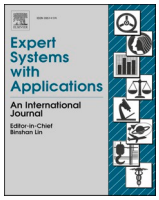
Expert Systems With Applications 202 (2022) 117433

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



A novel approach for text categorization by applying hybrid genetic bat algorithm through feature extraction and feature selection methods

Nazmiye Eligüzel [a,*,1], Cihan Çetinkaya [b,2], Türkay Dereli [c,3]

[a] Gaziantep University, Industrial Engineering, 27310 Gaziantep, Turkey
[b] Adana Alparslan Turkes Science and Technology University, Department of Management Information Systems, 01250 Adana, Turkey
[c] Hasan Kalyoncu University, Office of President, Gaziantep, Turkey

ARTICLE INFO

Keywords:
Bat algorithm
Feature extraction
Feature selection
Genetic algorithm
Uncapacitated P-median problem
Text categorization

ABSTRACT

Due to the rapid incline in the number of documents along with social media usage, text categorization has become an important concept. There are tasks required to be fulfilled during the text categorization, such as extracting useful data from different perspectives, reducing the high feature space dimension, and improving effectiveness. In order to accomplish these tasks, feature selection, and feature extraction gain importance. This paper investigates how to solve feature selection and extraction problems. Also, this study aims to decide which topics are the focus of a document. Moreover, the Twitter data-set is utilized as a document and an Uncapacitated P-Median Problem (UPMP) is applied to make clustering. In this study, UPMP is used on Twitter data collection for the first time to collect clustered tweets. Therefore, a novel hybrid genetic bat algorithm (HGBA) is proposed to solve the UPMP for our case. The proposed novel approach is applied to analyze the Twitter data-set of the Nepal earthquake. The first part of the analysis includes the data pre-processing stage. The Latent Dirichlet Allocation (LDA) method is applied to the pre-processed text. After that, a similarity (distance) matrix is generated by utilizing the Jensen Shannon Divergence (JSD) model. The study's main goal is to use Twitter to assess the needs of victims during and after a disaster. To evaluate the applicability of the proposed approach, experiments are conducted on the OR-Library data-set. The results demonstrate that the proposed approach successfully extracts topics and categorizes text.

1. Introduction

Increased internet usage leads to an exponential increase in the number of digital text documents. Therefore, it can be said that organizing text data gains importance. At this point, text categorization becomes a significant concept. Text categorization is characterized by determining documents into a set of pre-defined categories based upon the classification types (Uğuz, 2011). It is a significant process to reveal useful and beneficial information from the document. The high dimension of features, irrelevant and redundant terms, and noisy data are major problems of text categorization. To eliminate these problems and increase the effectiveness of text categorization, there are mainly two utilized dimensionality reduction methods in the literature, which are feature selection and feature extraction. Feature selection is a manner of selecting a minor subset of features that, preferably, are necessary and adequate to define the target concept (Yusta, 2009). Moreover, feature selection methods do not try to create new terms but try to choose the best ones from the primary set (Sebastiani, 2003). There are some feature selection methods applied to the text categorization problems, such as term frequency-inverse (TFI), document frequency (DF), mutual information (MI), Chi-square, information gain (IG), and term strength. Many feature selection techniques have been successfully implemented for text categorization (Galavotti et al., 2007; Kotcz, 2001; Moens & Dumortier, 2000; Soucy & Mineau, 2002; Taira & Haruno, 1999; Yang, 1997 Zhang et al. 2008). Feature extraction creates new features. It is the process in which the feature space dimension is reduced by integrating or transforming the original feature set. (Alsmadi & Gan, 2019). Principal component analysis (PCA), Latent Dirichlet Allocation (LDA),

* Corresponding author.
E-mail addresses: nazmiye@gantep.edu.tr (N. Eligüzel), ccetinkaya@atu.edu.tr (C. Çetinkaya), turkay.dereli@hku.edu.tr (T. Dereli).
[1] ORCID no: https://orcid.org/0000-0001-6354-8215.
[2] ORCID no: http://orcid.org/0000-0002-5899-8438.
[3] ORCID no: https://orcid.org/0000-0002-2130-5503.

https://doi.org/10.1016/j.eswa.2022.117433
Received 24 March 2020; Received in revised form 23 March 2022; Accepted 26 April 2022
Available online 29 April 2022
0957-4174/© 2022 Elsevier Ltd. All rights reserved.

Latent Semantic Indexing (LSI), clustering methods, etc. are feature extraction methods used for text categorization. Some feature extraction methods have been successfully applied to text categorization (Harrag et al., 2010; Li et al., 2008; Liu et al., 2010). In addition, biologically inspired algorithms such as genetic algorithm (Ghareb et al., 2016), ant colony (Joseph Manoj et al., 2018), artificial bee colony (Wang et al., 2018), and particle swarm optimization (Xue et al., 2012), have been implemented with feature selection and feature extraction techniques.

In this study, a novel hybrid genetic bat algorithm (HGBA) is demonstrated to extract center tweets in the document. The proposed algorithm is improved to solve the problem given in this study. It is the first paper that utilizes an Uncapacitated P-Median Problem (UPMP) to make feature selection in text categorization by applying the HGBA. Also, to the best of our knowledge, it is the first time that the bat algorithm has been applied to the UPMP. In the literature, many clustering algorithms were applied to Twitter data-sets and the average number of clusters was found to be 7, with a minimum of 2 and a maximum of 10 (Alnajran et al., 2017; Liang, 2010). Therefore, the scope of the study requires applying 2–10 centers due to effective clustering. To evaluate the applicability of the proposed novel approach, experiments are conducted on the P-median data-set with 5 centers and 10 centers from OR-Library. All in all, to the best of the authors' knowledge, LDA, JSD, and P Median applications are utilized for the first time sequentially in order to provide social insight into the earthquake by gathering center tweets. Topics are extracted by the application of LDA, and then JSD is utilized on the output of LDA to obtain a distance matrix. Finally, the P-Median model is used on the obtained distance matrix, and the model is solved by a novel HGBA *meta*-heuristic. The rest of this paper is organized as follows: Section 2 demonstrates the literature review; Section 3 presents research methodology (a brief overview of the pre-processing stage in the text categorization, the LDA model as feature extraction, the Jensen Shannon Divergence (JSD) model, UPMP, the genetic algorithm, and the bat algorithm). Section 4 presents the applicability of the proposed novel approach and the experimental results. Section 5 represents the computational experiments of the case study and includes a brief discussion of the results. Finally, the paper is concluded in section 6.

2. Literature review

In the literature, optimization algorithms are proved to be successful in various applications (Chiang et al., 2014; Eligüzel & Özceylan, 2021; Precup et al., 2021; Preitl et al., 2006). In addition, optimization and *meta*-heuristic algorithms have been utilized successfully in text categorization. In this section, the related works are presented.

