International Journal of Engineering
Journal Homepage: www.ije.ir

A Two-Level Semi-supervised Clustering Technique for News Articles

S. M. Sadjadi, H. Mashayekhi*[1], H. Hassanpour

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

PAPER INFO

Paper history:
Received 26 July 2021
Received in revised form 11 September 2021
Accepted 20 September 2021

Keywords:
News Clustering
Two-level Clustering
Semi-supervised
Word Embedding
Document Clustering

ABSTRACT

The web and social media are overcrowded with news pieces in terms of amount and diversity. Document clustering is a useful technique that is widely used in organizing and managing data into smaller groups. One of the factors influencing the quality of clustering is the way documents are represented. Some traditional methods of document representation depend on word frequencies and create sparse and large-sized document vectors. These methods cannot preserve proximity information between documents. In addition, neural network-based methods that preserve proximity information suffer from poor interpretability. Conceptual text representation methods have overcome the shortcomings of previous methods, but semi-supervised text clustering does not currently use concept-based document representation. This paper presents a two-level semi-supervised text clustering method that uses labeled and unlabeled data simultaneously to achieve higher clustering quality. In the first level, documents are represented based on the concepts extracted from the raw corpus. Second, the semi-supervised clustering process applies unlabeled data to capture the overall structure of the clusters and a small amount of labeled data to adjust the center of the clusters. Experiments on the Reuters-21578 and BBC News data collections show that the proposed model is superior to other semi-supervised approaches in both text classification and text clustering.

doi: *10.5829/ije.2021.34.12c.10*

1. INTRODUCTION

News documents on web pages as well as social networks are the main source of textual data due to the widespread use of Internet [1]. News articles flood the web every day through many major or minor news portals around the world. As the amount of online information resources increases rapidly, so does the content of available online news [2]. To analyze a large number of documents, text clustering is applied, which is a method of dividing a group of documents into different clusters based on content similarity [3]. This method has many applications in news recommender systems [4-5], news classification, emotion analysis [6], text summarization [7], etc.

The clustering function relies mainly on the representation of documents that aims to convert raw documents into numerical vectors. The most common way to represent a document, known for its interpretability and intuition, is the Bag-of-Words method [8], which represents a document vector with its word frequencies. However, although it is easy to interpret, it suffers from unreasonable dimensions. Deep neural network methods such as Convolutional Neural Networks (CNNs) [9] and Doc2Vec [10] create reasonable dimensional vectors to represent documents. Nevertheless, the resulting representations are not easy to interpret because the constituent values of the document vector are calculated through complex neural network weight structures.

*Corresponding Author Institutional Email: *hmashayekhi@shahroodut.ac.ir* (Hoda Mashayekhi))

Please cite this article as: S. M. Sadjadi, H. Mashayekhi, H. Hassanpour, Proposing a Model For a Resilient Supply Chain: A Meta-heuristic Algorithm, International Journal of Engineering, Transactions C: Aspects Vol. 34, No. 12, (2021)

Clustering is a primary method for discovering the natural structure of unlabeled data [11]. One of the newest methods is the use of labeled data to improve the performance of unsupervised clustering [12]. The basic idea is that unlabeled data form the overall structure of the clusters, and some labeled data set the center of the clusters. This method uses both labeled and unlabeled data, called semi-supervised clustering [13]. Nowadays many semi-supervised clustering methods have been proposed for various applications. In clustering methods such as SOM [14] and Naïve Bayes Expectation-Maximization [15], unlabeled data is first labeled, and then these new labeled data and the original labeled data train the model. But it is not clear how much data needs to be re-labeled and how reliable it is.

To solve these problems, we introduce a new two-level method for semi-supervised documents clustering, which makes full use of labeled and unlabeled data, while maintaining proximity information and high interpretability of documents. The words are represented in vectors using the Word2Vec [16] algorithm to utilizing the semantic similarity of the continuous space. In the first level, similar word vectors are grouped into clusters. In the second level, documents are represented based on these clusters. This proposed method can obtain the underlying components of documents while maintaining their interpretability. A semi-supervised clustering algorithm is applied on the documents in the new space to obtain the final document clusters. Because the model is explicable, it provides humans deeper understandings of texts and more explicit operation logic for reasoning.

This paper is organized as follows. In section 2, some related works of document representation and semi-supervised clustering are reviewed. Our proposed concept-based model for semi-supervised document clustering is presented in section 3. Section 4 presents the datasets used and the experimental results, and detailed analyses are presented in section 5. Ultimately, our work is concluded in section 6.

2. RELATED WORKS

2.1. Text Representation The Bag-of-Words (BoW) method has limitations such as large dimensionality and suffering from sparsity. Some succeeding representation techniques, such as Latent Semantic Analysis (LSA) [17] diminishes the term-document matrix into a low dimension matrix. Although it works more efficiently than the BoW method, it diminishes the matrix in linear space and fails to identify the non-linear semantic similarities between the words. The Word2Vec [18], a two-layer neural network, is a model for transforming large text into a multidimensional vector space. As the name implies, by training neural network weights, each word in the raw corpus is represented as a unique vector that can maintain a semantic similarity between words. One of the most important contributions of Word2Vec is that words that occurred in a similar context will be close in embedded space and will preserve the semantic similarities between the words. Also, while high dimensions and sparsity are weaknesses of BoW, the vectors produced by Word2vec have reasonable, optimal, and dense dimensions. For this reason, many machine learning and text data mining problems can be solved through Word2Vec [19].

