# Unsupervised Text Feature Selection Technique Based on Particle Swarm Optimization Algorithm for Improving the Text Clustering

Laith Mohammad Abualigah[1], Ahamad Tajudin Khader[1], Mohammed Azmi Al-Betar[2], and Essam Said Hanandeh[3]

[1] School of Computer Sciences, Universiti Sains Malaysia (USM), Pulau Pinang, Malaysia 11800
[2] Department of information technology, Al-Huson University College, Al-Balqa Applied University, Al-Huson, Irbid-Jordan
[3] Department of Computer Information System, Zarqa University, Zarqa-Jordan

**Abstract.** After incensing the amount of text information on internet web pages, the dealing with this information is very complex due to the volume of information. Text clustering technique is an appropriate task to deal with a huge amount of text documents by grouping set of documents into groups. Text documents contain uninformative features, which decrease the performance of the text clustering technique. Feature selection is an unsupervised technique used to select informative features by creating a new subset of informative features. This technique used to improve the performance of the underlying algorithm. Latterly, several complex optimization problems are success solved by meta-heuristic al-gorithms. In this paper, we proposed the Particle swarm optimization algorithm to solve the feature selection problem, namely, (FSPSOTC). The feature selection technique encourages the k-mean text clustering technique to obtain more accurate clusters. Experiments were conducted using four standard benchmark text datasets with different characteris-tics. Experimental results showed that the proposed method (FSPSOTC) is enhanced the performance of the text clustering technique by dealing with a new subset of informative features.

**Key words:** unsupervised feature selection, informative features, par-ticle swarm optimization algorithm, K-mean text clustering technique

## 1 Introduction

In the most recent years, the growth amount of digital text on internet web page and modern applications it affects the text analysis process. The text clus-tering is a suitable technique used to grouping or clustering a huge set of text documents into a predetermined number of groups [2]. This technique is impor-tant and used in many domains in the area of the text mining as text retrieval, text categorization, and image segmentation [3]. Vector Space Model (VSM) is a common popular model used in are of the text mining to represent each doc-ument features as a vector (row) of weight. In this model, each term weight

is represented as one dimension space. Thus, the performance of the clustering technique effects by the size of the dimension space and uninformative features
[4].

Text documents contain informative and uninformative features, where an uninformative feature are noisy, irrelevant, and redundant features [17]. Unsu-pervised feature selection is the main task used to find a new optimal subset of informative features for each document. This technique is used to enhance the k-mean clustering technique without any foreknowledge of the document class label. The feature selection technique gives the accurate results when defined as an optimization problem, with two objectives, (1) maximize the performance of the text clustering algorithm, (2) minimizing the number of uninformative fea-tures [4]. In general, several domains in the text mining area benefits the feature selection technique such as the text clustering [2] and text retrieval [7].

Text clustering is an effective unsupervised learning technique used to parti-tion a set of digital text documents into a subset of clusters to make the access tidy and easier for the users [24]. Text clustering algorithms seek to find an op-timal solution to partition a set of documents. The algorithm acts done based on some evaluation criteria such as objective function and fitness function. Re-cently, the internet web pages have become a major source for the humanity to find the information that they need. Also, an unorganized a huge amount of text documents, which are used in universities, hospitals, and digital libraries datasets [5].

Particle swarm optimization (PSO) algorithm was introduced by Kennedy and Eberhart in (1995) [1]. It is inspired by a swarm intelligence based meta-heuristic search and optimization method. PSO algorithm mimics the social behavior of the birds flocking and fish schooling and it is used the global best solution for achieving the optimal solution. In each iteration, the global best solution of the PSO is considered the near optimal solution to solve the feature selection problem so far [8].

Several models are proposed to enhance the performance of text feature se-lection based using meta-hubristic algorithm [17, 19, 15, 4, 18, 20, 21]. The authors applied three models [15]. The first model has used the original PSO algorithm, the second model has improved the PSO algorithm by the inertia weight to optimize the feature selection model, and the third model has added a new function to the original PSO algorithm. Experimental results showed that the second PSO model is the best model for improving the performance of text feature selection.