Most of the publications in the literature have utilized feature selection methods to decrease the dimensionality of the feature space in text categorization. Therefore, different *meta*-heuristic algorithms have been utilized with feature selection methods. Several studies demonstrated the application of Particle Swarm Optimization (PSO) as a feature selector in text categorization (Chen et al., 2012; Chuang et al., 2011; Xue et al., 2012; Zahran & Kanaan, 2009). A novel feature selection method that comprehends TFI, DF and, chi-square methods was proposed by the utilization of the PSO technique, and it was used for Chinese text categorization (Jin et al., 2010). The aforementioned method has a reduced dimension of feature space effectively. A study presented an improved Genetic Algorithm (GA) using a feature selection method for disaster-related tweet categorization (Benitez et al., 2018). Dimensionality reduction and classification accuracy were succeeded by this proposed method. Uguz presented a two-stage method for text categorization (2011). In the first level, the information gain method was used as a feature selector to rank every term in a document based on its importance. In the second level, GA and PCA methods were implemented separately as feature selectors and feature extractors to decrease the feature space dimension. To measure the effectiveness of proposed methods in two stages, C4.5 and k-Nearest Neighbor classifier techniques were utilized. Higher classification performance was obtained as a result. A novel clustering technique was proposed for tweets by integrating the K-Means clustering approach and the GA evolutionary approach (Dutta et al., 2017). Experiments were conducted on data-sets of real tweets from the Uttarakhand floods in 2013. This technique provided better results than existing clustering methods. The Variable Length Chromosome GA (VLCGA) was proposed to analyze Twitter sentiment (Fatyanosa et al., 2019). The VLCGA was combined with Naïve Bayes (VLCGA-NB). The comparison was conducted between Naïve Bayes and the VLCGA-NB. As a result, VLCGA-NB performed better accuracy than Naïve Bayes. A comparative study was proposed to solve the feature selection problem (Yusta, 2009). Different *meta*-heuristic strategies, including Grasp, Tabu Search, and Memetic Algorithm were compared with GA and with other ordinary feature selection methods. The acquired results indicated that GRASP and Tabu Search perform better accuracy than the other methods. The Ant Colony Algorithm (ACA) was applied as a feature selector for text categorization (Aghdam et al., 2009). The performance of the proposed algorithm was compared with other feature selection methods that are GA, IG and Chi-Square. The results demonstrated the superiority of the ACA as a feature selector among other algorithms. Several studies showed the application of the LDA method as a feature extractor in text categorization using the GA in the feature selection stage (Chen et al., 2017; Panichella et al., 2013; Sotiropoulos et al., 2014; Sotiropoulos et al., 2016). The clustering problem was reformulated by Sotiropoulos et al. (2016) as a discrete optimization problem within the n-dimensional standard simplex. They utilized LDA for topic clustering and a novel GA approach by integrating a centroid-based encoding scheme. Results showed the superiority of the proposed approach among other traditional clustering algorithms. Chen and Zhang et al. (2008) introduced a feature selection method by utilizing the GA, in which dimensions of LDA input were reduced, and generated topics were converged to be more meaningful. A novel LDA and GA was proposed to identify near-optimal configuration for LDA by considering different software engineering tasks: the first one is traceability link recovery, the second one is feature location, and the last one is software artifact labeling (Panichella et al., 2013). Results demonstrated that LDA-GA can determine robust LDA configurations. Sotiropoulos et al. (2014) proposed a group detection method within a subcommunity of Twitter micro-bloggers by utilizing a topic modeling. They utilized the LDA topic modeling technique to identify clusters of Twitter users considering both spatial and topical compactness.

2.1. Motivation and contribution of the our study

To the best of our knowledge, there is only one study that applied PMP to the text categorization (Liang, 2010). In the study, Liang proposed a new ACA which was applied to capacitated P-Median Problem (PMP) to make text categorization. Our study has a new contribution to the literature. To the best of our knowledge, this is the first study that applies UPMP to a Twitter data-set for gathering clustered tweets and extracting center tweets. Besides, a novel HGBA is proposed due to solving UPMP for our problem. Furthermore, LDA and JSD are utilized with PMP for the first time to generate a similarity (distance) matrix. The study aims to reduce the feature space dimension, determine center topics in earthquake tweets, remove irrelevant and redundant terms, and eliminate noisy data. The main contribution of the study is to determine the needs of victims during and after the disaster. The potential of the proposed method is demonstrated via a real-life application, namely the Nepal earthquake Twitter data-set. Another contribution of this paper is to provide social insight after an earthquake from the same perspective of the study proposed by Eligüzel et al. (2021).

3. Methodology

This section demonstrates the critical steps in the implementation of the LDA feature extraction technique by using the feature selection

method to extract important topics from the Nepal earthquake Twitter data-set. Fig. 1 illustrates the proposed text categorization structure, including sequential steps and techniques.

All in all, this is the first time that a combination of LDA, JSD, and P-median applications have been utilized together in the literature sequentially. The main aim and novelty of this study is to provide social insight into the earthquake by gathering center tweets by using a novel approach which is the combination of aforementioned methods for the first time. Stages of the novel proposed approach are as follows: Topics are extracted by the application of LDA, and then JSD is utilized on the output of LDA to obtain a distance matrix. Finally, the P-median model is used on the obtained distance matrix of a documents, and the model is solved by a novel HGBA *meta*-heuristic. In addition, another novelty of the proposed paper is to apply the HGBA to a UPMP in order to make a feature selection in text categorization.

In this study, the similarity of the documents is treated as a distance. Therefore, similarities are utilized in the P-median problem application to gather center points by aiming to have a general perspective on a group of documents. Lastly, the given problem is solved by a *meta*-heuristic approach, which is HGBA.

3.1. Research data-sets

The P-median data-set from OR-Library (Beasley, 1985) and human-labeled tweets gathered from the 2015 Nepal earthquake (Alam & Shafiq Joty, 2018) is utilized in the proposed study experiments.

3.2. Pre-Processing

Data pre-processing is the operation of cleaning data and making data ready for the other processes. This process is implemented by utilizing Python 3.6 and MATLAB software.

3.2.1. Removing of URLs, usernames, stop-words, punctuation, and infrequent words

URLs (Uniform Resource Locator) referred to as a web address and they may misclassify the texts. Usernames are not related to the content of the text. Stop-words such as pronouns and conjunctions do not carry meaning. Some of the more often utilized stop-words for English is ‘a’, ‘an’, ‘the’, ‘of’, ‘I’, ‘you’, ‘it’, etc., that are not related to the content of the text. Punctuation does not make sense alone. Therefore, applying all these removing steps is significant to make an accurate classification. It is also important to remove infrequent words in the text.

3.2.2. Tokenizing

Tokenization is the process of splitting the strings into pieces such as phrases, symbols, words, keywords, and other elements called tokens (Techopedia, 2019). In the proposed study, the tokenization step is applied.

3.2.3. Stemming

It is the process for obtaining the roots of the words. The words with the same roots seem to be different words because of the affixes they receive. To avoid this situation, Porter’s stemming algorithm (Porter, 1980) is utilized in the stemming process.

3.3. Latent dirichlet allocation (LDA)

Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus (Blei et al., 2003). The leading idea is that documents are demonstrated as random mixtures of latent topics, where every single topic is described by a distribution of words. It finds fundamental topics in a collection of documents and obtains probabilities of words in topics (MathWorks, 2019). LDA is an unsupervised learning algorithm. If data is labeled, it becomes supervised learning. In Fig. 2, a schematic matrix of LDA is shown.

According to Fig. 2: D is the documents to words matrix, θ is the topic distribution for each document, and β is the parameter of the each-topic word distribution. The LDA model involves a three-level model. The probability of a corpus can be seen in Eq. (1).

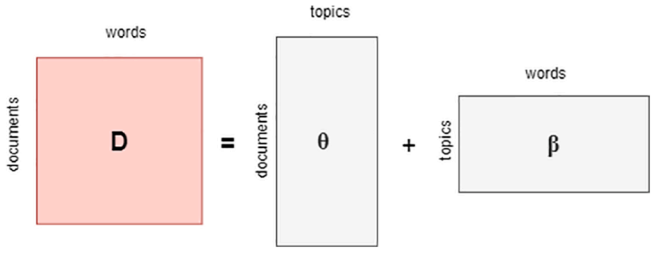


Fig. 2. Structure of LDA Model.

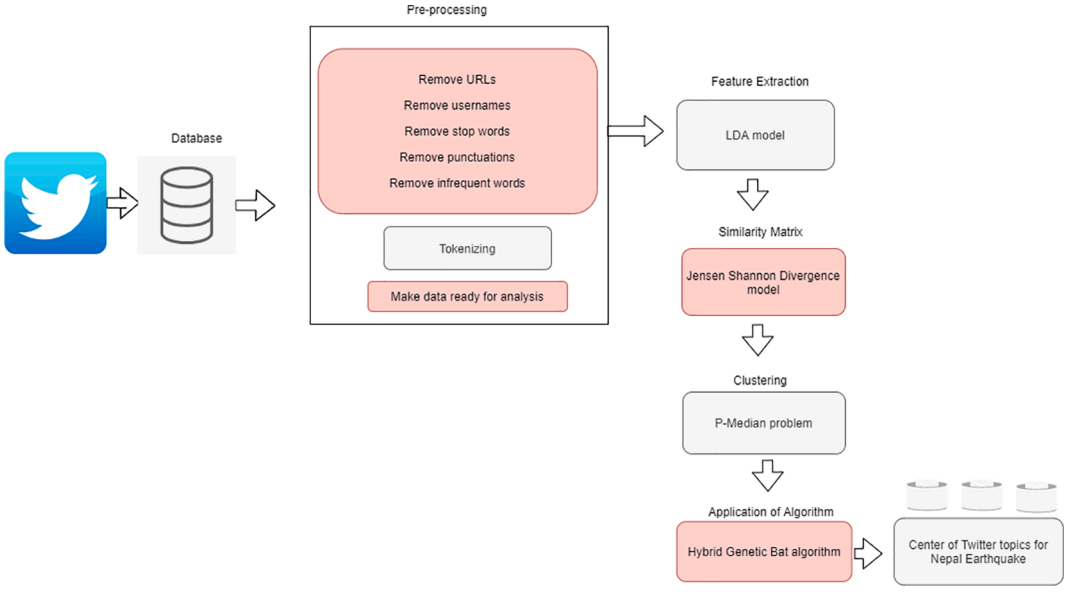


Fig. 1. Structure of proposed model.

$$p(D|\alpha,\beta)=\prod_{d=1}^{M}\int P(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}}p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d \quad (1)$$

α is the Dirichlet-previous concentration parameter of each document topic distribution, β is corpus level parameter, θ_d is the document-level variable, z_{dn} is the topic assignment for w_{dn}, w_{dn} is the n^{th} word in the d^{th} document, N is the number of words in the document, M is the number of documents to analyze, D is the corpus of collection M documents.