Le et al. [10] proposed the Doc2Vec model, which utilizes textual information from words and paragraphs mutually to obtain the representation of texts in a continuous vector space. Due to the fewer dimensions of the produced document vectors, it is more effective than BoW. In addition, research has shown that Doc2Vec is more effective than Word2Vec in solving clustering problems [20]. Nevertheless, low interpretability and unclear logic behind document vectors' generation procedure are the problems of the Doc2Vce method.

In this study, the documents are represented based on the concepts in the text. In this regard, Kim et al. [16] proposed the Bag-of-Concepts (BoC) method. It creates concepts through clustering word vectors generated by Word2Vec. Then, the document vector is formed considering the frequency of concepts in the documents. But this method does not suggest a solution for text clustering. Jia et al. [21] used the concept decompositions method to cluster short texts. They presented a decomposition approach to obtain concept vectors that generate by identifying the semantics of word communities in a weighted word co-occurrence network extracted from the short text set.

Lee et al. [22] proposed a new way for representing documents. Their method is based on concepts that automatically receive appropriate conceptual knowledge from an external knowledge base and then conceptualizes the words and terms of the documents with a probabilistic approach. Their method, using an external knowledge base, provides a better understanding of document representation for humans. They also diminish concept ambiguity through clustering concepts with related meanings to improve the BoC algorithm. To evaluate the performance of the proposed method, their model is evaluated in the field of document classification.

2.2. Semi-Supervised Clustering Semi-supervised clustering is considered an alternative to

conventional unsupervised methods. A complete review of some semi-supervised clustering algorithms is presented by Zhu et al. [23].

In a study, Dara et al. [14] used self-organizing map (SOM) for semi-supervised clustering of texts. First, unlabeled texts are labeled, and then these texts, along with the previously labeled texts, are used to train the classifiers. However, their proposed method does not specify how much re-tagging of unlabeled data is required, which is one of the disadvantages of this method. A combination of Naive Bayes and Expectation Maximization (NBEM) algorithms for semi-supervised clustering was also presented [15]. This model repeatedly tags unlabeled data in a loop and uses this newly labeled data to retrain the model. Basu et al. [24] suggested MCP KMEANS, a method that merges two similarity-based and search-based clustering approaches. Although a combination of these two approaches may enhance clustering quality, their objective function may fall to a local minimum. Zhang et al. [25] designed an algorithm named TESC for text classification using semi-supervised clustering. The main difference between this method and other semi-supervised methods is that this method uses labeled and unlabeled documents together. The TESC algorithm assumes that the document set consists of several components and uses a clustering process to obtain these text components. After clustering, the process of classifying test documents is based on calculating the distance to the clusters' centroids.

Lee et al. [26] proposed a distributed method for semi-supervised documents clustering similar to the TESC algorithm. The difference between this method and the TESC method is that clustering is distributed and performed by several sub-algorithm simultaneously. The results are then collected from sub-clusters. The advantage of this method is higher speed and accuracy that can compete with the TESC method. Gan et al. [27] state that prior knowledge can reduce the quality of semi-supervised clustering if incorrectly collected. The basic premise is that when the label of a labeled sample is identified as risky, the predictions of the labeled instance and the nearest homogeneous unlabeled instances should be similar. This is performed through unsupervised clustering then creating a local graph to model the similarities between the labeled and the nearest unlabeled instances.

In another algorithm, document clustering using automatic generation constraints is applied to classify documents [28]. The intrinsic structure of the text data is analyzed using a partial clustering algorithm. The clustering algorithm allows reaching a set of must-link/cannot-link constraints that can be applied in semi-supervised clustering. Constraints are then considered as a semi-supervision factor in a hierarchical clustering algorithm.

Lu et al. proposed a method that uses concept factorization to improve document clustering performance with supervisory data [29]. This approach involves pairwise penalty and reward constraints on conceptual factorization, which can guarantee that the data points of a cluster in the main space are still in the same cluster in the converted space.

In this paper, we present a method that uses labeled and unlabeled data simultaneously; however, our method is different from earlier approaches as well as the TESC method. In the TESC method, most data are labeled and only less than 3% of the data are unlabeled. In the proposed method of this research, large fractions of data are unlabeled and only a limited number of labeled data are used. This difference significantly reduces the cost of data tagging in real-world applications. In addition, most of the mentioned semi-supervised document clustering methods neglect the issue of document representation, which can greatly affect the clustering results. In this paper, a semi-supervised document clustering algorithm based on the conceptual representation of documents is presented that can be used in a variety of applications.

3. THE PROPOSED METHOD

This paper introduces an innovative semi-supervised clustering approach for news documents based on their conceptual representation. It is assumed that the input document set is split into unlabeled and labeled documents. Each document is constituted of a set of words. The purpose is to reach a clustering model $C = \{C_1, \dots, C_m\}$ of the documents, such that $\bigcup_{1 \le i \le m} C_i = D$ and $C_i \cap C_j = \emptyset$ $(1 \le i \ne j \le m), where(D) = Document\ set$.

Figure 1 presents the complete training procedure of the suggested model, expressed in terms of three steps: preprocessing, document representation, and clustering. This method, which represents documents based on their constituent concepts, takes advantage of the simultaneous use of both labeled and unlabeled data types for document clustering. In the following sections, we describe in detail all three steps of the proposed model.

