval, in our case 95%, leads the percentage of all such possible intervals which comprehends the true value of the parameter [38] with the 2% margin of error. In Table 1, utilized hashtags with their labels and sample size are demonstrated.

The modified SSO algorithm is run separately for each of the hashtags on the tweets in Turkish. The number of 1976, 2053, 2007, 1184, 2036, 1985, 1757, 989, 2024, 1914, 419, 1961 sentences under Hashtag-1, Hashtag-2, Hashtag-3, Hashtag-4, Hashtag-5, Hashtag-6, Hashtag-7, Hashtag-8, Hashtag-9, Hashtag-10, Hashtag-11, Hashtag-12 are considered, respectively, according to the sample size. In the modified SSO algorithm, a total of 21,005 sentences are utilized.

4.2 Latent semantic analysis (LSA)

LSA is associated with the semantic structure of documents to improve the determination of relevant documents considering terms found in queries [13]. The relationship between the documents can be found by utilizing LSA. There are some steps involved in LSA; decompose term-document matrix utilizing SVD, reducing the space, computing reduced document representations, computing similarity with all reduced documents, and output ranked list

[5] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.php.

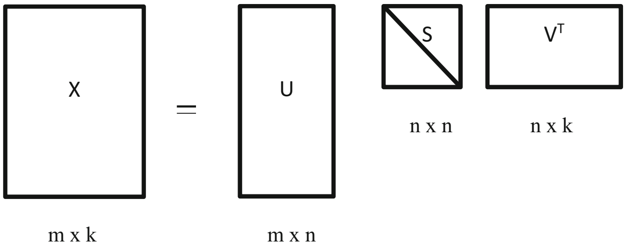


Fig. 5 The SVD representation to reduce dimension

of documents [39]. LSA does not utilize dictionaries constructed by humans. It uses input only raw text separated into words, parts, or samples such as paragraphs and sentences. First of all, the text is symbolized as a matrix where each row represents a unique word and each column represents a document. Each cell includes the frequency with which the word of its row seems in the document indicated by its column. After, the cell entries are exposed to a beginning transformation [40]. LSA uses SVD for dimension reduction. It is an information retrieval method founded on a spectral examination of the term-document matrix [41]. The words have similar meanings that take place in similar parts of the text. Let the corpus is indicated as a term-by-document matrix X (m × k) namely m different words in k documents collection. In Fig. 5, schematic SVD to reduce the dimension of the term-by-document matrix can be seen.

The SVD is given by Eq. 16.

$$X = USV^T \tag{16}$$

where U: $m \times n$ orthogonal matrix, S: $n \times n$ diagonal matrix, V: $n \times k$ orthogonal matrix.

The initial term by document matrix is approximated by utilizing the n largest singular values and their related singular vectors as seen in Eq. 17 [13].

$$X_n = U_n S_n V_n^T \tag{17}$$

where U_n is represented as the first n columns of the matrix U and V_n^T is comprised of the first n rows of matrix V^{T}. S_n is the first n factors [6].

LSA has been mainly utilized as an information retrieval method where a comparison can be made between the query vector and reconstituted matrix or query vector can be mapped to the reduced space [13]. Count data are utilized by LSA which is a statistical data type and observations can receive only nonnegative values.

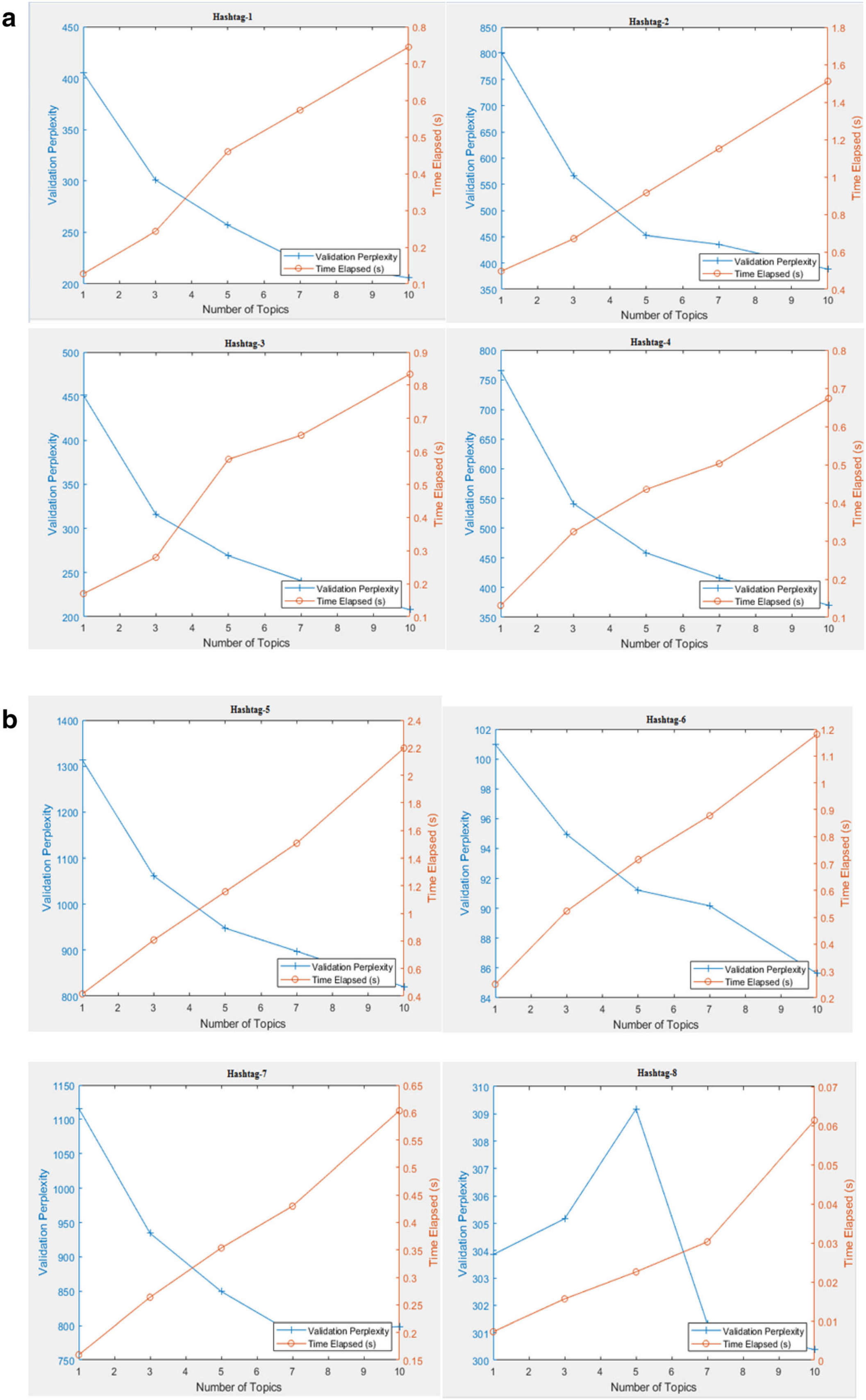
The various applications of LSA result from its ability to reveal the hidden topic structure in the input matrix, and creating a semantic space. Therefore, the comparison of parts of the text in various sizes can be allowed by the subsequent presentation of words and documents over this semantic space [42]. LSA can investigate linguistic properties as synonymy and polysemy of words [21].