The LDA process is implemented using MATLAB software. It is taken as M = 7000 documents and K = 50 topics for the Nepal earthquake model and Dirichlet hyper-parameters β = 0.1 and α = 50 / *K* (Tong & Zhang, 2016). In Table 1, a schematic probability matrix is shown in the model.

In the study, the document to the topic matrix is utilized to find the similarity matrix (distance matrix) for each document.

3.4. Jensen- Shannon Divergence model (JSD)

It is the popular method to find similarity between two probability distributions in probability theory and statistics. JSD is based on Kullback-Leibler Divergence but some useful differences. JSD is symmetric and it is all the time finite value. The square root of the JSD is considered as a metric (Fuglede & Topsøe, 2004). It is a symmetrized version of the Kullback-Leibler Divergence DKL(P\\Q). The formulation for JSD is shown in Eqs. (2) and (3).

$$\mathrm{JSD(P\backslash\backslash Q)}=\frac{1}{2}DKL(P\backslash\backslash M)+\frac{1}{2}DKL(Q\backslash\backslash M) \quad (2)$$

where M=.$\frac{1}{2}(P+Q)$

$$\mathrm{DKL(P\backslash\backslash Q)}=\sum_{x\in X}P(x)\log\left(\frac{Q(x)}{P(x)}\right) \quad (3)$$

P and Q are two probability distributions. The distribution of Q denotes instead a theory, a model, an illustration, or an approximation of P. *x* is probability space where P and Q are defined. The square root of the JSD is the metric of JSD distance. The smaller JSD means the more similar two distributions. In this study, the JSD formula is applied to the document to the topic matrix to find the similarity matrix of documents. At the end of the process, a 7000 × 7000 similarity (distance) matrix is obtained.

3.5. P-Median problem for clustering

The P-Median Problem (PMP) is a well-known facility location problem. The aim of the uncapacitated discrete PMP is the minimization of the total distance with the pre-decided number of medians (center) which supply rest of the points (demand nodes) under the condition of each demand nodes served from only one median. The problem formulation is given in Eqs. (4) to (8), respectively (Teixeira & Antunes, 2008).

Decision variables:

$$y_k=\begin{cases}1, if source node k is selected (\forall k\in K)\\ 0, otherwise\end{cases}$$

Table 1
Schematic probability matrix.

	$Topic_1$	$Topic_2$	$Topic_3$	⋯	$Topic_l$
D_1				⋯	$P_{1,K}$
D_2					$P_{2,K}$
⋯	⋯	⋯	⋯	⋯	⋯
D_M	$P_{M,1}$	$P_{M,2}$	$P_{M,3}$	⋯	$P_{M,K}$

$$x_{ik}=\begin{cases}1, if demand node i is assigned to source node k (\forall i\in K,\forall k\in K)\\ 0, otherwise\end{cases}$$

Parameters:

d_{ik} = Distance between source node k and demand node i.
p = Number of source node to be opened.

Objective function:

$$\mathrm{Min}\,Z=\sum_{i\in I}\sum_{k\in K}d_{ik}x_{ik} \quad (4)$$

Subject to.

$$\sum_{k\in K}x_{ik}=1\forall i\in K \quad (5)$$

$$x_{ik}\le y_k\forall i\in K,\ \forall k\in K \quad (6)$$

$$\sum_{k\in K}y_k=p \quad (7)$$

$$y_k,\ x_{ik}\in\{0,\ 1\forall i\in K,\ \forall k\in K \quad (8)$$

where the objective function (4) aims to minimize total distance. The assignment of each demand node to only one source node is provided by constraint (5), while constraint (6) ensures the assignment of demand nodes to the opened source nodes. Constraint (7) detects the number of source nodes that should be opened. Constraint (8) is related to the decision variables.

The major dimensionality of the feature space is the main problem for the text categorization. At the end of the pre-processing, LDA and JSD steps, documents are required to be clustered according to the similarity matrixes to eliminate insignificant or comparatively insignificant documents. In the study, center documents and documents around center documents are obtained with HGBA application on the UPMP.

3.6. Application of algorithms

A novel hybrid genetic bat approach is applied to make categorization for earthquake-related tweets, in which genetic algorithm and bat algorithm are used together to resolve combinatorial optimization problems and improved the searching mechanism to reduce high dimension feature space.

3.6.1. Genetic algorithm (GA)

A GA is an optimization method that has been inspired by biological progression (Oksuz et al., 2016). For the UPMP, the individuals will be an array with genes as the number of centers (medians). The basic procedure of the GA used in the study is shown below.

Begin:
Setting parameters
Initializing the entire population
for each individual in population
a. Selecting p-medians randomly
b. Appointing all demand points to nearest centroid (median)
c. ***for*** each centroid
  i. Allocate each node to its closest centroid
  ii. Replace the centroid with center point
  iii. Calculate fitness value for each individual
a. ***end***
end
Keep the best individual
Selection
Crossover
Mutation

Return the ***population***
end

3.7. Calculation of fitness value

The fitness value is computed for each individual by utilizing Eq.4.

3.8. Selection

The selection operator is chosen as a ranking-based method. This selection method ranks the population with respect to fitness values. Eq.9 represents the selected individual from the list (Correa et al.,2004).

$$\text{Select (R)} = \left\{ rj \in R \middle| j = L - \left\lfloor \frac{-1+\sqrt{1+4rnd\left(L^2+L\right)}}{2} \right\rfloor \right\} \tag{9}$$

where L represents population number, R is the list of individuals (R=(r_1, r_2,…,r_L)), and rnd is a random number between 0 and 1. It searches for high-quality solution.

3.9. Crossover

In the crossover process, a random value is generated and the crossover is performed for individual pairs coming from the selection stage. It is important to ensure that the same genes do not come together for PMP.

3.10. Mutation

The mutation process is conducted according to the mutation probability. A randomly selected center node is replaced with a randomly selected node.

3.10.1. Bat algorithm

The classical Bat Algorithm was introduced by Yang (2010) based on the echolocation behavior of bats. The bat algorithm is a continuous optimization algorithm. The mathematical equations for bat algorithms are shown from Eq. (10) to (14).

$$fi = fmin + (fmax - fmin)\beta \tag{10}$$

$$vi^t = vi^{t-1} + (xi^{t-1} - x^*)fi \tag{11}$$

$$xi^t = xi^{t-1} + vi^t \tag{12}$$

$$Ai^{t+1} = \propto Ai^t \tag{13}$$

$$ri^{t+1} = ri^0[1 - \exp(-\gamma t)] \tag{14}$$

where f, v, x, x*, A, t, and r represent frequency, velocity, location, the current best solution, loudness, iteration, and pulse rate respectively.

For each iteration, every bat is created by the application of Eq. 11–12 in orderly. $\beta \in [0\,1]$ is taken from a uniform distribution and α and γ are taken as constants. With a varying frequency, pulse rate and loudness, bats fly randomly for prey at a position xi with a velocity vi (Chansombat et al., 2018). To solve combinatorial optimization problems, discretization is essential.