3.1. Text Preprocessing Initially, documents are tokenized after removing stop-words and pre-processing the texts. The word embedding model Word2Vec [30] is utilized to train word relationships from the input document set. The tokenized words of documents set are employed as an input for training the Word2Vec model. Consequently, each token in the words set (W) is

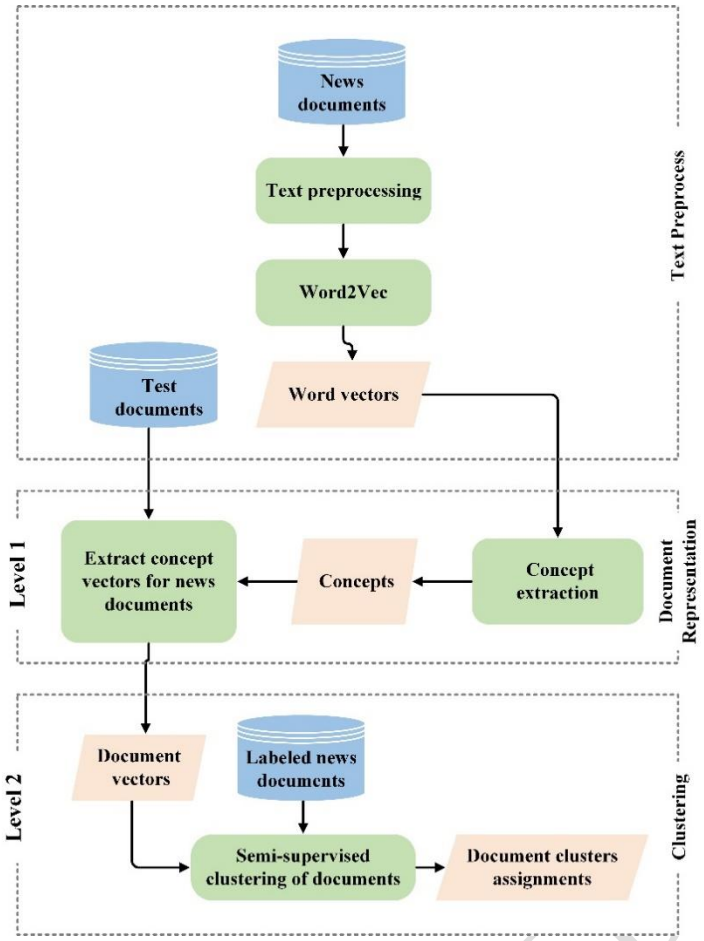

Figure 1. Proposed document clustering model

represented with a dense vector in the embedded space. The most notable contribution of the Word2Vec neural network model is that words that occur in a similar text are placed close to each other in the embedded space, after clustering these embedded words, words with related meanings are placed in the same cluster and concept, which helps to maintain semantic relationships between words.

3.2. Documents Representation Documents representation is based on the concepts extracted from the data corpus. In the document representation stage, firstly a set of concepts are extracted from the set of words W, such that each concept consists of an exclusive set of words. The main idea for deriving concepts is to implement a clustering algorithm on a set of words (W) to group it into several clusters, each of which represents a concept. Following the construction of the concepts, each document is represented by a vector formed by concepts ($\vec{d}$).

The Spherical K-Means clustering algorithm employing the cosine distance is applied to cluster word vectors. The procedure of the Spherical K-Means algorithm is the same as the K-Means clustering algorithm and it assigns each data to a cluster with a predetermined value for the number of clusters, and updates each cluster center according to the cluster data membership in the previous iteration. Since Word2Vec maximizes the cross-product of embedded vectors and context vectors, the cosine distance has been used as the proper criterion for clustering nearby word vectors into a common cluster and measuring distances between word vectors in semantic space.

Each cluster created by Spherical K-Means clustering is considered as a concept. Document vectors are constructed using these concepts. Words with similar meanings are divided into the same cluster according to the clustering efficiency and semantic space trained by the Word2Vec. Therefore, each word in the text corpus will be regarded as a concept's member. Because each word may be present in many documents, it is not a proper discriminator for machine learning applications [31], so Concept Frequency - Inverse Document Frequency (CF-IDF) Equation (1) is applied to the produced word vectors to eliminate the unfavorable effects of common words between concepts.

$$CF-IDF(c_i,d_j,D) = CF(c_i,d_j) \times \log \frac{|D|}{|d \in D\ ;\ c_i \in D|} \quad (1)$$

$$where\ (c_i,d_j,D) = (Concept_i, Document_j, Corpus)$$

The number of concepts and consequently the length of document vectors are arbitrary and defined by the user considering the processing complexity and storage constraints. It may also be determined experimentally according to the dataset. In this regard, the clustering accuracy may be evaluated for an increasing number of concept. As reported later in the experiments, it is observed that after a certain value, the accuracy does not change significantly. This value can determine number of concepts. After extracting the document vectors, it is time to cluster the documents. For this purpose, the clustering algorithm is implemented on the conceptual vectors of documents.