Fig. 6 **a** Hashtag-1, Hashtag-2, Hashtag-3, and Hashtag-4 for the number of topics, **b** Hashtag-5, Hashtag-6, Hashtag-7, and Hashtag-8 for the number of topics, **c** Hashtag-9, Hashtag-10, Hashtag-11, and Hashtag-12 for the number of topics ▶

In the proposed study, the LSA process is implemented using MATLAB R2018a software.[6] It is applied to the preprocessed text. The number of components, which specified as a positive integer, namely data-object (features) is considered as 4 by trial and error. Data-object stands for the dimensionality of the result vectors. Word embedding methods such as LSA are essentially based on bag-of-words models. Bag-of-words model in which the statement of each word reflects a weighted bag of context-words that co-occur with it [43].The proposed LSA model utilizes a bag-of-n-grams model in which text is not split into words. n-gram indicates a collection of n consecutive words. Euclidean distance is utilized to calculate the distance between the document scores vectors. By considering validation perplexity and time elapsed as performance metrics for topic numbers, the number of topics is estimated. Perplexity is one of the popular evaluation metrics for language models [44]. In the proposed study, the logarithmic probabilities function is utilized to calculate the validation perplexity. In the following Figures (Fig. 6a–c), the relationship between validation perplexity and time elapsed (second) to represent the number of topics based on all hashtags can be seen.

Based on an analysis of variation of validation perplexity and time elapsed during topic modeling, the most appropriate number of topics is estimated. Figure 6 indicates validation perplexity and time elapsed versus the number of topics. In the text analyses, better fit comes up with a lower validation perplexity. With lower perplexity in a language model, lesser confusion in recognition and higher speech-recognition accuracy is achieved [44]. Validation perplexity trends down when the number of topics is 3 in the proposed graphs except for Hashtag-8 and 11. Choosing a higher number of topics can provide a better fit; however, convergence takes a long time. In the points where validation perplexity is less, the elapsed time increases. Besides the graphics in Fig. 6, it is considered that the increase in the number of topics can cause deviation from the points that should be focused on. By taking the above graphics and literature [45] into consideration, the number of clusters for the modified SSO algorithm is taken as 3 based on all hashtags.

[6] MathWorks, http://www.mathworks.com/help/textanalytics/ref/fitlsa.html


a
Hashtag-1
Hashtag-2
Hashtag-3
Hashtag-4
b
Hashtag-5
Hashtag-6
Hashtag-7
Hashtag-8
Validation Perplexity
Time Elapsed (s)
Number of Topics
1 2 3 4 5 6 7 8 9 10
Validation Perplexity
Time Elapsed (s)

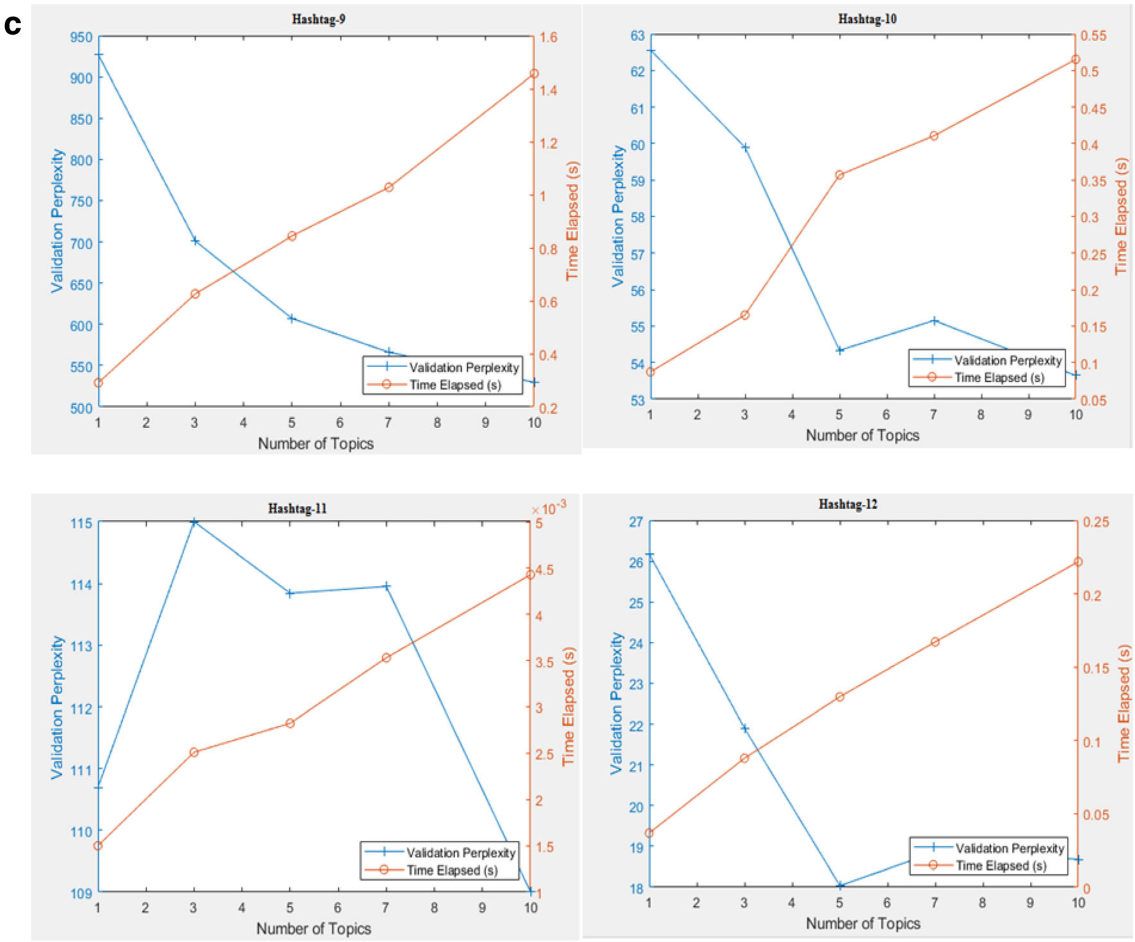


Fig. 6 continued

4.3 The proposed SSO algorithm

In our application, the fitness function f of spider s_i is calculated utilizing Eq. 18 [1]. Our problem is the minimization problem; therefore, the aim of proposed algorithm is to minimize fitness function.

$$f(s_i) = \frac{\sum_{i=1}^{K} \frac{\sum_{j=1}^{n_i} distance(center_i - document_j)}{n_i}}{K} = f(s_i = \{C_1, C_2, C_3, \ldots C_K\}) \quad (18)$$

where K is the number of clusters taken as 3. n_i is the number of data-objects in cluster C_i, $center_i$ is the center of cluster C_i, $document_j$ is the jth data-object in included in clusterC_i. Distance is the distance between two data-object vectors utilized as Euclidean distance in our study. The flowchart of the proposed SSO algorithm can be seen in Fig. 7.