A new technique using PSO with an opposition-based mechanism is applied for text clustering technique. It is to begin using a set of promising and varied solutions to achieve an optimal solution and a new dynamic into weight [4]. The authors investigated the proposed method using three text subset datasets. The experimental results proved the effectiveness of the selected features, increased the clustering accuracy and reduced the computation time.

In this paper, we proposed a new feature selection technique using particle swarm optimization algorithm, namely, FSPSOTC. This method is used for se-

lecting a new optimal subset of informative feature to improve the performance of the text clustering technique. The main aim of this paper is to propose a new feature selection method for enhancing the performance of the text clustering algorithm by eliminating uninformative features.

The rest of this paper is organized as follows: section 2 shows the term weight-ing. Section 3 illustrate text feature selection technique using the particle swarm optimization algorithm and its procedure. Section 4 show steps of the k-mean text clustering algorithm. Section 5 show the text clustering evaluations mea-surements. Experimental results of the proposed method are presented in Section
6. Finally, the conclusion is provided in Section 7.

## 2 Feature selection using particle swarm optimization algorithm

This section explains the proposed feature selection method based on the particle swarm optimization algorithm.

### 2.1 Mathematical model of the feature selection problem

The feature selection problem is formulated as an optimization problem to find an optimal subset of informative text features.

Given f a set of document features $f_i = \{f_{i1}, f_{i2}, ..., f_{it}\}$, t is the number of all unique features, i is the document number. let $sf_i = \{s_{i1}, s_{i2}, ..., s_{ij}, s_{im}\}$ is a new subset of features, m is a new dimension space $s_{ij} \in \{0, 1\}$, $j = 1, 2, ..., m$. If $s_{ij} = 1$ mean the $j_{th}$ feature is selected as useful feature in document i, if $s_{ij} = 0$ mean the $j^{th}$ feature is hidden or not useful in document $i$ [14, 2, 20, 21].

### 2.2 Solution representation

We apply the feature selection technique based on PSO algorithm, which be-gins with random initial solutions and improves the population to reach the global optimal solution [4, 22], which represent a new subset of features. Each unique feature in the given dataset considers as a dimension search space. Table 1 presents the solution representation of the feature selection technique.

**Table 1.** Solution representation of the feature selection technique

| X | *0* | *1* | *1* | *-1* | *-1* | *1* | *0* | *-1* | *1* | *-1* |
|---|---|---|---|---|---|---|---|---|---|---|

### 2.3 Fitness function

The fitness function (FF) is a type of objective function used to evaluate each solution. Each iteration has calculated the fitness function of each solution to decide if there is an improvement on the solutions to accept it or decline. Finally, the solution, which has a high fitness value is the optimal solution [10]. We used mean absolute difference (MAD) as a fitness function in PSO algorithm for feature selection technique using the weighting scheme (TF-IDF) as the objective function for evaluating the solution positions as the following[9]:

$$MAD_{(Xi)} = \frac{1}{a_i} \sum_{j=1}^{t} |x_{i,j} - \bar{x_i}|, \qquad \bar{x_i} = \left(\frac{1}{a}\right) \sum_{j=1}^{t} x_{ij}, \tag{1}$$

$MAD_{(Xi)}$ represents the fitness function of the solution i, $x_{i,j}$ is the value of the feature j in document i, $a_i$ is the number of selected features in document i, t is the number of unique features, and $x^{-}{}_i$ is the mean value of the vector i.

## 3 Particle swarm optimization algorithm for feature selection

PSO used the global best solution for achieving the optimal solution, also in each iteration, the global best solution is the best solution.