3.10.1.1. Proposed discrete bat algorithm. The Bat Algorithm cannot be directly utilized for PMP. Some modifications are performed in the light of literature (Osaba et al., 2016; Osaba et al., 2019). In the proposed algorithm, frequency (*fi*) (Eq. (10)) is not considered because it is specifically used for continuous problems. Eq. (11) is modified in a way that vi is taken as the distance between the bat i and the best bat of the population and Eq.15 is generated.

$$vi^t = Random[1, differencefunction(xi^t, x^*)] \tag{15}$$

The velocity vi is the random number generated between 1 and the number gathered from the difference functions of a bat i at iteration t by the discrete uniform distribution. The difference function calculates several non-corresponding genes for two individuals. Hamming Distance was previously utilized to measure distances in Travelling Salesman problem (Osaba et al., 2016). For the PMP, the Hamming Distance is not appropriate. It looks at the similarity of opposite genes in two individuals. It is inadequate for the PMP because it is required to make a comparison between one gene and the rest of the genes. In the proposed problem, the difference function is utilized instead of Hamming Distance. In difference function, all genes of two individuals are compared. The total different genes of the two individuals are calculated. The movement of the bat has changed in the light of velocity. The position of bat i is modified by considering the following three operators, called "movement functions" (Eq. (16)-Eq.18) with their mathematical demonstrations:

$$xi^t = \leftarrow SwappingOperator\left(xi^{t-1}, vi^t\right) \tag{16}$$

$$W_{best1} = [x_1^{best1}, x_2^{best1}, \cdots\cdots.., x_n^{best1} \tag{16.1}$$

$$W_{best2} = [x_1^{best2}, x_2^{best2}, \cdots\cdots.., x_n^{best2} \tag{16.2}$$

$$Random_{w_{best1}} = x_y^1 \tag{16.3}$$

$$Random_{w_{best2}} = x_j^2 \tag{16.4}$$

$$W_{best1}^{`} = [x_1^{best1}, x_2^{best1}, \cdots x_j^{best2} \cdots.., x_n^{best1} \tag{16.5}$$

$$W_{best2}^{`} = [x_1^{best2}, x_2^{best2}, \cdots x_y^{best1} \cdots.., x_n^{best2} \tag{16.6}$$

$$xi^t = \leftarrow ReplacementOperator\left(xi^{t-1}, vi^t\right) \tag{17}$$

$$W_{best} = [x_1^{best}, x_2^{best}, \cdots x_j^{best} \cdots.., x_n^{best} \tag{17.1}$$

$$Random_{gene} = Y \tag{17.2}$$

$$W_{best}^{`} = [x_1^{best}, x_2^{best}, \cdots Y_j^{best} \cdots.., x_n^{best} \tag{17.3}$$

$$xi^t = \leftarrow RandomnessOperator\left(xi^{t-1}, vi^t\right) \tag{18}$$

According to the pulse rates and loudness, utilized movement functions can be varied. For the swapping operator, the genes are exchanged between the two best individuals in a population. For the replacement operator, a random gene assignment is made to replace a random gene of the best individual. In addition to the swapping and replacement operators, we have also used the randomness operator. In this operator, completely random individuals are obtained. A new initial matrix is obtained with the help of these operators. Loudness and pulse rate formulas are given in Eqs. (19) to (20).

$$pulserate = \alpha^* pulserate \tag{19}$$

$$loudness = loudness^*(1 - \exp(-\gamma^* i)) \tag{20}$$

where α and γ are the constants. In the proposed study, we use $\alpha = \gamma =$ 0.98 (Osaba et al., 2019).

The proposed approach has both intensification and exploration steps. Unlike in the literature, the loudness is increased and the pulse rate is decreased while iterations are pursued. As the loudness increases, the intensification (convergence) increases and the swapping operator is activated. As the pulse rate decreases, exploration decreases and the replacement operator is activated. The purpose of intensification is to increase focus. Namely, it searches for other good neighbor solutions. Exploration is important for global search and intensification is

important for local search. Then, to balance intensification and exploration in the bat algorithm, the randomness operator is activated.

3.10.2. Hybrid algorithm

Our hybrid algorithm consists of two parts. The first part includes the application of a bat algorithm to provide a better starting point, and the second part comprises a modified GA. The genetic part of the proposed novel hybrid algorithm works for finding near-optimal results. Therefore, the presented hybrid approach leads to an improved starting point in solution space and achieves optimal results more quickly. This is accomplished by managing the balance between intensification and exploration. To increase the understandability of the algorithm, its flowchart is shown in Fig. 3.

Flowchart begins with the initialization part which includes uploading data, setting parameters, initializing population, and evaluation of the initial population. The parameters, which are loudness, pulse rate, velocity, alpha, number of population, mutation rate, and number of the iteration are user defined. Randomly selected parameters are "rnddiff", "rnd", rndl". Then, best individuals from the population are considered in order to provide fast convergence to optimal or near optimal solution. After that multiplication part is as follows; random selection of two individuals is implemented and different gene(s) between the selected individuals are extracted. Therefore, the range of random number generation is decided. Next, swapping, replacement, and randomness operators are used. In that way, three random numbers are generated.The first one (rnddiff) is generated between 1 and the calculated number of different genes, the second one (rnd) is generated between 1 and the number of medians, and the last one (rndl) is generated between 0 and 1. If rnddiff is lesser than rnd, and rndl is lesser than loudness, swapping operator is activated. So, pulse rate decreases, loudness increases, and convergence is increased. If pulse rate is lesser than rndl, replacement operator is activated. Therefore, exploration is decreased. Finally, randomness operator is activated if aforementioned two conditions are not met. After this step, a new matrix is generated for GA application and best solution from this part is stored. In the GA application, steps consist of selection, crossover, and mutation operations. After GA application, new solutions are gathered and the best of these solutions is stored. Lastly, two of the stored best solutions are compared and better one is kept to pass GA application for the next iteration till number of maximum iteration is reached. Therefore, augmented algorithm by selecting population with the best fitness values is proposed with integration of discrete bat algorithm with GA application.

In a nutshell, Applied algorithm in the proposed study is a mixed approach which includes bat and genetic algorithm. Bat algorithm is used to find initial population. The bat algorithm is modified in order to apply to discrete problems. To accomplish modification process, each bat in the population indicates a possible solution which is the set of center points as an index for the P-median problem. P-median objective function is used as objective function in the proposed algorithm. In that way, distance between center point and connected points to given center is calculated. The aim of the objective function is minimize the sum of the distances for each bat. In this process, given number of population is generated and each population contains number of pre-decided bat. For each population, one of the movement functions is selected in accordance with mentioned rules. Therefore, objective values are gathered and bat with minimum value is selected to processed by genetic algorithm in order to converge optimum or near optimum value. In that way, it is provided a better initial solution in the solution space for genetic algorithm process. The paper is inspired by improved bat algorithm for the TSP (Osaba et al., 2016). The parameter frequency is not taken into account and velocity has been modified by using difference function that compares all genes each other. We integrated crossover part of genetic algorithm into bat algorithm by utilizing swapping, replacement, and randomness operator to convert bat algorithm into discrete space.

The pseudo-code of the HGBA is presented as follows.

set parameters: loudness,pulse rate, velocity, alfa, number of population, mutation rate, and, number of iteration;
load the distance matrix;
generate initial matrix;
for i = 1: number of population
a. calculate fitness value for each individual (ith)
end
choose two of the best individuals
for i = 1:number of population
a. diff = calculate the number of different genes for between the chosen two individuals;
b. rnddiff = generate random number 1 to diff;
end
for i = 1:number of population (generation of new initial matrix)
a. rnd = generate a random number between 1 to number of medians;
b. rndl = generate a random number from 0 to 1;
c. ***if*** rnddiff < rnd && rndl < loudness
  i. utilize swapping operator;
  ii. decrease the pulserate by multplying with alfa;
  iii. increase loudness by multiplying with (1-exp(-alfa*i));
a. ***elseif*** pulserate < rndl
  iv. utilize replacement operator;
d. ***else***
  i. utilize randomness operator;
e. ***end***
end
for i = 1: number of population
a. ***calculate*** fitness value for each individual in new initial matrix (ith)
end
for iteration = 1:number of iteration
a. increase the iteration by 1;
b. select ***the*** p1 (parent 1) from the best of the new initial matrix;
c. ***calculate*** the fitness value of p1;
d. select the p2 (parent 2) by utilizing ranking based method;
e. ***calculate*** the fitness value of p2;
f. apply ***the*** crossover operator utilizing p1 and p2;
g. ***calculate*** the fitness values of offsprings (y1, y2);
h. (best1, best2) = ***select*** the best two individuals from p1, p2, y1, and y2;
  i. (m1, m2) = apply the mutation operator on best1 and best2;
i. ***newindividual*** = min(m1,m2);
j. ***if*** newindividual < p1;
  i. update p1 with newindividual
k. ***else***
  i. continue
l. ***end***
end

The code and data-set is provided at https://github.com/neliguzel/BatGenetic.