3.3. Semi-supervised Clustering Once the documents have been created based on the conceptual representation, it is time to cluster the constructed document vectors. In the clustering process, two documents with similar concepts are expected to have the same vectors. In this step, which uses both unlabeled and labeled data types, labeled documents are used as supervisors of the

clustering process. Labeled and unlabeled document vectors are entered as input to the semi-supervised clustering algorithm. Spherical K-Means clustering is used to partition vectors. The resulting clusters may contain data from several different labels, so in a purification process described below, the gross clusters are broken down into smaller pure clusters.

Based on the data labels in each cluster, the proposed algorithm decides whether the cluster needs to be purified or is already pure. Also, the decision on how many smaller clusters to break the gross cluster is one of the tasks of the clustering algorithm, which is performed according to the following purification procedure which is repeated until all clusters have a label:

1. The cluster contains data from only one label: the cluster is transferred to the final clustering result.

2. The cluster contains both unlabeled data and data from one label: The label of labeled data is selected as the cluster label.

3. The cluster contains several different labels: The cluster is divided into the number of labels and each of these sub-clusters contains only one type of data label.

4. The cluster contains several labels and unlabeled data: Purification is performed according to procedure 3, with the difference that unlabeled data will also have a separate sub-cluster.

5. The cluster is composed entirely of unlabeled data: Using the cosine distance, the nearest center of the labeled cluster is selected and its label is assigned as the label of this unlabeled cluster.

Once the purification is complete, all clusters will have an appropriate label.

After the document vectors are clustered using the semi-supervised clustering described in Figure 1, the document output clusters are identified as components of the text corresponding to the document categories. Each of these clusters is labeled and can be used in the test data clustering process. Each test data uses the cosine distance to find the nearest center of the cluster and chooses the label of that cluster as its label.

The method proposed in this research has the following contributions:

Previous methods of document representation have disadvantages such as not maintaining non-linear semantic relationships between words. Also, neural network-based methods such as Doc2Vec suffer from low interpretability. The method proposed in this paper is based on the conceptual representation of documents, in addition to maintaining non-linear relationships, has high interpretability and intuition. Since the proposed method is a semi-supervised clustering method, large amounts of data can be clustered and categorized with acceptable accuracy with low overhead and low cost. This point is beneficial in the application of social networks and stream data where the amount of unlabeled data is large. One of the most important advantages of this method over deep learning methods is that, unlike deep learning networks, the logic of the proposed method is clear, and with the addition of new data, there is no need to re-train the model.

3.4. Complexity Analysis Since the proposed model consists of two levels, the time complexity of each level is calculated separately and the training total time complexity is obtained from the sum of these two values. At the first stage, to form the concept vectors of documents, due to the existence of the Word2Vec model, the time complexity value is equal to $O(N * \log(V))$, where N is the total corpus size and V is the unique-words vocabulary count [17]. Also, the time complexity of the concept extraction part is equal to $O(tkV)$, where t is the number of iterations of the algorithm, and k is the number of concepts.

In the second step, the semi-supervised clustering algorithm is calculated with time complexity $O(t'mN + N)$, where t' is the number of iterations of the algorithm, and m is the number of final clusters. As a result, the overall time complexity of the architecture presented in this method is a maximum of $O(N * \log(V) + tkV + t'mN)$.

4. DATASET

In this paper, two datasets consisting of news documents are used to evaluate the proposed model. The Reuters-21578 news dataset includes a collection of news items published on the news agencies' websites. The Reuters-21578 set of documents is related to the news that was published on the Reuters website in 1987, which was collected by Reuters' staff in 1991. In this study, 2110 documents from four different categories are chosen as "agriculture" (571 documents), "crude" (580 documents), "trade" (483 documents), and "interest" (476 documents) are randomly selected after deleting the uncommon words.

The second dataset is the BBC News documentation, which includes 2,225 news documents published from 2004 to 2005, and was compiled in five categories in 2006. In this study, all 2225 documents from five different categories are chosen as follows: "tech" (401 documents), "sport" (511 documents), "politics" (417 documents), "entertainment" (386 documents), and "business" (510 documents).

Some common natural language pre-processing tasks, such as case folding (converting uppercase to

lowercase letters), removing punctuations, removing stop-words, and tokenization, are applied to the document collection. For fast Word2Vec training, words that have occurred less than 5 times in the entire datasets are removed.

5. EXPERIMENTS

Various experiments have been designed and performed to observe the performance of the proposed model. The proposed model is compared with K-Means, Bag-of-Concepts (BoC) [16], TESC (BoW) [25], and Doc2Vec [10]. To compare, there is a need for criteria to measuring the efficiency of the mentioned methods, which are described in the following.

The Normalized Mean Squared Error (NMSE) criterion expresses the quality of the clustering performed. This criterion calculates the average squares of the errors, and the normalized numerical value gives an output between 0 and 1, and smaller values show a lower variance within the cluster. The NMSE metric is defined in Equations (2) and (3), in which μ is the set of cluster centers, X is the set of data points, and μ_{c_i} is the cluster centroid of the data point x_i.

$$MSE\ (X,\mu) = \frac{1}{N}\sum_i (x_i - \mu_{c_i})^2 \quad (2)$$

$$NMSE(X,\mu) = \frac{1}{N}\frac{\sum_i (x_i - \mu_{c_i})^2}{\sum_i (x_i)^2} \quad (3)$$