In the proposed SSO algorithm, a new spider is generated by utilizing the weight and vibrations of each spider. Among the dominant males, one male that has maximum weight namely with minimum fitness value is selected and mating operation is applied between this dominant male and females within the specific range. The working principle of the proposed algorithm is based on updating the three best spiders in the population. First of all, the three best spiders are chosen according to the minimum fitness values and the sum of intra-cluster distances (SICD) is calculated (Eq. 19). After that, new spider fitness is calculated in order to assign the new spider to the nearest point in the distance matrix as we need exact points to obtain center tweets. Among the best three spiders, the first spider is removed and the new spider is appended. If the SICD of the combination of the new three spiders is better than the previous SICD, the new combination is kept, otherwise, the process continues till the maximum iteration is met by updating with generated better combinations. In the original SSO, weights are taken for updating process by eliminating the worst spider. In the proposed SSO, fitness values are considered for the updating process and the updated process is implemented via the best three spiders. The reason behind the aforementioned approach is that the original SSO focuses on middle points between two exact

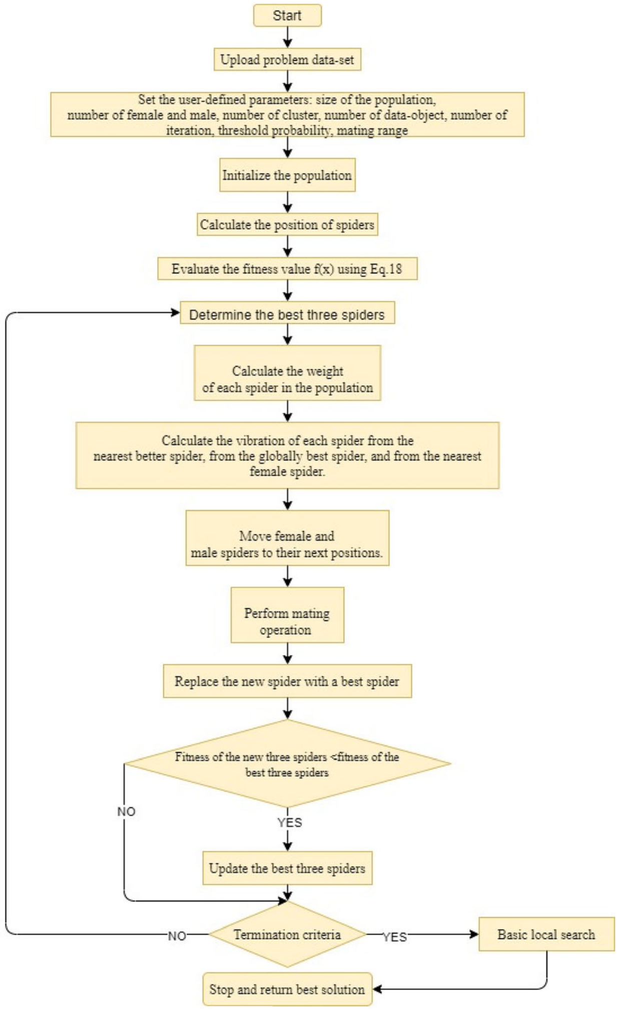

Start
Upload problem data-set
Set the user-defined parameters: size of the population, number of female and male, number of cluster, number of data-object, number of iteration, threshold probability, mating range
Initialize the population
Calculate the position of spiders
Evaluate the fitness value f(x) using Eq.18
Determine the best three spiders
Calculate the weight of each spider in the population
Calculate the vibration of each spider from the nearest better spider, from the globally best spider, and from the nearest female spider.
Move female and male spiders to their next positions.
Perform mating operation
Replace the new spider with a best spider
Fitness of the new three spiders <fitness of the best three spiders
NO
YES
Update the best three spiders
NO
Termination criteria
YES
Basic local search
Stop and return best solution

Fig. 7 Flowchart of the SSO algorithm

points, while the proposed SSO focuses on the exact points. In addition to the given reasons, it can be said that weight calculation adaptation for exact points is difficult to implement and weight calculation leads to complexity in our case. Therefore, the proposed study focuses on the fitness value of spiders and takes the best three spiders for the update process. Moreover, the exact solution is conducted via GUROBI, and a comparison is done with the proposed SSO. After the maximum number of iteration, a basic local search is applied by removing one spider among the best spiders and appending a random spider in order to get closer to the optimum every time and ensure stability. If the new generated combination is better than the previous one, the new combination is kept.

$$\mathrm{SICD} = \sum_{i=1}^{K} \sum_{j=1}^{n} \mathrm{distance}\left(\mathrm{center}_i, \mathrm{data\_object}_j\right) \tag{19}$$

During the application of the SSO algorithm, new data-objects (features) are generated. In our case, it is required to find the central tweets among the others by utilizing existing data-objects. Therefore, new data-objects and their distances to existing nodes are calculated by Euclidean distance and new data-objects attended to their nearest node.

The pseudo-code of the SSO is presented as follows:

4.4 Experimental design

Iris and Wine data-sets from UCI Machine Learning Repository are used for experimenting on the proposed SSO algorithm since the performance of the SSO algorithm in data clustering has been tested on these data-sets in the literature [1, 8] and these data-sets are analyzed with 3 clusters as it is in our data-set. All processes are applied by the R2018a MATLAB software package. Iris and wine data-sets are run five times for the number of iterations equal to 100, 200, and 300, respectively, when the population is 50. Features of data-sets results according to the SICD including optimal clusters and data per cluster, and elapsed time of these data-sets are given in Tables 2 and 3, and 4, respectively. All experiments are run on a computer with Intel ® Core ™ i7-4790 CPU with a 3.60 GHz processor, 8 GB of RAM, and the Windows 8 operating system.

The Euclidean distance function is utilized in the proposed SSO algorithm; better results are produced when the population number is 50 and 100 iterations in both Iris and Wine data-sets. Optimum SICD value is found as 98.2136 and 16,375.88 for Iris and Wine data-sets, in accordance. The amount of time MATLAB spends to terminate processes is referred to as elapsed time that represents the time in seconds. An optimal solution can be found less than 1 s for the Iris data-set and less than 4 s for the Wine data-set.