4. Experimental design

The P-median data-sets from OR-Library (Beasley, 1985) are used for experimenting on the proposed novel hybrid algorithm. All 5 and 10 centered data-sets are tested. All processes are applied by the R2018b Matlab software package. The algorithm is run five times. All experiments are run on a computer with intel ® Core ™ i7, 2 GHz CPU, 8 GB of RAM, and the Windows 8 operating system. The results of the 5- and 10 -centered data-sets are given in Table 2 and Table 3, respectively. Iteration numbers are taken as 5000 for 5-centered data-sets and 50,000 for 10-centered data-sets, steadily.

To make an experimental design, an algorithm is run with different populations. Moreover, optimal results are obtained in different

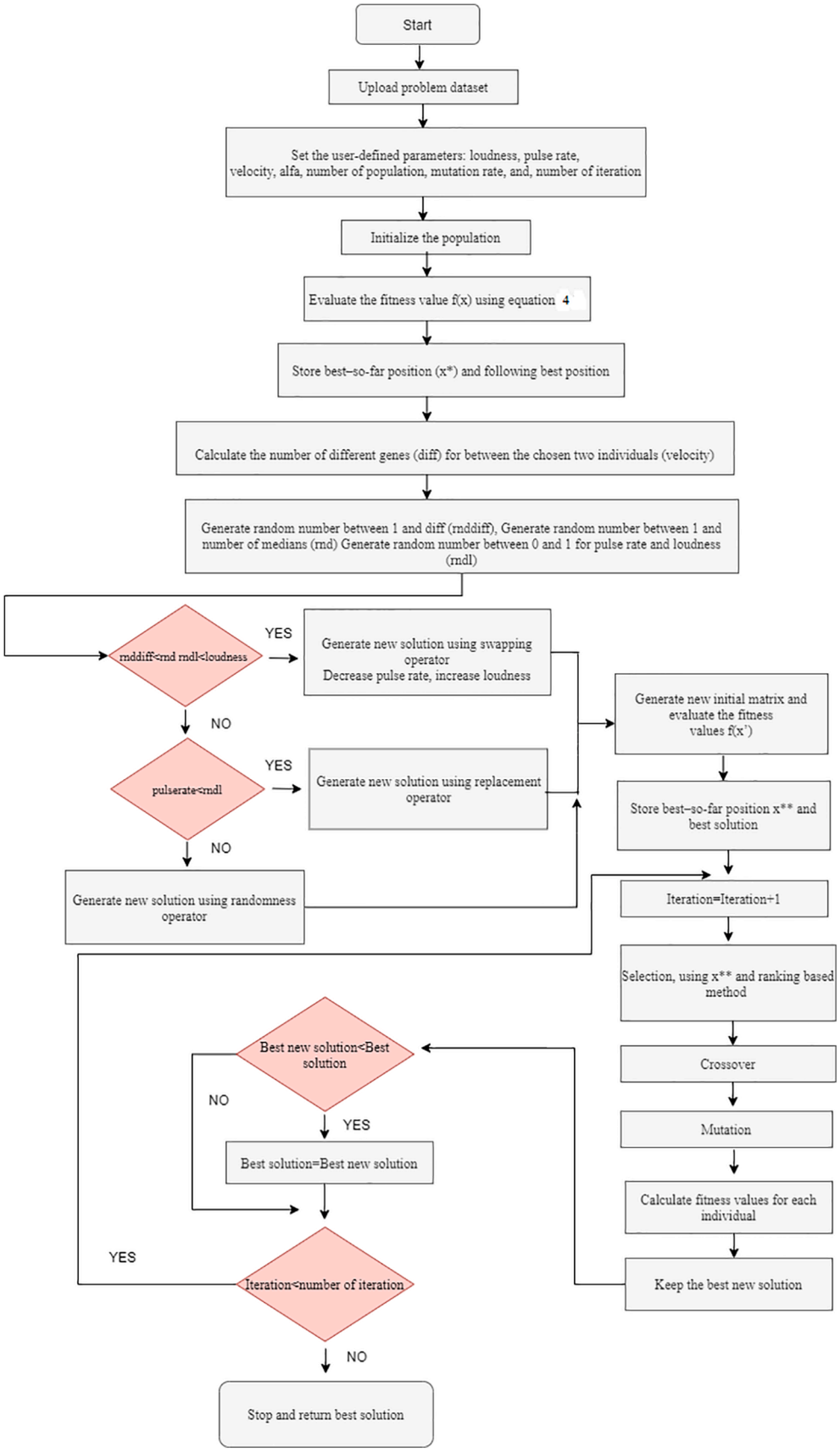


Fig. 3. Flowchart of the proposed hybrid algorithm.

Table 2
Results of 5-centered P-median data-sets.

Data Set	Number of population	Minimum found	Optimum	Gap (%)	Iteration number	Elapsed time (second)
P- median 1	100	5819	5819	0	281	0.524940
P- median 6	100	7824	7824	0	1442	3.95824
P- median 11	100	7696	7696	0	2288	10.931223
P- median 16	100	8162	8162	0	1592	31.515512
P- median 21	100	9292	9138	1.6	1500	52.518273
	500	9179	9138	0.44	2000	74.606992
	1000	9138	9138	0	2127	96.183492
P- median 26	100	9979	9917	0.62	3991	195.31403
	500	9917	9917	0	3340	215.39194
P- median 31	100	10,112	10,086	0.257	3509	267.43427
	500	10,086	10,086	0	2242	158.21278
P- median 35	100	10,406	10,400	0.05	3180	252.31718
	500	10,515	10,400	1.10	2571	243.61755
	1000	10,482	10,400	0.788	1828	225.29375
	2000	10,400	10,400	0	4824	398.80736
P- median 38	100	11,164	11,060	0.94	1436	163.86167
	500	11,101	11,060	0.37	2574	271.54102
	1000	11,128	11,060	0.61	3340	279.44711
	2000	11,072	11,060	0.108	3580	408.75727
	5000	11,060	11,060	0	3216	403.53612

Table 3
Results of 10-centered P-median data-sets.

Data Set	Number of population	Minimum found	Optimum	Gap (%)	Iteration number	Elapsed time (second)
P- median 2	500	4102	4093	0.21	4012	8.605438
	1000	4093	4093	0	6036	13.47548
P- median 3	500	4250	4250	0	4800	10.04650
P- median7	500	5639	5631	0.14	6905	34.44492
	1000	5631	5631	0	7442	38.91450
P- median 12	500	6634	6634	0	5476	69.17323
P- median 17	500	7050	6999	0.72	2773	139.6304
	1000	6999	6999	0	9967	374.7505
P- median 22	500	8680	8579	1.17	4461	303.2037
	1000	8605	8579	0.30	4024	314.3202
	2000	8586	8579	0.08	3297	317.4152
	5000	8579	8579	0	3892	335.2682
P- median 27	500	8332	8307	0.30	4895	421.1516
	1000	8322	8307	0.18	4253	407.8592
	2000	8310	8307	0.036	4561	418.2961
	5000	8307	8307	0	4114	416.1256
P- median 32	500	9301	9297	0.04	5684	601.1259
	1000	9315	9297	0.19	6012	612.1587
	2000	9301	9297	0.04	4965	607.1236
	5000	9297	9297	0	4309	604.8629
P- median 36	500	9966	9934	0.32	5128	671.1589
	1000	9955	9934	0.21	5216	673.1596
	2000	9953	9934	0.19	3589	665.1741
	5000	9934	9934	0	4312	687.1233
P- median 39	500	9456	9423	0.35	5697	716.4129
	1000	9485	9423	0.65	5869	724.5874
	2000	9460	9423	0.39	5124	712.4893
	5000	9466	9423	0.45	5128	798.4125
	10000	9423	9423	0	5931	928.3987

populations and different iterations for each 5-centered and 10-centered P-median data-set. In Figs. 4-5, the results of the 5–10 centered data-sets and experimental design for HGBA are given.

The amount of time MATLAB spends terminating operations is referred to as elapsed time, which demonstrates the time in seconds.