The Normal Mutual Information (NMI) is a cluster criterion that evaluates the quality of data clustering according to their pre-given labels. The NMI evaluates how the clustering algorithm can reconstruct the original data labels [32]. This criterion can be used when the data label is available. The output of this numerical criterion is in the range [0,1], which shows the statistical similarity between the labels of the generated clusters and the original labels of the data. A value of zero indicates a failed cluster assignment, while values close to one indicate that clustering can recreate real data classes. The NMI criterion shows better performance in presenting the quality of clusters than the entropy criterion. This is because the entropy criterion depends on the number of clusters, and the higher the number, the better the entropy criterion. But the NMI standard is not like this and does not necessarily increase as the number of clusters increases. Equation (4) shows the mathematical definition of this criterion.

$$NMI = \frac{I(C;K)}{(H(C) + H(K))/2} \quad (4)$$

$$I(X;Y) = H(X) - H(X|Y) \quad (5)$$

Equation (5) is the mutual information between the random variables X and Y , $H(X)$ is the Shannon entropy of X, $H(X|Y)$ is the conditional entropy of X given Y, C is the set of class labels and K is the set of cluster labels.

In this paper, not only the quality of the generated clusters is evaluated, but also the real application of this method in the classification of news documents is evaluated. For this purpose, the classification accuracy criterion is introduced, which is a criterion that expresses the performance of a classifier with a percentage value. This value shows that of all the test data, how many data are rightly classified. By dividing the number of rightly classified samples by the total number of samples, the amount of accuracy is obtained. Equation (6) shows the measure of accuracy. In this regard, $\hat{y}_l$ is the class prediction for example l.

$$Accuracy = \frac{\sum_i 1(\hat{y}_l = y_i)}{|X|} \quad (6)$$

5.1. Results

5.1.1. Effect of the Number of Concepts In document representation, the number of concepts determines the length of the document vector. Therefore, it would have a significant effect on the performance quality of the proposed model. The performance of the proposed model, in terms of clustering quality and classification accuracy, when the number of concepts varies, is shown in Table 1. According to the results of this table, the best performance occurs when the number of concepts is 300 and after that, there is no noticeable increase in both clustering and classification accomplishments. Compared to BoW method, which depends on the number of words in the text, a significant improvement in classification accuracy is observed. Also, compared to BoC method, which displays the text conceptually, it is observed that with the addition of labeled documents, the proposed model shows its superiority. In subsequent experiments, the length of the document vector is assumed to be 300.

5.1.2. Effect of Window Size In the suggested model, to maintain nonlinear semantic relations between words, a word embedding method Word2Vec is used. Word2Vec neural network training depends on parameters that one of the most important parameters is the size of the window. At each stage of the neural network training, a slider window is moved on the text so that the words in this

window can be used as input and output of the neural network. Experiments have shown that the larger the window size, the model would be trained better, and the generated word vectors would be more effective as a result of clustering.

Tables 2 and 3 examine the effect of window size changes on clustering quality and document classification accuracy. As shown in Tables 2 and 3, the performance of the proposed model improves as expected by increasing the window size. This performance improvement is obtained because the neural network encounters more words at each stage and can predict output more likely. Semantic relationships between words are more discovered and have a significant effect on the weight of the neural network. In this experiment, 80 percent of data is used for training with 200 labeled documents.

TABLE 1. Performance of the proposed model when the number of concepts varies - (Reuters-21578)

Number of Concepts	**100**	**200**	**300**	**400**	**500**	**600**
Proposed Model Classification Accuracy (%)	69.33	74.20	76.32	76.12	77.02	77.13
BoC Classification Accuracy(%) [16]	66.31					
TESC (BoW) Classification Accuracy (%) [25]	62.65					
Proposed Model Clustering NMSE	0.124	0.114	0.106	0.107	0.105	0.103
BoC Clustering NMSE [16]	0.1340					
TESC (BoW) Clustering NMSE [25]	0.1803					

TABLE 2. Performance of the proposed model when the size of the window varies - (Reuters-21578).

Window Size	**4**	**8**	**20**
Classification Accuracy	63.9 %	67%	72%
NMI	0.274	0.33	0.38
NMSE	0.1191	0.1031	0.1007

TABLE 3. Performance of the proposed model when the size of the window varies - (BBC News)

Window Size	**4**	**8**	**20**
Classification Accuracy	74.5 %	76.7%	76.8%
NMI	0.421	0.449	0.511
NMSE	0.1038	0.0961	0.0904

The values mentioned for NMI and NMSE indicate that as the window size increases in word embedding, the quality of the clusters also improves. For example, in Table 2 (Reuters-21578), when the window size changes from 4 to 20, the NMI value increases from 0.274 to 0.33, which indicates better quality. The mean squared error also shows a decreasing trend. As the evaluation metrics are negligibly improved in larger window sizes, in order to avoid additional overhead and reduce the time complexity, a window size of 8 is considered to train the Word2Vec model in subsequent experiments.

5.1.3. Effect of the Number of Labeled Documents

Another factor influencing the quality of semi-supervised text clustering is the number of labeled documents. To observe the effects of labeled data on the quality of clustering, an experiment is designed in which the number of labeled data changes though the number of unlabeled data is kept constant. In this analysis, 80% of the documents are used for training and the remaining 20% for testing. Tables 4 and 5 show the NMI values of proposed model clustering for various numbers of labeled documents compare to other methods when the number of unlabeled documents is fixed.