4.5 Exact solution

SICD formulation is used in order to find the exact solution. The formulation is applied to the Iris and Wine data-sets by utilizing GUROBI 9.0 package program. The run is taken on a computer with a 3.60 GHz Intel Core processor and 8 GB of RAM. The optimal solutions are gathered as 98.2136 and 16,375.88, respectively. The optimal

Table 2 Features of Iris and Wine data-sets

	Iris	Wine
Number of data	150	178
Number of classes	3	3
Number of data-objects	4	13

Table 3 SICD variations based on the solutions

	Iris data-set			Wine data-set		
	100 Iterations	200 Iterations	300 Iterations	100 Iterations	200 Iterations	300 Iterations
Best	98.2136	98.2136	98.2136	16,375.88	16,375.88	16,375.88
Average	98.7963	99.0368	98.8008	16,395.14	16,400.17	16,396.52
Worst	99.4777	99.7040	99.8854	16,417.45	16,426.68	16,454.84
Optimal clusters		[8 79 113]			[51 73 136]	
Data per cluster for optimal solution		[50 62 38]			[48 68 62]	

Table 4 Elapsed time variations based on the number of iterations

	Iris data-set			Wine data-set		
	100 Iterations	200 Iterations	300 Iterations	100 Iterations	200 Iterations	300 Iterations
Best	0.6034	0.6965	0.7990	2.9187	3.4078	3.9521
Average	0.6471	0.7581	0.9422	2.9680	3.4930	4.0345
Worst	0.6780	0.8399	1.0588	2.9836	3.5746	4.2031

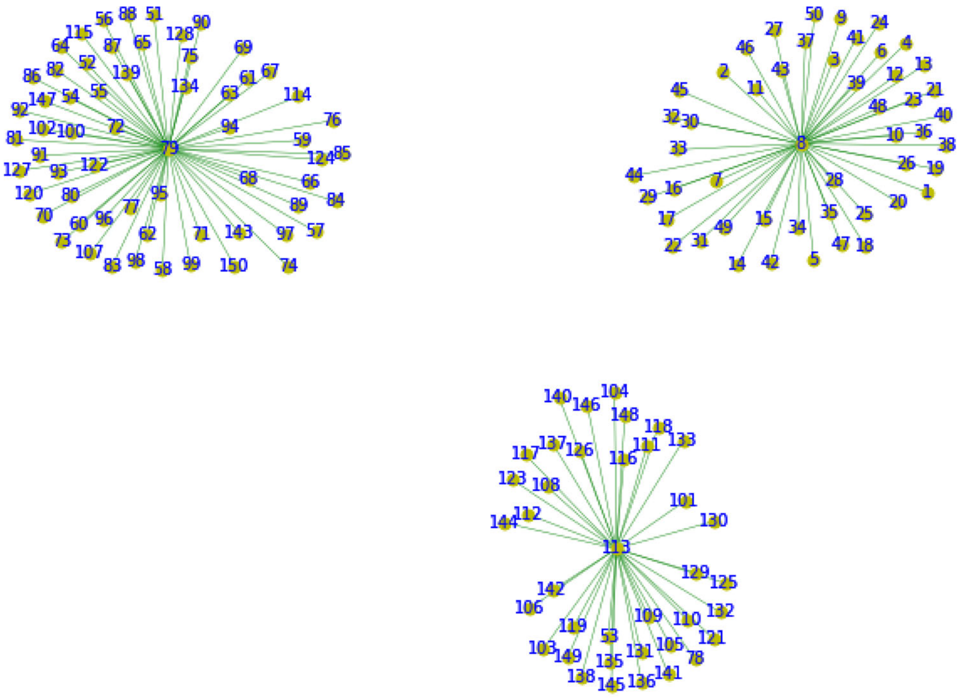


Fig. 8 The optimal assignments for iris data-set

assignments for data-sets by considering center nodes are shown in Figs. 8 and 9, in accordance.

By considering the exact solution, it is seen that the proposed SSO algorithm gives optimal results.

5 Results and discussions (application of proposed method on real data-set)

The proposed algorithm by using the Elazig earthquake data-set is run ten times with the population size 50 and 500 iterations for each hashtag. Namely, the proposed algorithm is run 120 times for all hashtags in total. In Table 5, SICD values and elapsed times by considering