### 3.1 Particle swarm optimization algorithm search space

Meta-heuristic algorithm generated the population with random positions ac-cording to the upper and lower bounds, each candidate solution called particle. And it pursues to enhance the swarm for achieving an optimal global best solu-tion. Each position is a dimension of the search space. The solutions, evaluated by the cost function as Eq. 1 to obtain the global best solution. The PSO algo-rithm includes a store of solutions, which filled by generating $S$ random solutions as in the matrix 2.

$$\mathbf{PSOM} = \begin{bmatrix} x_1^1 & \cdots & x_t^1 & f(\boldsymbol{X}_1) \\ \vdots & \ddots & \vdots & \vdots \\ x_1^{S-1} & \cdots & x_t^{S-1} & f(\boldsymbol{X}_{S-1}) \\ x_1^S & \cdots & x_t^S & f(\boldsymbol{X}_S) \end{bmatrix} \tag{2}$$

PSO algorithm works by using two main features to update each particle position: velocity and particle positions based on two equation 4 and 3. The velocity of each particle updated according to the particle movement effect, each

particle trying to move to an optimal position based on the following equation [11]:

$$x_{ij} = x_{ij} + V_{ij} \tag{3}$$

where,

$$V_{ij} = w * V_{ij} + c_1 * r_1 * (Lb_{ij} - x_{ij}) + c_2 * r_2 * (Gb_{ij} - x_{ij}) \tag{4}$$

The inertia weight is determined using the following linear equation:

$$w = w_{max} - w_{min} * (\frac{I_{max} - I}{I_{max}}) + w_{min} \tag{5}$$

$w_{max}$ represents the maximum inertia weight 0.9, $w_{min}$ is the minimum iner-tia weight 0.4, so often the value of inertia weight changing based on iteration in the range of $(0.1, 0.9)$. Lb is the best particle's personal experience during each iteration, and Gb is the global best from the global best solution, $r_1$ and $r_2$ are random number in range $(0, 1)$, $c_1$ and $c_2$ are constant values, it usually taken (1.49) based on the literature of the feature selection technique.

## 4 Text clustering technique

### 4.1 Mathematical model of the text clustering problem

The text clustering technique is defined as: given $D$ a set of text documents $D = d_1$, $d_2$, ..., $d_j$ , ..., $d_n$, where, n represents the number of all documents in the given dataset, $d_1$ represents the document number 1, $Cos(d_i)$ is an objective function to maximize the similarity measure of the document $d_i$ [5, 12].

### 4.2 Compute clusters centroid

In order to partition a set of text documents into a subset of clusters, each cluster has one centroid, which needs updating in each iteration using Eq. 6. Imposed, each document assigns to the similar cluster based on the similarity with the cluster centroid. Where, $C_k$ are clusters centroid of $k$ clusters $C_k = (c_{k1}, c_{k2}, ...., c_{kj} , ..., c_{kK})$, $c_{kj}$ is the centroid of cluster j [5, 23]. The following equation is used to calculate clusters centroid:

$$c_{kj} = \frac{\sum_{i=1}^{n} (a_{ki}) d_i}{\sum_{j=1}^{r_i} a_{kj}}, \tag{6}$$

.

$d_i$ is the document $i$ that belongs to $c_j$ centroid of the cluster $j$, $a_{kj}$ is the number of documents that belong to cluster j, $r_i$ is the number of documents in cluster i [5].

### 4.3 K-mean algorithm

The K-mean algorithm is introduced in (1967) as local search technique [13]. It is a common and suitable technique used in the domain of the text clustering. K-mean clustering technique is considered a proper algorithm to choose initial clusters centroid [5, 23]. These procedures are reviewed in the algorithm Pseudo code 1 as the following:

**Algorithm 1** K-mean clustering algorithm

1: **Input**: $D$ is a collection of text documents , $K$ is the number of clusters. 2:
**Output**: Assign $D$ to $K$.
3: **Termination criteria**
4: Randomly choosing $K$ document as clusters centroid $C = (c_1, c_2, ...., c_K)$ 5: Initialize matrix X as zeros
6: **for** all $d$ in $D$ **do**
7: let $j = arg_{min}\ k$ based on $Cos(d_i, c_k)$
8: **end for**
9: Update clusters centroid using equation 6
10: **End**