Clustering with the proposed model on Reuters-21578 is of better quality than other methods. It is also noteworthy that as the number of labeled documents increases, the NMI value and therefore the clustering quality increases significantly. For example, in a case, when 9% of all documents are labeled (200 documents), the value of The proposed method is 0.370, TESC (BoW) 0.189, Doc2Vec 0.292, and K-Means 0.261. In the worst case, when only 4% of all documents are labeled (100 documents), the NMI value of the proposed model does not fall below 0.331, while other methods produce far fewer NMIs and lower quality clusters. The same trend and performance for the BBC News dataset can be seen in Table 5.

As can be concluded from Tables 4, 5, and Figure 2, with the increase of labeled documents, the quality of the resulting clustering has an increasing trend. Comparing the

values in Tables 4 and 5, it can be seen that the architecture presented in this paper for semi-supervised clustering of documents has significantly improved the quality of news document clustering. Because of the concepts and components of the text have been extracted, the proposed method can create cluster labels corresponding to documents classes, which is why the NMI in the proposed method is higher than other methods.

TABLE 4. NMI scores of news document clustering for proposed model compared with K-Means, TESC (BoW), and Doc2Vec at the various percentage of labeled documents – (Reuters-21578)

The percentage of labeled documents (# of labeled documents)	**4% (100)**	**9% (200)**	**14% (300)**	**18% (400)**	**22% (500)**	**27% (600)**
K-Means	0.198	0.261	0.305	0.342	0.367	0.345
TESC [25]	0.165	0.189	0.306	0.397	0.413	0.388
Doc2Vec [10]	0.243	0.292	0.362	0.375	0.413	0.421
Proposed model	**0.331**	**0.370**	**0.435**	**0.457**	**0.460**	**0.481**

TABLE 5. NMI scores of news document clustering for proposed model compared with K-Means, TESC(BoW), and Doc2Vec at the various percentage of labeled documents – (BBC News)

The percentage of labeled documents (# of labeled documents)	**4% (100)**	**9% (200)**	**14% (300)**	**18% (400)**	**22% (500)**	**27% (600)**
K-Means	0.236	0.301	0.338	0.361	0.389	0.390
TESC [25]	0.200	0.237	0.352	0.432	0.463	0.469
Doc2Vec [10]	0.339	0.365	0.393	0.420	0.437	0.494
Proposed model	**0.421**	**0.453**	**0.471**	**0.478**	**0.529**	**0.555**

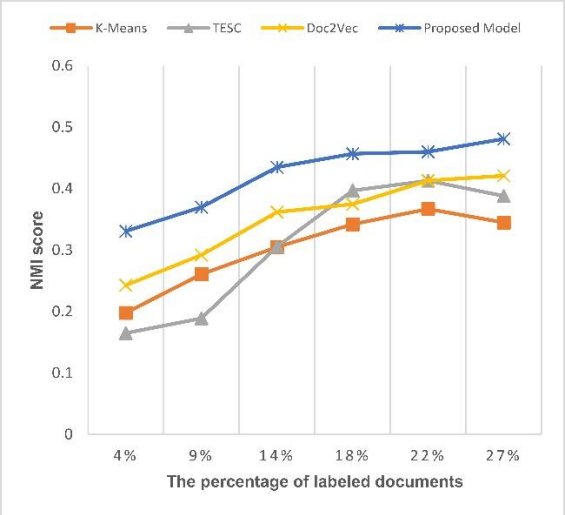

Figure 2. NMI scores of news document clustering for proposed model compared with K-Means, TESC(BoW), and Doc2Vec at the various percentage of labeled documents – (Reuters-21578)

Tables 6 and 7 show the accuracy of the classification of news documents for the proposed method compared to other methods. As can be seen from these tables, the classification accuracy of the proposed method is at least 4% superior to other methods. It is clear that with the increase of labeled documents, the accuracy of classifying news texts has increased.

TABLE 6. News document classification accuracy of the proposed model compared with TESC (BoW), and Doc2Vec at the various number of labeled documents – (Reuters-21578)

Number of labeled documents	**200**	**250**	**300**	**350**	**400**	**450**
TESC (BoW) (%) [25]	63.32	64.02	64.23	65.36	66.32	66.90
Doc2Vec (%) [10]	64.00	67.23	70.64	72.23	71.62	73.17
BoC [16]			66.31			
Proposed model (%)	**72.06**	**74.45**	**75.12**	**76.01**	**76.11**	**77.37**

TABLE 7. News document classification accuracy of the proposed model compared with TESC (BoW), and Doc2Vec at the various number of labeled documents – (BBC News)

Number of labeled documents	200	250	300	350	400	450
TESC (BoW) (%) [25]	64.68	68.19	69.46	71.37	72.11	72.45
Doc2Vec (%) [10]	67.81	69.60	73.42	76.35	76.57	77.92
BoC [16]	69.45					
Proposed model (%)	**75.51**	**76.67**	**77.84**	**78.71**	**81.34**	**81.92**

6. CONCLUSION

In this research, a concept-based method for semi-supervised clustering of news documents is presented. The main idea is that the way documents are represented affects the quality of clustering and classification of documents. For this purpose, a two-level semi-supervised clustering is proposed that extract concepts from corpus words, and represents documents based on the concepts. This method of document representation overcomes the weaknesses of previous methods and has high interpretability by describing documents in low dimensions. The method proposed for clustering document vectors is a semi-supervised method that uses a limited amount of labeled data. This method uses unlabeled data to capture the overall structure of clusters and labeled data to set cluster centers. It also identifies the structure and components of the text and creates clusters corresponding to the data classes. Experiments have shown that the method proposed in this paper has a significant advantage over other methods of semi-supervised clustering of the text. Also, the effect of various parameters such as window size, document length (number of concepts), and number of labeled documents have been studied and evaluated. The results are satisfactory but more studies can be done in the future. For example, the use of N-Grams in training the Word2Vec neural network model may produce better results.