Fig. 9 The optimal assignments for wine data-set

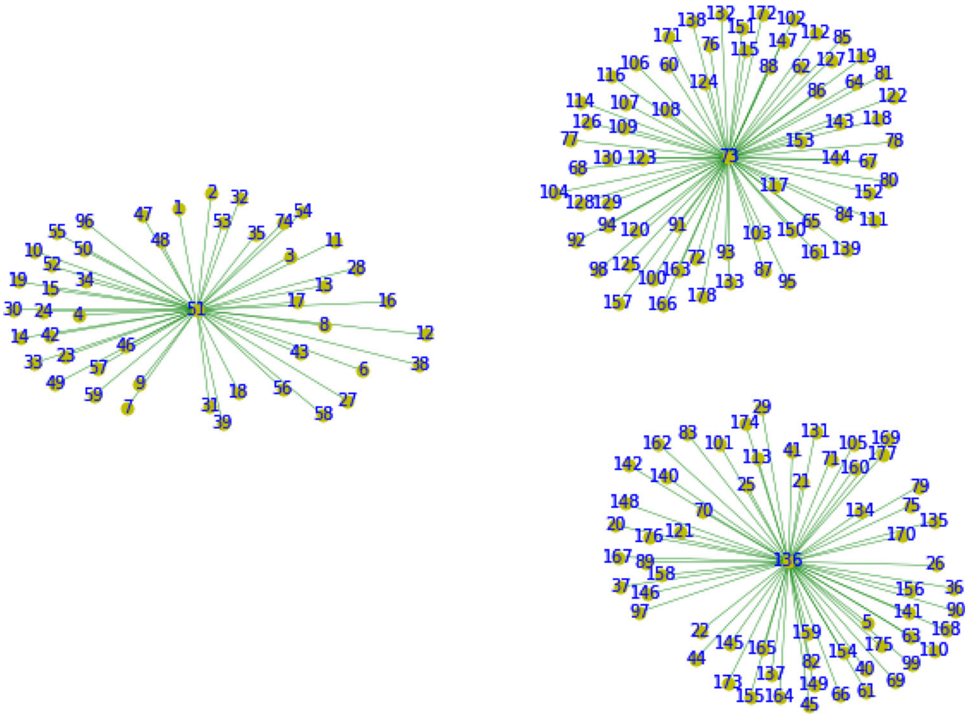


Table 5 Analyses results for proposed SSO algorithm

	SICD			Elapsed Time			Best Solution Centers
	Best	Average	Worst	Best	Average	Worst	
Hashtag-1	47.26	47.30	47.35	111,61	114.67	116.95	[11253 4417 3287]
Hashtag-2	38.83	38.96	39.11	119.01	121.16	124.24	[13348 1525 67818]
Hashtag-3	49.27	49.53	49.99	117.13	119.11	120.94	[12038 5424 5025]
Hashtag-4	43.12	43.34	44.22	103.61	105.54	107.68	[7622 9245 5447]
Hashtag-5	40.44	40.70	40.89	120.31	122.18	123.61	[401 13,317 12363]
Hashtag-6	64.93	65.89	67.28	122.86	125.61	127.42	[1092 1404 5087]
Hashtag-7	44.02	44.18	44.56	96.56	98.10	100.13	[5748 4227 5266]
Hashtag-8	58.40	58.93	59.41	28.19	29.55	31.22	[1250 1782 1099]
Hashtag-9	51.52	51.85	52.17	118.31	120.00	122.48	[6199 10,860 1590]
Hashtag-10	75.69	77.014	78.82	106.07	109.45	113.71	[10993 1030 6611]
Hashtag-11	45.00	45.078	45.086	4.34	5.81	11.03	[10 7 201]
Hashtag-12	49.27	50.12	50.62	109.95	112.46	119.45	[2600 3578 1506]

best, average, and worst values, as well as centers of optimal solutions for each hashtag, is demonstrated.

From Table 5, it can be said that there is no considerable deviation between best, worst, and average values for each hashtag. The biggest deviation between average value and worst value is found in Hashtag-10, and the smallest value is found in Hashtagh-11. Also, when the elapsed time is taken into consideration, it can be easily said that it is consistent for each run. Therefore, from Table 5, it can be concluded that the proposed SSO algorithm works consistently according to best, average, and worst SICDs and elapsed times. In Table 6, English translations of centers for each hashtag are demonstrated. However, tweets in Turkish are used for analyses.

Elazig earthquake data are collected under twelve hashtags. One of these hashtags is about a victim who was under the debris labeled as Hashtag-1. Healthcare personnel made a phone call with a survivor under the debris and managed the rescue operation, which is why Hashtag-1 was brought to the forefront. Hereby, the victim and her relatives survived the debris with conversation. Another hashtag (Hashtag-11) is related to a person who scraped the debris with his nails and rescued the people. Hashtag-9 and Hashtag-10 are districts in which are affected by the

Table 6 Center tweets for each hashtag (Proposed SSO algorithm)

Hashtag-1	Center 1	The center tweet which expresses conversation between health personnel and victim during rescue operation
	Center 2	The center tweet which indicates the saving victim alive from debris after considerable time
	Center 3	The center tweet which says there is no discrimination between ethnic groups in Turkey and underlines brotherhood, unity, togetherness and humanity
Hashtag-2	Center 1	The center tweet which criticizes the perception by not considering human life but occurred responses
	Center 2	The center tweet which expresses the help of an immigrant during the rescue operations
	Center 3	The center tweet which includes a pray for protection of institutions which helps the recovering processes
Hashtag-3	Center 1	The center tweet informs that President of the Republic of Turkey is participated search and rescue operations in Elazig
	Center 2	The center tweet which includes full support of State after earthquake
	Center 3	The center tweet which says sharing mourning and providing help for whom in need
Hashtag-4	Center 1	The center tweet which claims there is an intention for gathering votes after the earthquake
	Center 2	The tweet which demonstrates the bad weather condition in Elazig
	Center 3	The center tweet which includes the message for relatives of victims who lost their lives and for victims who injured in Elazig
Hashtag-5	Center 1	The center tweet which includes offering accommodation for victims in Elazig
	Center 2	The center tweet which make criticism on people who criticize the government during the operations for recovery after the earthquake
	Center 3	The center tweet which expresses the earthquake in Elazig
Hashtag-6	Center 1	The center tweet which indicates the help and motivation of national football team for Elazig earthquake victims
	Center 2	The center tweet which says Elazig consists of several ethnic groups, but brave man comes out from there, there will be more, if and only if, people get rid of the bad thoughts
	Center 3	The center tweet which consists of information about the help of one of the associations in Turkey for Elazig
Hashtag-7	Center 1	The center tweet which makes criticism on a mayor of one of the metropolitan municipalities in Turkey for his/her objectiveness
	Center 2	The center tweet which makes criticism on a mayor of one of the metropolitan municipalities in Turkey for his/her lack of information about holly scripts
	Center 3	The center tweet which claims that a mayor of one of the metropolitan municipalities in Turkey is the most reliable politician
Hashtag-8	Center 1	The center tweet which includes the report of earthquake observatory about magnitude of earthquake in Elazig province
	Center 2	The center tweet which includes criticism on journalists and politics
	Center 3	The center tweet which includes announcement of AFAD (Disaster and Emergency Management Presidency) and earthquake observatory about magnitude of earthquake in Elazig
Hashtag-9	Center 1	The center tweet which says memory of the first time hearing of Elazig earthquake and its effect on person
	Center 2	The center tweet which expresses the announcement of Presidency of Religious Affairs which is about guiding people to shelters and meeting points and works done by its personnel
	Center 3	The center tweet which includes the help of people from different ethnicity after the earthquake in Elazig
Hashtag-10	Center 1	The center tweet which says local reporters made interview with citizens in the earthquake region
	Center 2	The center tweet which says whole country shall see
	Center 3	The center tweet which says "17 s that tears the hearths!"
Hashtag-11	Center 1	The center tweet which expresses the effort of immigrant during the rescue of victim
	Center 2	The center tweet which includes the speaking of victim after rescued by immigrant
	Center 3	The center tweet which expresses the effort of immigrant during the rescue of victim and treatment against the immigrants in general
Hashtag-12	Center 1	The center tweet which reports that blood needs in the region were met by blood centers
	Center 2	The center tweet which says help for the earthquake region, mobile kitchen with the capacity of 5000 people, coming from Erzurum city
	Center 3	The center tweet which comprehends response for criticism on Turkish Red Crescent Society

earthquake. The other hashtag is Hashtag-8 (Earthquake Observatory) where earthquake data are published. Hashtag-7 is about a mayor of one of the metropolitan municipalities in Turkey. In addition, Hashtag-3 regards the state's support to the nation. Finally, tweets under the Hashtag-12 are collected. There are also some other hashtags related to Elazig.

When it is considered the highlights in the center tweets, several points can be expressed. First of all, it was talked about brotherhood, unity, and togetherness through rescuing the victim (Hashtag-1, Hashtag-11) and her relatives at the end of 17 h by a phone conversation between the healthcare personnel and victim. Also, a person who scraped the debris with his nails and rescued the people is mentioned with the aforementioned perspective. In addition, it was discussed that Turkey is in unity with help of the non-governmental organizations. State supported to the nation, and the Presidency of Religious Affairs opened the mosques to earthquake victims at night. Some citizens made announcements about their empty houses; they want to allocate them to earthquake victims. The society became one and help was collected through the Turkish Red Crescent. In a short time, 9 trucks were ready and set off for Elazig, 10 tons of potatoes, 5000 boxes of water, 7000 blankets were distributed. In addition, announcements were made about the need for blood. A mobile kitchen with a capacity of 5 thousand people was sent from Erzurum city to serve in the region where the earthquake hit. From time to time, politicians were discussed. It was spoken that the head of the state immediately went to the region.