## 5 Evaluation measures

### 5.1 Precision, Recall and F-measure

F-measure (F): is a common measurement used on the text clustering domain. It works to measure the percentage of the matched clusters which depends on two measurements: Precision (P) and Recall (R). The Precision and Recall are common measurements used in area of text mining and used in order to calculate the F-measure value for cluster $j$ and class $i$ that is will by combined [9].

$$P(i,j) = \frac{n_{i,j}}{n_j}, \qquad R(i,j) = \frac{n_{i,j}}{n_i} \tag{7}$$

$n_{ij}$ is the number of members of class i in cluster j, $n_j$ is the number of members of cluster j and $n_i$ is the number of members of class i.

$$F(i,j) = \frac{2P(i,j)R(i,j)}{P(i,j)R(i,j)}, \qquad F = \sum_j \frac{n_j}{n} \max_i \{n(i,j)\} \tag{8}$$

P(i,j) is the precision of members of class i in cluster j, R(i,j) is the recall of members of class i in cluster j and F-measure for all clusters calculate by the following equation:

### 5.2 Accuracy

The accuracy (AC) measurement is one of the common external measurements that used precisely to compute the percentage of correct assigned documents to each cluster according to the following equation [20, 21]:

$$Ac = \sum_{i=1}^{k} \frac{1}{n} P(i,j) \tag{9}$$

Where, $P(i,j)$ is the precision value for class i in cluster j, n is the number of all documents in each cluster, k is the number of all clusters.

## 6 Experimental results

We have programmed PSO algorithm for feature selection technique, then the k-mean algorithm for text clustering technique using Matlab (version 7.10.0) software. This section provides the details of the given datasets, the evaluation criteria, and experiments results and discussion. Table 1 shows four standard text datasets that used to investigate the performance of the proposed method (FSPSOTC) and compare with other well-known in the domain of the feature selection technique. Text clustering benchmark standard datasets are available at `http://sites.labic.icmc.usp.br/text_collections/` by numerical form after the terms extraction.

**Table 2.** Text document datasets characteristics

| Datasets | Resources | # of Documents | # of Terms | # of Clusters |
|---|---|---|---|---|
| DS1 | Reuters−21578 | 200 | 2935 | 4 |
| DS2 | 20Newsgroups | 100 | 3263 | 5 |
| DS3 | Reuters−21578 | 100 | 2063 | 8 |
| DS4 | 20Newsgroups | 200 | 5773 | 10 |

### 6.1 Results and discussions

We applied the k-mean text clustering algorithm to investigate the influence of the feature selection technique on the text clustering technique. On the other hand, the feature selection technique is used to improve the performance of text clustering algorithm by using a new subset which contains the informative fea-tures. The proposed FSPSO is compared with other well-known optimization algorithms in the domain of the feature selection technique. Overall, it is used to obtain efficient text clusters. The experimental was tested using four benchmark datasets shown in Table 2. In order to make fair comparisons, we applied the experiments over 20 runs. This number is selected based on the previous stud-ies in text clustering domain which sufficient to evaluate the proposed method.