Considering Eq. (21), the gap between minimum found and the optimal result is.

$$Gap = \frac{f(current) - f(opt)}{f(opt)} x100 \tag{21}$$

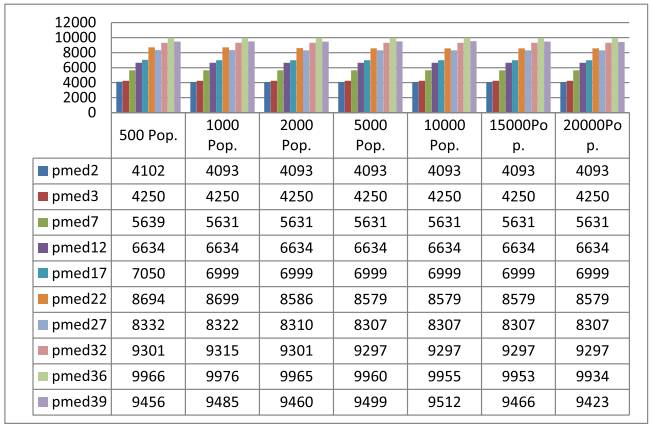
where f(current) indicates the solution found by HGBA and f(opt) refers to the optimum solution for a given data-set. As is seen in Table 2, optimal results are found using the proposed algorithm for 5 and 10 centered data-sets.

In Fig. 4, it can be easily seen that optimal results are obtained in a low number of the population compared to the rest of the 5 centered data-sets for Pmed1 to Pmed16. For the rest of the 5 centered data-sets, the number of population increases to gather optimal results concerning the size of the data-sets. Furthermore, it is necessary to state that optimal results are found at 5000 population without taking the size of the data-sets into account.

In Fig. 5, the same application with 5 centered data-sets is applied on the 10 centered data-sets. 10 centered data-sets can be taught as more complex data-sets due to the difference between the numbers of centers. Therefore, it is expected to reach optimal results with a larger population compare to 5 centered data-sets. It is observed that the optimal results for 10 centered data-sets from Pmed2 to Pmed17 are reached

	100 Pop.	500 Pop.	1000 Pop.	2000 Pop.	5000 Pop.
pmed1	5819	5819	5819	5819	5819
pmed6	7824	7824	7824	7824	7824
pmed11	7696	7696	7696	7696	7696
pmed16	8162	8162	8162	8162	8162
pmed21	9292	9179	9138	9138	9138
pmed26	9979	9951	9924	9945	9917
pmed31	10193	10087	10102	10095	10086
pmed35	10406	10515	10535	10400	10400
pmed38	11164	11101	11128	11072	11060

Fig. 4. Experimental design for 5-centered data-sets.



	500 Pop.	1000 Pop.	2000 Pop.	5000 Pop.	10000 Pop.	15000Pop.	20000Pop.
pmed2	4102	4093	4093	4093	4093	4093	4093
pmed3	4250	4250	4250	4250	4250	4250	4250
pmed7	5639	5631	5631	5631	5631	5631	5631
pmed12	6634	6634	6634	6634	6634	6634	6634
pmed17	7050	6999	6999	6999	6999	6999	6999
pmed22	8694	8699	8586	8579	8579	8579	8579
pmed27	8332	8322	8310	8307	8307	8307	8307
pmed32	9301	9315	9301	9297	9297	9297	9297
pmed36	9966	9976	9965	9960	9955	9953	9934
pmed39	9456	9485	9460	9499	9512	9466	9423

Fig. 5. Experimental design for 10-centered data-sets.

with a population between 500 and 1000. However, as seen from the previous application for 5 centered data-sets, the increase in the size of the data-set leads to an increase in the number of population to obtain an optimal result. Also, for Pmed36 and Pmed39, the optimal result is obtained with the same number of populations, which means 20,000 populations is sufficient for 10 centered data-sets to have an optimal result.

5. Results and discussion from application of proposed method on a case study: Nepal earthquake

In order to find important tweets and focus on the critical subjects, we determine the number of centers to be five (Alnajran et al., 2017; Liang, 2010). An existing Twitter data-set for the Nepal earthquake is utilized, consisting of 7000 tweets (Jin et al., 2010). The algorithm is run with a different number of populations and a different number of iterations. The point where the improvement stopped is recorded as the minimum result. The centers at this point are taken as the best centers. A proposed novel hybrid algorithm is run five times. The obtained results are demonstrated in Table 4.

As is seen in Table 4, the same objective values are obtained in the 1st, 3rd and 5th runs. However, centers can be varied for the same objective value. The reason for this is that tweets can fall into the same point in the distance matrix, or the locations of the centers can be close to each other, or there can be retweets in the document. Therefore, these different centers lead us to the same topic. Results in iterations until finding the minimum objective value are demonstrated in Table 5.

Table 5 indicates that genetic algorithm is constructed on best new solution which is starting point. In Fig. 6, the convergence of both bat algorithm and genetic algorithm is given separately. The consistency of

Table 4
Results of the Twitter data-sets for the Nepal earthquake.

(Population/ iteration)	Minimum found	Centers	Iteration number for minimum found
20000/15000	**1.941870877309351e + 02**	**[727 4165 1668 5825 5727]**	5084
20000/10000	1.942723444459188e + 02	[604 3610 39 5432 670]	1250
15000/20000	**1.941870877309351e + 02**	**[3548 6783 59,275,399 5825]**	11,441
20000/10000	1.943699098854494e + 02	[4556 5729 59,273,368 3610]	7019
20000/10000	**1.941870877309351e + 02**	**[4181 4214 5927 4441 4619]**	2623

the results in all runs indicates the optimal or near-optimal solution is obtained. In Fig. 7, comparisons of each run are demonstrated. Centers in the 3rd run and some of the closest nodes to these centers are demonstrated in Fig. 8 as an example.

Fig. 6 demonstrates the results for the first run. The algorithms, which are bat and genetic algorithms, are run separately and in the first stage, after the iteration number of 14111, the objective value of the bat algorithm is stabilized and found to be 1.96254e + 02. In the second stage, the genetic algorithm is started from the value of 1.96254e + 02, which is the best value of the bat algorithm, and it is stabilized after the iteration number of 11441. The optimum objective value is found to be 1.941871e + 02.

For 3 runs in HGBA, there is no improvement after the iteration number of 11441.

In Table 6, center tweets are shown.

In Table 7, the topic for center tweets is demonstrated.

50 topics are selected for the LDA model. Topic 37 is identified as the center topic. Furthermore, for the other runs (other center tweets), the center topic is the 37th topic again. According to the words in Table 7, the featured topic is labeled as “assistance” and “help” which are determined with respect to center tweets. Also, it is required to keep in mind that tweets from the Nepal earthquake cover the tweets a few months after the earthquake. Therefore, the tweets that are receive during or right after the earthquake can lead us to a topic that includes the needs of victims.

6. Conclusion

In this study, a novel HGBA is introduced to extract center tweets by utilizing Twitter data-sets for the Nepal earthquake. Text categorization is an important task for feature selection and feature extraction steps, to reduce the feature space dimension and remove irrelevant terms or topics in the classification process. Therefore, during the pre-processing stage, URLs, usernames, stop words, punctuation, and infrequent words are removed; tokenizing and stemming steps are applied to reduce the complexity of documents. After the pre-processing stage, the LDA method is applied to extract important topics from the Twitter data-sets. Therefore, document-topic probabilities are provided for the next stage. The JSD model is implemented to find the distribution matrix between each document, and the gathered distribution matrix is taken into consideration as a distance matrix. To evaluate extracted topics, tweets are needed to be clustered according to the distribution matrix. So, insignificant or comparatively insignificant tweets can be eliminated. Center tweets and tweets around center tweets are obtained utilizing HGBA through the application of UPMP. In the proposed study, the bat algorithm is embedded in the genetic algorithm to make a better initial matrix (starting point). Considerable improvement is achieved with this approach. Since the Continuous Bat Algorithm can not be directly

Table 5
Results from the iterations for 3 runs.