7. REFERENCES

1. Forsati. R, Mahdavi. M, Kangavari. M, Safarkhani. B, "Web page clustering using harmony search optimization", In *2008 Canadian Conference on Electrical and Computer Engineering* IEE 1601–1604, https://doi.org/10.1109/CCECE.2008.4564812.
2. Bouras. C, Tsogkas. V, "A clustering technique for news articles using WordNet", ***Knowledge-Based Systems***, Vol. 36, (2012) 115–128, https://doi.org/10.1016/J.KNOSYS.2012.06.015.
3. Karypis. M, Kumar. V, Steinbach. M, "A comparison of document clustering techniques", (2000). the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/215421.
4. Bobadilla. J, Ortega. F, Hernando. A, Gutiérrez. A, "Recommender systems survey", ***Knowledge-Based Syst.***, Vol. 46, (2013), 109–132, https://doi.org/10.1016/j.knosys.2013.03.012.
5. Barzegar Nozari. R, Koohi. H, Mahmodi. E, "A Novel Trust Computation Method Based on User Ratings to Improve the Recommendation", ***International Journal of Engineering, Transactions C: Aspects***, Vol. 33, (2020), 377–386, https://doi.org/10.5829/IJE.2020.33.03C.02.
6. Djenouri. Y, Belhadi. A, Fournier-Viger. P, Lin. J, "Fast and effective cluster-based information retrieval using frequent closed itemsets", ***Information Sciences***, Vol. 453, (2018), 154–167, https://doi.org/10.1016/j.ins.2018.04.008.
7. Joty. S, Carenini. G, Ng. R, "Topic segmentation and labeling in asynchronous conversations", ***J. Artif. Intell. Res.***,Vol. 47, (2013), 521–573, https://doi.org/10.1613/jair.3940.
8. Li. Y, Guo. H, Zhang. Q, Gu. M, Yang. J, "Imbalanced text sentiment classification using universal and domain-specific knowledge", ***Knowledge-Based Systems***,Vol. 160, (2018), 1–15, https://doi.org/10.1016/j.knosys.2018.06.019.
9. Jacovi. A, Shalom. O, Goldberg. Y, "Understanding Convolutional Neural Networks for Text Classification", ***ArXiv***, (2018), arXiv preprint arXiv:1809.08037.
10. Le. Q, Mikolov. T, "Distributed Representations of Sentences and Documents", ***International Conference on Machine Learning. PMLR***, Vol. 32, (2014), 1188–1196.
11. Zhang. W, Yoshida. T, Tang. X, Wang. Q, "Text clustering using frequent itemsets", ***Knowledge-Based Systems***, Vol. 23, (2010), 379–388, https://doi.org/10.1016/j.knosys.2010.01.011.
12. Cozman. F, Cesar Cirelo. M, "Semi-Supervised Learning of Mixture Models" , Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, Vol. 4, (2003), 4-24.
13. Luo. X, Liu. F, Yang. S, Wang. X, Zhou. Z, "Joint sparse regularization based Sparse Semi-Supervised Extreme Learning Machine (S3ELM) for classification" , ***Knowledge-Based Systems***, Vol. 73, (2015), 149–160, https://doi.org/10.1016/j.knosys.2014.09.014.
14. Dara. R, Kremer. S, Stacey. D, "Clustering unlabeled data with SOMs improves classification of labeled real-world data", ***Proc. International Joint Conference on Neural Networks***, (2002), 2237–2242, https://doi.org/10.1109/ijcnn.2002.1007489.
15. Zhang. W, Yang. Y, Wang. Q, "Using Bayesian regression and EM algorithm with missing handling for software effort prediction", ***Information and Software Technology***, Vol. 58, (2015), 58–70, https://doi.org/10.1016/j.infsof.2014.10.005.
16. Kim. HK, Kim. H, Cho. S, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation", ***Neurocomputing.***, Vol. 266, (2017), 336–352, https://doi.org/10.1016/j.neucom.2017.05.046.
17. Deerwester. S, Dumais. S.T, Furnas. G.W, Landauer. T.K, Harshman. R, "Indexing by latent semantic analysis", ***Journal of the American society for information science,*** Vol. 41, (1990), 391–407, https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.
18. Mikolov. T, Chen. K, Corrado. G, Dean. J, "Efficient estimation of word representations in vector space", ***International Conference on Learning Representations, ICLR***, (2013).