**Table 3.** Algorithm performance based on clusters quality

| Dataset | Method | text clustering | FSHSTC | FSGATC | FSPSOTC |
|---|---|---|---|---|---|
| DS1 | Dimention | 2935 | 738 | 805 | 1985 |
| | Accuracy | 0.5565 | 0.5365 | **0.5955** | 0.5845 |
| | Precision | 0.5201 | 0.5274 | 0.5690 | **0.5754** |
| | Recall | 0.5077 | 0.5046 | **0.5681** | 0.5518 |
| | F-measure | 0.5244 | 0.5011 | 0.5679 | **0.5690** |
| | Best values | 0 | 0 | 2 | 2 |
| | Rank | 4 | 3 | **1** | **1** |
| DS2 | Dimention | 3263 | 328 | 382 | 1930 |
| | Accuracy | 0.3520 | 0.3595 | **0.4070** | 0.4040 |
| | Precision | 0.2852 | 0.3166 | 0.3346 | **0.3551** |
| | Recall | 0.2718 | 0.3159 | 0.3446 | **0.3595** |
| | F-measure | 0.3057 | 0.3150 | 0.3386 | **0.3559** |
| | Best values | 0 | 0 | 1 | 3 |
| | Rank | 4 | 3 | 2 | **1** |
| DS3 | Dimention | 2063 | 469 | 547 | 1316 |
| | Accuracy | 0.5070 | 0.5025 | 0.4705 | **0.5170** |
| | Precision | 0.4721 | 0.4611 | 0.4262 | **0.4768** |
| | Recall | 0.4709 | 0.4644 | 0.4261 | **0.4758** |
| | F-measure | 0.4751 | 0.4610 | 0.4262 | **0.4844** |
| | Best values | 0 | 0 | 0 | 4 |
| | Rank | 2 | 3 | 4 | **1** |
| DS4 | Dimention | 5773 | 779 | 869 | 3616 |
| | Accuracy | 0.2707 | 0.2692 | 0.2762 | **0.2862** |
| | Precision | 0.2422 | 0.2502 | 0.2578 | **0.2581** |
| | Recall | 0.2514 | 0.2499 | 0.2479 | **0.2626** |
| | F-measure | 0.2349 | 0.2491 | 0.2526 | **0.2707** |
| | Best values | 0 | 0 | 0 | 4 |
| | Rank | 3 | 3 | 2 | **1** |
| **Mean rank** | | 3.25 | 3.00 | 2.25 | **1.00** |
| **Final rank** | | 4 | 3 | 2 | **1** |

The PSO is a global search algorithm, runs 500 iterations in each run. Exper-imentally noted that 500 iterations are enough for the convergence of global search algorithm. The k-mean is a local search algorithm, runs 100 iterations in each run. Experimentally noted that 100 iterations are enough for reaching the convergence of local search clustering algorithm [5].

Table 3 shows that the performance of text clustering based on the feature se-lection technique. The proposed feature selection technique using PSO algorithm improved the text clustering performance almost overall given datasets based on evaluation measurements of the text clustering technique. Our proposed method is used to select an optimal subset of informative text features for improving the text clustering by finding the optimal clusters. The proposed FSPSOTC performs very well to improve the text clustering technique, and it reduced the number of features. The proposed FSPSOTC overcomes the others comparative methods to deal with a huge collection of text documents with multi-dimension

feature space, sparse features. At the end, the FSPSOTC got the best method overall method based on the final ranking with rank 1. The FSGATC got the second best method based on the final ranking with rank 2. The FSHSTC got the third best method based on the final rank with rank 3. But, the pure k-mean clustering without feature selection technique got the worst method based on the final rank with rank 4.

## 7 Conclusion

In this paper, feature selection technique using PSO algorithm to find a new op-timal subset of informative text features is proposed and compared with the well-known algorithm in the domain of the feature selection. The proposed method is called feature selection technique using particle swarm optimization algorithm for the text clustering technique (FSPSOTC). It overwhelms all other compar-ative algorithms (ex. harmony search and genetic algorithm) by enhancing the performance of text clustering results by obtaining more accurate clusters. FSP-SOTC is evaluated using four text datasets. The results showed that the per-formance text clustering technique is improved by using the proposed method "FSPSOTC" and in term of time it reduce the computational time. For future work, the proposed FSPSOTC can be modified to improve the global search in hope to obtain more accurate clusters.

## References

1. Eberhart, Russ C., and James Kennedy. "A new optimizer using particle swarm theory." Proceedings of the sixth international symposium on micro machine and human science. Vol. 1. 1995.
2. Bharti, Kusum Kumari, and Pramod Kumar Singh. "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering." Expert Systems with Applications 42.6 (2015): 3105-3114.
3. Wang, Xingheng, et al. Text clustering based on the improved TFIDF by the it-erative algorithm. Electrical and Electronics Engineering (EEESYM), 2012 IEEE Symposium on. IEEE, 2012.
4. Bharti, Kusum Kumari, and Pramod Kumar Singh. "Opposition chaotic fitness mu-tation based adaptive inertia weight BPSO for feature selection in text clustering." Applied Soft Computing 43 (2016): 20-34.
5. Forsati, Rana, et al. "Efficient stochastic algorithms for document clustering." In-formation Sciences 220 (2013): 269-291.
6. Geem, Zong Woo, Joong Hoon Kim, and G. V. Loganathan. "A new heuristic op-timization algorithm: harmony search." Simulation 76.2 (2001): 60-68.
7. Abualigah, Laith Mohammad Qasim, and Essam S. Hanandeh. "APPLYING GE-NETIC ALGORITHMS TO INFORMATION RETRIEVAL USING VECTOR SPACE MODEL." International Journal of Computer Science, Engineering and Applications 5.1 (2015): 19.