1st run		3rd run		5th run	
Best solution: 1.997408923117008e + 02	Best new solution (starting point) 1.962549419484962e + 02	Best solution: 1.995042060686592e + 02	Best new solution (starting point) 1.964502509326044e + 02	Best solution: 2.002224338545593e + 02	Best new solution (starting point) 1.969624417109948e + 02
Minimum found	Iteration number	Minimum found	Iteration number	Minimum found	Iteration number
1.957941e + 02	115	1.953584e + 02	11	1.967798e + 02	83
1.951836e + 02	139	1.948684e + 02	331	1.967454e + 02	85
1.949631e + 02	252	1.947654e + 02	575	1.966880e + 02	187
1.945818e + 02	820	1.946579e + 02	1406	1.959090e + 02	208
1.944659e + 02	913	1.943708e + 02	1665	1.958259e + 02	455
1.942723e + 02	1368	1.942723e + 02	2092	1.957603e + 02	499
1.942223e + 02	9189	**1.941871e + 02**	**2623**	1.956951e + 02	695
1.941871e + 02	**11,441**			1.955109e + 02	740
				1.953716e + 02	880
				1.948684e + 02	932
				1.943699e + 02	1245
				1.941871e + 02	**5084**

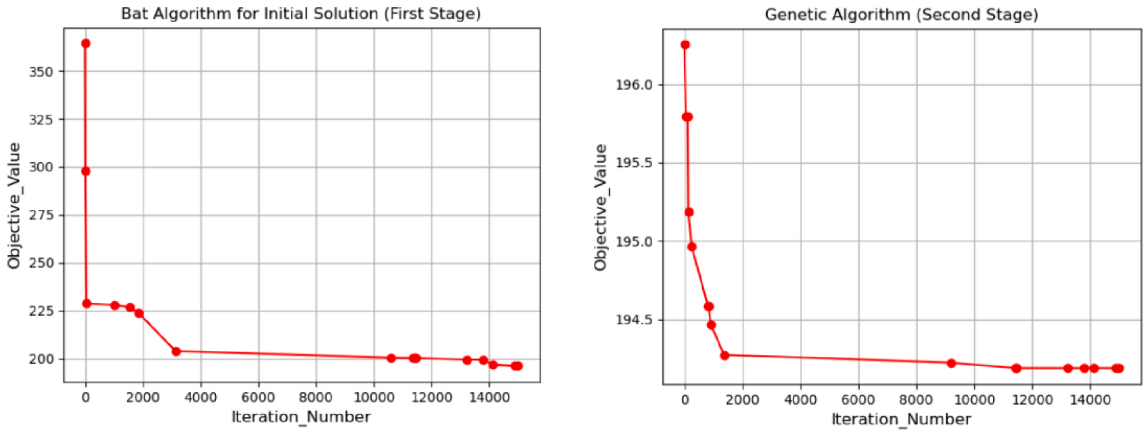


Fig. 6. The convergence of bat and genetic algorithm.

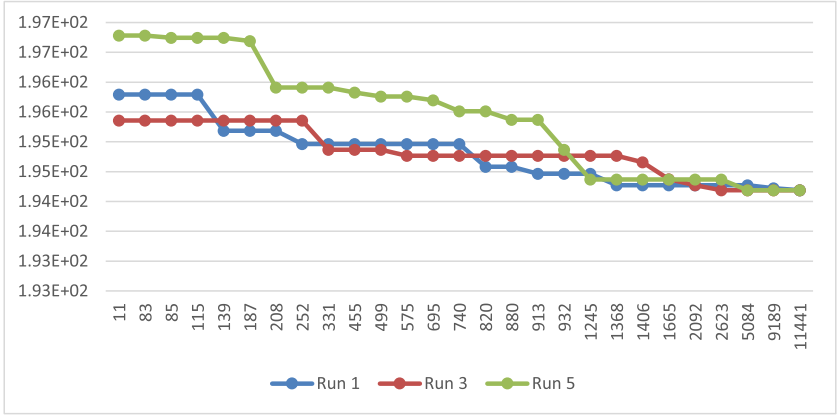


Fig. 7. Comparison of the runs.

utilized for PMP, some modifications are performed to make the algorithm discrete. Experiments are conducted on the P-median data-set with 5 and 10 centers from OR-Library. As a result of the experimental studies, it is seen that the proposed approach achieves optimal results. Finally, the proposed novel approach is tested on the Twitter data-set for the Nepal earthquake. For future work, locations of the center tweets and the tweets allocated to center tweets are going to be extracted with methods in the literature. After the location of extracted tweets that are connected to the centers, a network model is going to be designed to provide the right requirements in the right place.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

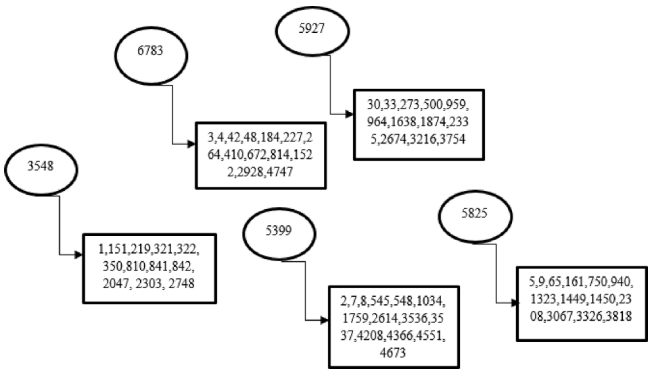


Fig. 8. Center tweets with closest tweets.

Table 6
Center tweets.

Center tweets
It's easy to get caught in the #MMA bubble, but the shit that's happening in Nepal and Baltimore is way more important
Wake up LA. There are over 1000 buildings that are not quake reinforced. Learn from the tragedy in Nepal. Spend now or pay in blood later.
Hope and strength. For all who were, and continue being affected by #NepalEarthquake
At least 3,617 people have been confirmed dead in #Nepal after massive #earthquake police say -
Earthquake in Nepal is one of the worst human tragedies i have come across in my life. it will take a lot to reconstruct nepal

Table 7
Center topic.

Center topic	
Word	score
Help	0,176510604526418
Donat	0,0875764565291334
Need	0,0848609100253996
Know	0,0651731978733290
People	0,0590632182399277

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert Systems with Applications, 36*(3 part 2), 6843–6853. https://doi.org/10.1016/j.eswa.2008.08.022

Alnajran, N., Crockett, K., McLean, D., & Latham, A. (2017). Cluster analysis of twitter data: A review of algorithms. *ICAART 2017 - Proceedings of the 9th International Conference on Agents and Artificial Intelligence*, *2*(Icaart), 239–249. 10.5220/0006202802390249.

Alsmadi, I., & Gan, K. H. (2019). Review of short-text classification. *International Journal of Web Information Systems, 15*(2), 155–182. https://doi.org/10.1108/IJWIS-12-2017-0083

Beasley, J. (1985). A note on solving large p-median problems. *European Journal of Operational Research, 21*(2), 270–273.

Benitez, I. P., Sison, A. M., & Medina, R. P. (2018). An improved genetic algorithm for feature selection in the classification of Disaster-related Twitter messages. In *ISCAIE 2018–2018 IEEE Symposium on Computer Applications and Industrial Electronics* (pp. 238–243). https://doi.org/10.1109/ISCAIE.2018.8405477

Blei, D., Jordan, M., & Ng, A. Y. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993

Chansombat, S., Musikapun, P., Pongcharoen, P., & Hicks, C. (2018). A hybrid discrete bat algorithm with krill herd-based advanced planning and scheduling tool for the capital goods industry. *International Journal of Production Research, 7543*(May), 1–22. https://doi.org/10.1080/00207543.2018.1471240

Chen, L. F., Su, C. T., & Chen, K. H. (2012). An improved particle swarm optimization for feature selection. *Intelligent Data Analysis, 16*(2), 167–182. https://doi.org/10.3233/IDA-2012-0517

Chen, L., Li, J., & Zhang, L. (2017). A method of text categorization based on genetic algorithm and LDA. *Chinese Control Conference, CCC*, 10866–10870. 10.23919/ChiCC.2017.8029089.

Chiang, H. S., Shih, D. H., Lin, B., & Shih, M. H. (2014). An APN model for Arrhythmic beat classification. *Bioinformatics, 30*(12), 1739–1746. https://doi.org/10.1093/bioinformatics/btu101

Chuang, L. Y., Tsai, S. W., & Yang, C. H. (2011). Improved binary particle swarm optimization using catfish effect for feature selection. *Expert Systems with Applications, 38*(10), 12699–12707. https://doi.org/10.1016/j.eswa.2011.04.057

Correa, E. S., Steiner, M. T. A., Freitas, A. A., & Carnieri, C. (2004). A genetic algorithm for solving a capacitated p-median problem. *Numerical Algorithms, 35*(2–4), 373–388. https://doi.org/10.1023/B:NUMA.0000021767.42899.31

Eligüzel, İ. M., & Özceylan, E. (2021). Application of an improved discrete crow search algorithm with local search and elitism on a humanitarian relief case. *Artificial Intelligence Review, 54*(6), 4591–4617. https://doi.org/10.1007/s10462-021-10006-2

Eligüzel, N., Çetinkaya, C., & Dereli, T. (2021). A state-of-art optimization method for analyzing the tweets of earthquake-prone region. *Neural Computing and Applications, 33*(21), 14687–14705. https://doi.org/10.1007/s00521-021-06109-0

Fatyanosa, T. N., Bachtiar, F. A., & Data, M. (2019). Feature Selection using Variable Length Chromosome Genetic Algorithm for Sentiment Analysis. In *3rd International Conference on Sustainable Information Engineering and Technology*. https://doi.org/10.1109/SIET.2018.8693190

Alam, F., & Shafiq Joty, M. I. (2018). Domain Adaptation with Adversarial Training and Graph Embeddings. In *Accepted for Publication at the 56th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1077–1087).