5.1 Comparison to K-means clustering

K-means clustering technique belongs to a group based on the iterative reposition of data points between clusters called partitioning-based techniques. It is utilized to split either the events or the variables of a data-set into non-overlapping clusters [46]. The aim of the algorithm is to generate clusters with a high degree of similarity within each cluster and a low degree of similarity between clusters. K-means clustering has some strong and weak aspects. The strength of the algorithm is: easy to implement, computationally faster than most of the clustering methods with large populations, the adaptation process is easy for new examples. The weakness of the algorithm is: performance of the K-means clustering result is affected negatively by noise, outliers, and empty clusters. Clustering result is also sensitive to the initial points [47]. Moreover, global optimization is not considered by K-means to produce the optimal number of clusters [48]. In order to overcome the aforementioned weaknesses, the SSO algorithm is applied and compared with K-means clustering.

Experiments are conducted to compare the performance of the SSO algorithm with the K-means clustering technique, which is the most popular clustering method, to show the differentiation of center tweets. Tweets are clustered after the LSA processing. All processes are applied by the R2018a MATLAB software package as in the proposed SSO algorithm. SICD is utilized as a metric. K-means clustering technique with the Elazig earthquake data-set is run ten times with the 500 iterations for each hashtag. SICD values and elapsed times are taken into consideration with best, average, and worst values, and centers of optimal solutions can be seen in Table 7.

When Table 7 is compared with Table 5, it is seen that the K-means clustering technique gives faster but poor results. As the number of iterations increases, the results do not improve. The results are not stable; the reason for this can be getting stuck into local optima. The Hashtag-12 has given the same SICD value in both algorithms (Proposed SSO and K-means). In Table 8, centers for each hashtag are demonstrated as a result of the K-means clustering technique.

In total, 12 of the 36 center tweets are matched in two applications. These centers are 'Center 3' in Hashtag-2, 'Center 2,3' in Hashtag-3, 'Center 1' in Hashtag-4, 'Center 1' in Hashtag-6, 'Center 1,3' in Hashtag-10, 'Center 1,2' in Hashtag-11, 'Center 1,2,3' in Hashtag-12.

We compare center tweets of the proposed SSO algorithm with center tweets of the K-means clustering technique. To begin with, instead of dialogues between victims and rescuers and an emphasis on race as seen in the proposed SSO algorithm, the Hashtag-1 of K-means clustering contains only the victim's origin and praise for the rescue team. Despite the fact that the Hashtag-2 has identical centers, a center tweet about a mother and her children can be seen in K-means clustering, which is distinct from the proposed SSO. Moreover, there are different center tweets in the Hashtag-3; centers of both algorithms demonstrate the same content. Unlike the proposed SSO algorithm, the Hashtag-4 has a center tweet about animals. The Hashtag-5 emphasizes good people in the K-means algorithm. Although there are similar centers in the Hashtag-6, a center tweet about financial support for traders can be seen in K-means clustering. While Hashtag-7 in the proposed SSO algorithm includes only the mayor of one of the metropolitan municipalities in Turkey, the K-means algorithm includes also information about dead and injuries. The SSO algorithm has both criticism and praise on the mayor, while the K-means has only praise. In Hashtag-8, there are similar centers, but in the SSO, there is criticism against journalists, while the K-means emphasizes the fear of earthquakes. While Hashtag-9 includes help and aid topics in both algorithms, the K-means has a center tweet related to the wish that the earthquake will not happen

Table7 Analyses results for K-means clustering

	SICD			Elapsed time			Best solution centers
	Best	Average	Worst	Best	Average	Worst	
Hashtag-1	56.13	59.01	62.58	16.22	22.88	76.32	[7998 8661 12510]
Hashtag-2	46.39	51.36	62.23	16.52	17.14	17.81	[17029 12,882 1415]
Hashtag-3	56.54	61.78	67.04	15.75	16.42	18.63	[8720 285 10741]
Hashtag-4	50.58	57.46	67.26	15.43	15.84	16.78	[10158 1865 1761]
Hashtag-5	44.52	48.47	51.55	16.42	17.40	18.86	[2304 5200 11600]
Hashtag-6	66.53	69.80	76.03	15.67	16.24	17.63	[952 7046 5552]
Hashtag-7	54.81	61.22	72.51	14.57	15.33	16.29	[1980 5287 383]
Hashtag-8	63.55	66.47	72.32	9.38	9.97	10.59	[1796 625 215]
Hashtag-9	54.40	58.26	62.11	15.84	16.71	17.19	[11750 9195 13000]
Hashtag-10	79.24	84.52	88.38	15.75	16.15	16.50	[1626 1082 396]
Hashtag-11	50.81	53.18	57.82	5.76	6.12	6.56	[477 298 407]
Hashtag-12	49.27	55.52	77.15	15.54	16.71	20.06	[3687, 13,078, 1506]

again. Hashtag-10 emphasizes the same things, except for minor details. Although Hashtag-11 includes the issues of unity and togetherness in both algorithms, the K-means algorithm contains a center tweet about worship under the debris. Hashtag-12 gives the same results in both algorithms.

Although both algorithms cover similar centers and topics, there are many hashtags have irrelevant centers in the K-means clustering technique. The comparison of results shows that the proposed SSO algorithm is a better clustering method compared with the K-mean clustering technique when SICD and content integrity are taken into consideration.

6 Conclusion

In a nutshell, social media accumulates a huge amount of just-in-time data. Therefore, it can be utilized for emergencies and disaster situations. In the proposed study, Twitter data are collected within a three-day earthquake in Elazig, Turkey. Analyses of collected data consist of three main steps which are data-set preparation including preprocessing stage, LSA application, and implementation of the SSO algorithm. During the aforementioned processes, our main focus is to extract center tweets in order to deduce important topics for the current situation of disaster. In the SSO algorithm, new data-objects (features) are generated. In our case, we applied the modified SSO algorithm to solve the clustering problem in order to find the central tweets among others. The first modification is to select the best three spiders instead of the worst one for updating operations. Therefore, intensification in SSO is augmented. The second modification is to take fitness values into consideration instead of the weight of spiders for updating the procedure of the best three spiders. Next, newly generated spiders are assigned to the nearest existing spider by calculating Euclidean distance in order to identify a center tweet. Lastly, local search is applied to ensure avoid local optima. Moreover, we compare the proposed SSO algorithm with the K-means clustering technique. K-means clustering technique gives faster but poor results with SICD.