8. Jin, Yaohong, Wen Xiong, and Cong Wang. "Feature selection for Chinese text categorization based on improved particle swarm optimization." Natural Language Processing and Knowledge Engineering (NLP-KE), 2010 International Conference on. IEEE, 2010.
9. Bharti, Kusum Kumari, and Pramod Kumar Singh. "A three-stage unsupervised dimension reduction method for text clustering." Journal of Computational Science 5.2 (2014): 156-169.
10. Uuz, Harun. "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm." Knowledge-Based Systems 24.7 (2011): 1024-1032.
11. Ramos, Caio CO, et al. "A novel algorithm for feature selection using harmony search and its application for non-technical losses detection." Computers and Elec-trical Engineering 37.6 (2011): 886-894.
12. Zhao, Zheng, et al. "On similarity preserving feature selection." Knowledge and Data Engineering, IEEE Transactions on 25.3 (2013): 619-632.
13. MacQueen, James. "Some methods for classification and analysis of multivariate observations." Proceedings of the fifth Berkeley symposium on mathematical statis-tics and probability. Vol. 1. No. 14. 1967.
14. Zhao, Wei, and Yafei Wang. "Notice of Retraction An Improved Genetic Algorithm for Text Feature Selection." Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference on. IEEE, 2010.
15. Lu, Yonghe, et al. "Improved particle swarm optimization algorithm and its application in text feature selection." Applied Soft Computing 35 (2015): 629-636.
16. Diao, Ren. Feature selection with harmony search and its applications. Diss. Aberystwyth University, 2014.
17. Lin, Kuan-Cheng, et al. "Feature selection based on an improved cat swarm optimization algorithm for big data classification." The Journal of Supercomputing (2016): 1-12.
18. Shamsinejadbabki, Pirooz, and Mohammad Saraee. "A new unsupervised feature selection method for text clustering based on genetic algorithms." Journal of Intel-ligent Information Systems 38.3 (2012): 669-684.
19. Hong, Sung-Sam, Wanhee Lee, and Myung-Mook Han. "The Feature Selection Method based on Genetic Algorithm for Efficient of Text Clustering and Text Classification." International Journal of Advances in Soft Computing and Its Applications 7.1 (2015).
20. Abualigah, L. M., Khader, A. T., AI-Betar, M. A., AI-Huson, I. J. Unsupervised Feature Selection Technique Based on Genetic Algorithm for Improving the Text Clustering.
21. Abualigah, L. M., Khader, A. T., AI-Betar, M. A., AI-Huson, I. J. Unsupervised Feature Selection Technique Based on Harmony Search Algorithm for Improving the Text Clustering.
22. Bolaji, A. L. A., Al-Betar, M. A., Awadallah, M. A., Khader, A. T., Abualigah, L. M. (2016). A comprehensive review: Krill Herd algorithm (KH) and its applications. Applied Soft Computing.
23. Abualigah, L. M., Khader, A. T., AI-Betar, M. A., AI-Huson, I. J. Multi-objectives-based text clustering technique using K-mean algorithm.
24. Abualigah, L. M., Khader, A. T., Al-Betar, M. A., Awadallah, M. A. (2016, May). A krill herd algorithm for efficient text documents clustering. In Computer Applica-tions and Industrial Electronics (ISCAIE), 2016 IEEE Symposium on (pp. 67-72). IEEE.