Fuglede, B., & Topsøe, F. (2004). Jensen-Shannon divergence and Hubert space embedding. *IEEE International Symposium on Information Theory -*.

Galavotti, L., Sebastiani, F., & Simi, M. (2007). *Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization*. 59–68. 10.1007/3-540-45268-0_6.

Ghareb, A. S., Bakar, A. A., & Hamdan, A. R. (2016). Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications, 49*, 31–47. https://doi.org/10.1016/j.eswa.2015.12.004

Harrag, F., Al-Salman, A. M. S., & Mohammed, M. B. (2010). A comparative study of neural networks architectures on Arabic text categorization using feature extraction. In *2010 International Conference on Machine and Web Intelligence*. https://doi.org/10.1109/ICMWI.2010.5648051

Jin, Y., Xiong, W., & Wang, C. (2010). Feature selection for Chinese text categorization based on improved particle swarm optimization. In *Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering*. https://doi.org/10.1109/NLPKE.2010.5587844

Joseph Manoj, R., Anto Praveena, M. D., & Vijayakumar, K. (2018). An ACO–ANN based feature selection algorithm for big data. *Cluster Computing, 0123456789*, 1–8. https://doi.org/10.1007/s10586-018-2550-z

Kotcz, A. (2001). *Summarization as Feature Selection for Text*. 1–6. papers2://publication/uuid/A9F66FC6-2B71-4345-8C8F-98D13FB70055.

Li, X. F., Zhao, L. L., & Wu, L. H. (2008). A feature extraction method using base phrase and keyword in Chinese text. *Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering, ISKE 2008*, *1*, 680–684. 10.1109/ISKE.2008.4731016.

Liang, C. (2010). An ant colony algorithm for text clustering. *2010 International Conference on Computing, Control and Industrial Engineering, CCIE 2010*, *2*, 249–252. 10.1109/CCIE.2010.180.

Liu, H., Su, Z., Yao, Z., & Zhang, X. (2010). A method of text feature extraction based on weighted scatter difference. *Proceedings - 2010 2nd WRI Global Congress on Intelligent Systems, GCIS 2010*, *3*, 83–86. 10.1109/GCIS.2010.49.

MathWorks. (2019).

Moens, M. F., & Dumortier, J. (2000). Text categorization: The assignment of subject descriptors to magazine articles. *Information Processing and Management, 36*(6), 841–861. https://doi.org/10.1016/S0306-4573(00)00012-1

Oksuz, M., Satoglu, S., & Kayakutlu, G. (2016). A Genetic Algorithm for the P-Median Facility Location Problem. *Researchgate.Net*, *September*. https://www.researchgate.net/profile/Sule_Satoglu/publication/305380696_A_Genetic_Algorithm_for_the_p-Median_Facility_Location_Problem/links/57ed51a808ae03fa0e82946d/A-Genetic-Algorithm-for-the-p-Median-Facility-Location-Problem.pdf.

Osaba, E., Yang, X., Diaz, F., Lopez-garcia, P., & Carballedo, R. (2016). *An Improved Discrete Bat Algorithm for Symmetric and Asymmetric Traveling Salesman Problems arXiv : 1604 . 04138v1 [cs . NE] 14 Apr 2016*. *1985*, 1–28.

Osaba, E., Yang, X. S., Diaz, F., Lopez-Garcia, P., & Carballedo, R. (2016). An improved discrete bat algorithm for symmetric and asymmetric Traveling Salesman Problems. *Engineering Applications of Artificial Intelligence, 48*(1985), 59–71. https://doi.org/10.1016/j.engappai.2015.10.006

Osaba, E., Yang, X. S., Fister, I., Del Ser, J., Lopez-Garcia, P., & Vazquez-Pardavila, A. J. (2019). A Discrete and Improved Bat Algorithm for solving a medical goods distribution problem with pharmacological waste collection. *Swarm and Evolutionary Computation*, *44*(March 2018), 273–286. 10.1016/j.swevo.2018.04.001.

Panichella, A., Dit, B., Oliveto, R., Di Penta, M., Poshynanyk, D., & De Lucia, A. (2013). How to effectively use topic models for software engineering tasks? An approach

based on Genetic Algorithms. *Proceedings - International Conference on Software Engineering, 522–531*. https://doi.org/10.1109/ICSE.2013.6606598

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137. https://doi.org/10.1108/eb046814

Precup, R. E., David, R. C., Roman, R. C., Szedlak-Stinean, A. I., & Petriu, E. M. (2021). Optimal tuning of interval type-2 fuzzy controllers for nonlinear servo systems using Slime Mould Algorithm. *International Journal of Systems Science*. https://doi.org/10.1080/00207721.2021.1927236

Preitl, Z., Precup, R. E., Tar, J. K., & Takács, M. (2006). Use of multi-parametric quadratic programming in fuzzy control systems. *Acta Polytechnica Hungarica, 3*(3), 29–43.

Sebastiani, F. (2003). Text Categorization. *Encyclopedia of Database Systems*, *October 2003*, 0–5. 10.1007/978-0-387-39940-9.

Sotiropoulos, D. N., Kounavis, C. D., & Giaglis, G. M. (2014). *Semantically meaningful group detection within sub-communities of Twitter blogosphere*. *August*, 734–738. 10.1145/2492517.2492613.

Sotiropoulos, D. N., Pournarakis, D. E., & Giaglis, G. M. (2016). A genetic algorithm approach for topic clustering: A centroid-based encoding scheme. In *IISA 2016–7th International Conference on Information, Intelligence, Systems and Applications* (pp. 1–8). https://doi.org/10.1109/IISA.2016.7785378

Soucy, P., & Mineau, G. W. (2002). *A simple KNN algorithm for text categorization*. 647–648. 10.1109/icdm.2001.989592.

Dutta, S., Ghatak, S., Saptarshi Ghosh, A. K., & Das.. (2017). A Genetic Algorithm based tweet clustering Technique. *2017 International Conference on Computer Communication and Informatics (ICCCI*.

Taira, H., & Haruno, M. (1999). Feature Selection in SVM Text Categorization. *Proceedings of AAAI99 16th Conference of the American Association for Artificial Intelligence*, *41*, 480–486. http://www.springerlink.com/index/9rkk15dfy3rrcx41.pdf.

Techopedia. (2019).

Teixeira, J. C., & Antunes, A. P. (2008). A hierarchical location model for public facility planning. *European Journal of Operational Research, 185*(1), 92–104. https://doi.org/10.1016/j.ejor.2006.12.027

Tong, Z., & Zhang, H. (2016). *A Text Mining Research Based on LDA Topic Modelling*. 201–210. 10.5121/csit.2016.60616.

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*(7), 1024–1032. https://doi.org/10.1016/j.knosys.2011.04.014

Wang, Y., Feng, L., & Zhu, J. (2018). Novel artificial bee colony based feature selection method for filtering redundant information. *Applied Intelligence, 48*(4), 868–885. https://doi.org/10.1007/s10489-017-1010-4

Xue, B., Zhang, M., Member, S., & Browne, W. N. (2012). Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Transactions on Cybernetics*, 1–16.

Yang, Y. M., & J. o. p.. (1997). A comperative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning, 2*, 412–420.

Yang, X.-S. (2010). A New Metaheuristic Bat-Inspired Algorithm. In *In Nature inspired cooperative strategies for optimization (NICSO 2010)* (Issue April 2010, pp. 65–74). Springer. 10.4018/978-1-59904-885-7.ch129.

Yusta, S. C. (2009). Different metaheuristic strategies to solve the feature selection problem. *Pattern Recognition Letters, 30*(5), 525–534. https://doi.org/10.1016/j.patrec.2008.11.012

Zahran, B. M., & Kanaan, G. (2009). Text Feature Selection using Particle Swarm Optimization Algorithm. *World Applied Sciences JournalSpecial Issue of Computer & IT, 7*, 69–74.

Zhang, W., Yoshida, T., & Tang, X. (2008). Text classification based on multi-word with support vector machine. *Knowledge-Based Systems, 21*(8), 879–886. https://doi.org/10.1016/j.knosys.2008.03.044