The results of the proposed SSO algorithm can be considered from four main perspectives which are humanistic, humanitarian, political, and charity. From the humanistic perspective, Hashtag-1, Hashtag-9, and Hashtag-11, which indicate brotherhood, unity, and togetherness during rescue operations, can be given. From the political perspective, it can be seen that some of the criticisms are made with Hashtag-7 and Hashtag-3. Tweets about a mayor of a metropolitan municipality and the President of the Republic of Turkey are seen in these hashtags. In addition, the tweets demonstrate an expression that includes the state's support for the nation. Hashtag-6 can be considered from both humanitarian and humanistic perspectives. Lastly, Hashtag-12 includes the charity perspective. Namely, the society became one, and help was collected through the Turkish Red Crescent and other governmental or non-governmental organizations. Although the K-means clustering includes similar centers with the proposed SSO algorithm, irrelevant center tweets in the hashtags are seen distinctively. In terms of content integrity, the proposed SSO algorithm gives better results.

All in all, the general situation is analyzed and results for different topics are gathered. Therefore, comprehensive insight is received with an analysis of the tweets in order to understand the general situation of society during and after the emergencies.

Table 8 Center tweets for each hashtag (K-means clustering)

Hashtag-1	Center 1	The center tweet indicates the origin of the victim
	Center 2	The center tweet includes thanks to the search and rescue teams for their efforts
	Center 3	The center tweet includes what the rescuer told about the victim
Hashtag-2	Center 1	The center tweets which includes criticism on the head of the Turkish Red Crescent Society
	Center 2	The center tweet which is about a mother who was rescued from debris and described the location of her children
	Center 3	The center tweet which expresses the help of an immigrant during the rescue operations
Hashtag-3	Center 1	The center tweet indicates food and shelter aid in the park
	Center 2	The center tweet informs that President of the Republic of Turkey is participated search and rescue operations in Elazig
	Center 3	The center tweet which includes full support of State after earthquake
Hashtag-4	Center 1	The center tweet which claims there is an intention for gathering votes after the earthquake
	Center 2	The center tweet says the animals that are treated badly by people, are waiting for people on the earthquake day
	Center 3	The center tweet indicates a wish that disasters do not happen again
Hashtag-5	Center 1	The center tweet expresses that 35% of people within an ethnic group are more patriotic than other citizens
	Center 2	The center tweet talks about the existence of good-hearted people
	Center 3	The center tweet says that history will record a father who sacrifices his life as a shield for his children
Hashtag-6	Center 1	The center tweet which indicates the help and motivation of national football team for Elazig earthquake victims
	Center 2	The center tweet emphasizes the wishes to get well soon
	Center 3	The center tweet which is about financial support for trader who is damaged by the earthquake
Hashtag-7	Center 1	The center tweet informs about the number of dead, injured, and missing in the earthquake
	Center 2	The center tweet expresses opposition on comments of group, which is not in favor of mentioned mayor administration, about metropolitan municipalities major’s visit
	Center 3	The center tweet which claims that a mayor of one of the metropolitan municipalities in Turkey should be the President of the Republic of Turkey
Hashtag-8	Center 1	The center tweet includes the possibility of earthquakes of this magnitude recurring in the region
	Center 2	The center tweet emphasizes the fear of an earthquake at any moment
	Center 3	The center tweet refers to the 5.1 magnitude earthquake that occurred in the region
Hashtag-9	Center 1	The center tweet which expresses the announcement about the donation to the disaster area
	Center 2	The center tweet which includes expression that the aid campaign is ongoing
	Center 3	The center tweet indicates a wish that disasters do not happen again
Hashtag-10	Center 1	The center tweet which says local reporters made interview with citizens in the earthquake region
	Center 2	The center tweet about wishes for getting well soon
	Center 3	The center tweet which says whole country shall see
Hashtag-11	Center 1	The center tweet which expresses the effort of immigrant during the rescue of victim
	Center 2	The center tweet which expresses the effort of immigrant during the rescue of victim and treatment against the immigrants in general
	Center 3	The center tweet includes worship under the debris
Hashtag-12	Center 1	The center tweet which reports that blood needs in the region were met by blood centers
	Center 2	The center tweet which says help for the earthquake region, mobile kitchen with the capacity of 5000 people, coming from Erzurum city
	Center 3	The center tweet which comprehends response for criticism on Turkish Red Crescent Society

For future work, the system that gathers tweets simultaneously finds center tweets and gives an opportunity to evaluate the current situation in the disaster will be provided via generating an application.