19. Edara. D.C, Vanukuri. L.P, Sistla. V, Kolli. V.K.K, "Sentiment analysis and text categorization of cancer medical records with LSTM", ***Journal of Ambient Intelligence and Humanized Computing***, (2019), 1–17, https://doi.org/10.1007/s12652-019-01399-8.
20. Dai. A.M, Olah. C, Le. Q, "Document Embedding with Paragraph Vectors", ***Arxiv***, (2015), 1–8, arXiv preprint arXiv:1507.07998.
21. Jia. C, Carson. M.B, Wang. X, Yu. J, "Concept decompositions for short text clustering by identifying word communities", ***Pattern Recognition***, Vol. 76, (2018), 691–703, https://doi.org/10.1016/j.patcog.2017.09.045.
22. Li. P, Mao. K, Xu. Y, Li. Q, Zhang. J, "Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base", ***Knowledge-Based Systems***, Vol. 193, (2020), https://doi.org/10.1016/j.knosys.2019.105436.
23. Zhu. X.J, "Semi-Supervised Learning Literature Survey", (2005). http://digital.library.wisc.edu/1793/60444
24. Basu. S, Bilenko. M, Mooney. R.J, "Comparing and Unifying Search-Based and Similarity-Based Approaches to Semi-Supervised Clustering", ***Proceedings of the ICML-2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining***, (2003), 42-49.
25. Zhang. W, Tang. X, Yoshida. T, "TESC: An approach to TExt classification using Semi-supervised Clustering", ***Knowledge-Based Systems***, Vol. 75, (2015), 152–160, https://doi.org/10.1016/j.knosys.2014.11.028.
26. Li. P, Deng. Z, "Use of distributed semi-supervised clustering for text classification", ***Journal of Circuits, Systems and Computers,*** Vol. 28, No. 8 (2019), 1–13, https://doi.org/10.1142/S0218126619501275.
27. Gan. H, Fan. Y, Luo. Z, Zhang. Q, "Local homogeneous consistent safe semi-supervised clustering", ***Expert Systems with Applications,*** Vol. 97, (2018), 384–393, https://doi.org/10.1016/j.eswa.2017.12.046.
28. Diaz-Valenzuela. I, Loia. V, Martin-Bautista. M.J, Senatore. S, Vila. M.A, "Automatic constraints generation for semisupervised clustering: experiences with documents classification", ***Soft Computing***, Vol. 20, No. 6 (2016), 2329–2339, https://doi.org/10.1007/s00500-015-1643-3.
29. Lu. M, Zhao. X.J, Zhang. L, Li. F.Z, "Semi-supervised concept factorization for document clustering", ***Information Sciences***, Vol. 331, (2016), 86–98, https://doi.org/10.1016/j.ins.2015.10.038.
30. Mikolov. T, Sutskever. I, Chen. K, Corrado. G, Dean. J, "Distributed Representations of Words and Phrases and their Compositionality", In ***Advances Neural Information Processing Systems,***, (2013) 3111-3119.
31. Robertson. S, "Understanding inverse document frequency: On theoretical arguments for IDF", ***Journal of Documentation,*** Vol. 60, No. 5 (2004), 503–520, https://doi.org/10.1108/00220410410560582.
32. Strehl. A, Ghosh. J, Mooney. R, "Impact of Similarity Measures on Web-page Clustering", ***Workshop on artificial intelligence for web search***, Vol. 58, (2000), 64.

Persian Abstract

چکیده

صفحات وب و رسانه‌های اجتماعی از نظر مقدار و تنوع مملو از اخبار هستند. خوشه‌بندی اسناد یک روش مفید است که به طور گسترده‌ای در سازماندهی و مدیریت داده‌ها به گروه‌های کوچکتر استفاده می‌شود. یکی از عوامل تأثیرگذار بر کیفیت خوشه‌بندی ، نحوه بازنمایی اسناد است. برخی از روش‌های سنتی بازنمایی اسناد به تکرارهای کلمه در متن بستگی دارند و بردارهای سند پراکنده و بزرگی را ایجاد می‌کنند. این روش‌ها نمی‌توانند اطلاعات مجاورتی بین اسناد را حفظ کنند. علاوه بر این ، روش‌های مبتنی بر شبکه عصبی که اطلاعات مجاورتی را حفظ می‌کنند، از تفسیرپذیری ضعیف رنج می‌برند. روش‌های بازنمایی متنِ مبتنی بر مفاهیم بر کاستی‌های روش‌های قبلی غلبه می‌کنند، اما روش‌های خوشه‌بندی نیمه‌نظارتی متن در حال حاضر از نمایش اسناد مبتنی بر مفهوم استفاده نمی‌کنند. در این مقاله یک روش خوشه‌بندی نیمه‌نظارتی متون خبری مبتنی بر مفهوم ارائه شده است که برای دستیابی به کیفیت خوشه‌بندی بالاتر از داده‌های دارای برچسب و بدون برچسب به طور همزمان استفاده می‌کند. در مرحله اول اسناد بر اساس مفاهیم استخراج شده از مجموعه اسناد نمایش داده می‌شوند. سپس، فرآیند خوشه‌بندی نیمه‌نظارتی داده‌های بدون برچسب را برای گرفتن ساختار کلی خوشه‌ها و مقدار کمی از داده‌های دارای برچسب را برای تنظیم مرکز خوشه‌ها به طور همزمان اعمال می‌کند. آزمایش‌های انجام شده بر روی مجموعه داده های Reuters-21578 و BBC News نشان می‌دهد که مدل پیشنهادی هم در طبقه‌بندی متن و هم در خوشه‌بندی متن از سایر روش‌های نیمه نظارتی بهتر عمل می‌کند.