In addition, multiple initial populations can be integrated into SSO. Therefore, a more comprehensive search in the solution space can be provided, and skipping good solutions can be avoided. SSO will be utilized for different social networks such as Facebook and other openly available sources in order to have a more detailed insight into emergency situations. Also, SSO can be used to handle irrelevant data by separating them into a specific cluster. Therefore, more reliable data can be gathered.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Thalamala RC, Venkata Swamy Reddy A, Janet B (2020) A novel bio-inspired algorithm based on social spiders for improving performance and efficiency of data clustering. J Intell Syst 29(1):311–326
2. Thalamala R, Barnabas J, Reddy AV (2019) A novel variant of social spider optimization using single centroid representation and enhanced mating for data clustering. PeerJ Comput Sci 5:201
3. Bharti KK, Singh PK (2016) Chaotic gradient artificial bee colony for text clustering. Soft Comput 20(3):1113–1126
4. Abualigah LM, Khader AT, Hanandeh ES (2018) A new feature selection method to improve the document clustering using particle swarm optimization algorithm. J Comput Sci 25:456–466
5. Liu X, Fu H (2010) An effective clustering algorithm with ant colony. J Comput 5(4):598–605
6. Song W, Park SC (2009) Genetic algorithm for text clustering based on latent semantic indexing. Comput Math with Appl 57(11–12):1901–1907
7. Hong SS, Lee W, Han MM (2015) The feature selection method based on genetic algorithm for efficient of text clustering and text classification. Int J Adv Soft Comput its Appl 7(1):22–40
8. TR Chandran, AV Reddy, and B Janet (2019) Performance comparison of social spider optimization for data clustering with other clustering methods. In: Proceedings 2nd International Conference Intelligent Computer Control Systems ICICCS 2018, no. Iciccs, pp 1119–1125
9. A Aghamohseni and R Ramezanian (2015) An efficient hybrid approach based on K-means and generalized fashion algorithms for cluster analysis. In: 2015 AI Robot. IRANOPEN 2015 - 5th Conference Artificial Intelligence Robotics, pp 1–7
10. Nandwalkar JR, Pete DJ (2021) Social spider optimization based optimized heat management for wet-electrospun polymer fiber. Microw Opt Technol Lett 63(2):670–678
11. Yu JJQ, Li VOK (2015) A social spider algorithm for global optimization. Appl Soft Comput J 30:614–627
12. R Zhao, A Zhou, and K Mao (2016) Automatic detection of cyberbullying on social networks based on bullying features. In: ACM International Conference Proceeding Series, vol 04–07, pp. 1–6
13. Deerwester S, Dumais ST, Furnas GW, Landauer TK (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407
14. Yilmaz S, Toklu S (2020) A deep learning analysis on question classification task using Word2vec representations. Neural Comput Appl 32(7):2909–2928
15. Corallo A et al (2020) Sentiment analysis of expectation and perception of MILANO EXPO2015 in twitter data: a generalized cross entropy approach. Soft Comput 24(18):13597–13607
16. Aaron Sonabend W et al (2020) Integrating questionnaire measures for transdiagnostic psychiatric phenotyping using word2-vec. PLoS One 15(4):1–14
17. T. Hofmann (1999) Probabilistic latent semantic analysis. In: Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)
18. E Altszyler, M Sigman, S Ribeiro, and DF Slezak, (2016) Comparative study of LSA vs Word2vec embeddings in small corpora: a case study in dreams database. arXiv preprint 1–14
19. J Pennington, R Socher, and CD Manning (2014) GloVe: Global Vectors forWord Representation Jeffrey. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543
20. Naili M, Chaibi AH, Ben Ghezala HH (2017) Comparative study of word embedding methods in topic segmentation. Procedia Comput Sci 112:340–349
21. Aguilar J, Salazar C, Velasco H, Monsalve-Pulido J, Montoya E (2020) Comparison and evaluation of different methods for the feature extraction from educational contents. Computation 8(2):1–20
22. C Hua and W Wei, (2019) A particle swarm optimization k-means algorithm for mongolian elements clustering. In: 2019 IEEE Symposium Series Computer Intelligence SSCI 2019, pp. 1559–1564
23. Janani R, Vijayarani S (2019) Text document clustering using spectral clustering algorithm with particle swarm optimization. Expert Syst Appl 134:192–200
24. P Nema and V Sharma, (2016) Multi-label text categorization based on feature optimization using ant colony optimization and relevance clustering technique. In: Proceedings - 2015 International Conference Computer Communication Systems ICCCS 2015, pp. 1–5
25. P Hailong, Z Hui, L Wanglong, and M Ying, (2017) The research on the improved ant colony text clustering algorithm. In: 2017 IEEE 2nd International Conference Big Data Analysis ICBDA 2017, pp. 323–328
26. Cuevas E, Cienfuegos M, Zaldívar D, Pérez-cisneros M (2013) A swarm optimization algorithm inspired in the behavior of the social-spider. Expert Syst Appl 40(16):6374–6384
27. Abirami E (2019) Social spider optimization algorithm: theory and its applications. Int J Innov Technol Explor Eng 8(10):327–331
28. HM Zawbaa, E Emary, AE Hassanien, and B Parv, (2016) A wrapper approach for feature selection based on swarm optimization algorithm inspired from the behavior of social-spiders. In: Proceedings 2015 7th International Conference Soft Computer Pattern Recognition, SoCPaR 2015, pp. 25–30
29. Baş E, Ülker E (2020) An efficient binary social spider algorithm for feature selection problem. Expert Syst Appl 146:113185
30. Abd El Aziz M, Hassanien AE (2018) An improved social spider optimization algorithm based on rough sets for solving minimum number attribute reduction problem. Neural Comput Appl 30(8):2441–2452
31. TR Chandran, AV Reddy, and B Janet, (2016) A social spider optimization approach for clustering text documents. In: Proceeding IEEE - 2nd International Conference Advance Electrical and Electronical Information, Communication Bio-Informatics, IEEE - AEEICB 2016, pp. 22–26
32. Chandran TR, Reddy AV, Janet B (2017) Text clustering quality improvement using a hybrid social spider optimization. Int J Appl Eng Res 12(6):995–1008
33. Hart EM, Avile L (2014) reconstructing local population dynamics in noisy metapopulations — the role of random catastrophes and allee effects. PLoS One 9(10):110049
34. Ochoa I, Juárez-Casimiro A, Olivier K, Camarena T, Vázquez R (2017) Social spider algorithm to improve intelligent drones used in humanitarian disasters related to floods. Nature-inspired design of hybrid intelligent systems. Springer, Cham, pp 457–476
35. Wang W, Chau K, Xu D, Qiu L, Liu C (2017) The annual maximum flood peak discharge forecasting using hermite projection pursuit regression with SSO and LS method. Water Resour. Manag 31:461–477
36. Cuevas E, Cienfuegos M (2014) A new algorithm inspired in the behavior of the social-spider for constrained optimization. Expert Syst Appl 41(2):412–425

37. L Webb and Y Wang, (2013) Techniques for sampling online text-based data sets. In: Advances in Data Mining and Database Management (ADMDM), no. May 2015
38. Indrayan A, Gupta P (2000) Clinical research methods sampling techniques, confidence intervals, and sample size. Natl Med J India 13:29–36
39. Pawde K, Purbey N, Gangan S, Kurup L (2014) Latent semantic analysis in information retrieval. Int J Eng Tech Res 2(10):243–246
40. Landauer TK, Foltz PW, Laham D (1998) An introduction to latent semantic analysis. Discourse Process 25(2–3):259–284
41. Papadimitriou CH, Raghavan P, Tamaki H, Vempala S (2000) Latent semantic indexing: a probabilistic analysis. J Comput Syst Sci 61(2):217–235
42. JC Valle-Lisbo and E Mizraji, (2006) The uncovering of hidden structures by latent semantic analysis. *arXiv*
43. Levy O, Goldberg Y, Dagan I (2015) Improving distributional similarity with lessons learned from word embeddings. Trans Assoc Comput Linguist 3:211–225
44. Chueh C-H, Wang H-M, Chien J-T (2006) A maximum entropy approach for semantic language modeling. Comput Linguist Chin Lang Process 11(1):37–56
45. N Alnajran, K Crockett, D McLean, and A Latham (2017) Cluster analysis of twitter data: a review of algorithms. In: ICAART 2017 - Proceedings 9th International Conference Agents Artificial Intelligence, vol. 2, no. Icaart, pp. 239–249
46. Morissette L, Chartier S (2013) The k-means clustering technique: general considerations and implementation in Mathematica. Tutor Quant Methods Psychol 9(1):15–24
47. Haq EU, Hussain A, Ahmad I (2019) Performance evaluation of novel selection processes through hybridization of k-means clustering and genetic algorithm. Appl Ecol Environ Res 17(6):14159–14177
48. AP Bhopale and KS Sowmya (2017) Novel hybrid feature selection models for unsupervised document categorization.In: 2017 International Conference Advance Computer Communications Informatics, ICACCI 2017, vol. 2017–January, pp. 1471–1477

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.